7 Schriften aus der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

# Fehlende Werte in den Sozialwissenschaften

## Analyse und Korrektur mit Beispielen aus dem ALLBUS

von Martin Messingschlager

UNIVERSITY OF
BAMBERG
PRESS

Schriften aus der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg 7

Schriften aus der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg

Band 7


University of Bamberg Press 2012

# Fehlende Werte in den Sozialwissenschaften

Analyse und Korrektur mit Beispielen aus dem ALLBUS

von Dipl.-Pol. Martin Messingschlager

University of Bamberg Press 2012

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.ddb.de/ abrufbar

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.
1. Gutachter: Prof. Dr. Susanne Rässler
2. Gutachter: Prof. Dr. Olaf Struck
Tag der mündlichen Prüfung: 24. September 2012



Herstellung und Druck: docupoint GmbH, Barleben
Umschlaggestaltung: Dezernat Kommunikation und Alumni der Otto-Friedrich-Universität Bamberg



ISSN: 1867-6197
ISBN: 978-3-86309-122-4 (Druckausgabe)
eISBN: 978-3-86309-123-1 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-16185

*meinen Eltern*

## Danksagung

Die Begeisterung für Statistik hat mich als Politikwissenschaftler bereits während des Studiums gepackt und mich schließlich entscheiden lassen, eine Promotion im Fach Statistik zu beginnen. Diese Begeisterung wurde maßgeblich von einem Mann genährt, der leider die Abgabe dieser Promotion selbst nicht mehr miterleben kann. Ohne die ideelle Unterstützung von Professor Friedrich Vogel, wäre diese Arbeit wahrscheinlich nie begonnen worden. Ihm spreche ich daher zuvörderst größten Dank aus. So wie er mich auf den Kurs der Promotion gebracht hat, hat Professorin Susanne Rässler mit ganzem Eifer und Hingabe dafür gesorgt, dass ich die letzten Jahre – auch bei schwerem Seegang – Kurs gehalten habe. Ihr danke ich von ganzem Herzen für die fachliche Betreuung wie menschliche Unterstützung. Daran schließt sich ein herzlicher Dank an Professor Olaf Struck an, der als Arbeitswissenschaftler ohne Zögern die Zweitbetreuung für meine Promotion zugesagt hat.
Neben den großen Stützen für ein so langwieriges Projekt wie eine Dissertation gab es immer wieder viele Helferinnen und Helfer. Namentlich möchte ich Hans Walter Steinhauer für die Hilfe bei der Programmierung diverser R-Codes danken. Ebenso möchte ich mich bei Marcel Preising und Carolin Fleischmann für Zeit und Geduld bei der Korrektur der Arbeit bedanken.
Die Replikation der Veröffentlichungen von Eike Hennig, Armin Schäfer sowie von Claus Schnabel und Joachim Wagner wäre ohne deren uneingeschränkte Unterstützung wesentlich schwerer gewesen; vielen Dank hierfür. Schließlich sei noch auf einen wichtigen Punkt hingewiesen: Der erfolgreiche Abschluss dieser Arbeit wurde nicht zuletzt von dem kollegialen und großartigen Klima am Lehrstuhl für Statistik und Ökonometrie getragen. Allen Kolleginnen und Kollegen, Sekretärinnen und wissenschaftlichen Hilfskräften ein herzliches Dankeschön für ihre Unterstützung.

## Inhaltsübersicht

**1 Fehlende Werte im Kontext sozialwissenschaftlicher Erhebungen** . . . . . 1

**2 Übersicht und Definitionen** . . . . . 3

2.1 Begriff . . . . . 3

2.2 Definitionen . . . . . 5

2.3 Kontext: Erhebung und Fehler . . . . . 6

2.4 Umgang mit fehlenden Werten . . . . . 10

2.5 Zwischenfazit . . . . . 14

**3 Item Nonresponse: Theorie und Determinanten** . . . . . 15

3.1 Einleitung . . . . . 15

3.2 Theorie zur Entstehung von Item Nonresponse . . . . . 15

3.2.1 Übersicht: Item Nonresponse begünstigende und mindernde Faktoren . . . . . 16

3.2.2 Interview: Interaktion von Interviewer und Befragten . . . . . 19

3.2.3 Theorie zu Entscheidungsprozessen bei Item Nonresponse . . . . . 22

3.2.4 Konsequenzen für den Umgang mit Item Nonresponse . . . . . 25

3.3 Determinanten . . . . . 28

3.3.1 Analyse von Item Nonresponse . . . . . 29

3.3.2 Item Nonresponse als Zähldaten . . . . . 32

3.3.3 Erklärungsmodell für Item Nonresponse im ALLBUS 2006 . . . . . 39

**4 Item Nonresponse: Korrekturmethoden im Vergleich** . . . . . 47

4.1 Einleitung . . . . . 47

4.2 Ausgewählte Beispiele . . . . . 47

4.2.1 Beispiel 1: Anteilswerte . . . . . 48

4.2.2 Beispiel 2: Multivariates Probitmodell . . . . . 52

4.2.3 Beispiel 3: Multivariates Logitmodell und individuelle Eintrittswahrscheinlichkeiten . . . . . 55

4.3 Verfahren zum Vergleich von Korrekturmethoden . . . . . 58
4.3.1 Konstruktion eines Stresstests . . . . . 61
4.3.2 Ausgewählte Korrekturverfahren . . . . . 65
4.3.3 Ergebnisse des Methodenvergleichs . . . . . 67
4.3.3.1 Ergebnis 1: Anteilswerte . . . . . 68
4.3.3.2 Ergebnis 2: Parameter des Probitmodells . . . . . 76
4.3.3.3 Ergebnis 3: Parameter des Logitmodells und individuelle Eintrittswahrscheinlichkeiten . . . . . 90
4.3.3.4 Zusammenfassung . . . . . 103
4.4 Zwischenfazit Item Nonresponse . . . . . 104

**5 Unit Nonresponse: Theorie und Determinanten** . . . . . 105

5.1 Kontext: Unit Nonresponse im Survey Lifecycle . . . . . 105
5.1.1 Nichterreichbarkeit im weiteren Sinne (Undercoverage) . . . . . 109
5.1.2 Nichterreichbarkeit im engeren Sinne . . . . . 111
5.1.3 Nichtbefragbarkeit . . . . . 113
5.1.4 Verweigerung . . . . . 113
5.1.4.1 Konkretisierung von RC in habitualisierten Verhaltenstendenzen und skripttheoretischer Spezifizierung . . . . . 115
5.1.4.2 Leverage-Salience-Theorie . . . . . 116
5.1.4.3 Konzept einer wertrationalen Erklärung für die Teilnahme . . . . . 120
5.1.5 Überlegungen zum Ausfallmechanismus bei Unit Nonresponse . . . . . 122
5.2 Messung von Unit Nonresponse . . . . . 123
5.2.1 Problematik der Ausschöpfungsquote: generelle Trends . . . . . 123
5.2.2 Ausschöpfungsquote beim ALLBUS . . . . . 125
5.2.2.1 Veränderung der Ausschöpfungsquote beim ALLBUS . . . . . 127
5.2.2.2 Analyse der Ausschöpfungsquote . . . . . 128
5.2.3 Zusammenfassung . . . . . 134
5.3 Praxis: ALLBUS 2008 . . . . . 135
5.3.1 Ausfallgründe und Ausschöpfungsquote . . . . . 135
5.3.2 Erklärungsmodell für Unit Nonresponse . . . . . 139

**6 Unit Nonresponse: Korrekturmethoden im Vergleich** . . . . . 147

6.1 Verfahren zum Vergleich von Korrekturmethoden: Modifikation für Unit Nonresponse . . . . . 147

6.2 Ausgewählte Parameter . . . . . 150

6.2.1 Uni- und multivariate Parameter . . . . . 150

6.2.2 Stresstest . . . . . 151

6.2.3 Ausgewählte Korrekturverfahren . . . . . 154

6.2.3.1 Gewichtung . . . . . 154

6.2.3.2 Multiple Imputation . . . . . 156

6.3.4 Ergebnisse . . . . . 158

6.3.4.1 Ergebnis 1: Anteilswert . . . . . 158

6.3.4.2 Ergebnis 2: Mittelwert . . . . . 161

6.3.4.3 Ergebnis 3: OLS-Modell . . . . . 164

6.3.4.4 Ergebnis 4: Logitmodell . . . . . 174

6.3.4.5 Zusammenfassung . . . . . 179

6.3 Zwischenfazit Unit Nonresponse . . . . . 180

**7 Keine Angst vor Problemen mit Zähnen** . . . . . 181

**Anhang und Verzeichnisse**

A Verzeichnis regulärer Anhänge . . . . . iv

B Abbildungsverzeichnis . . . . . vi

C Tabellenverzeichnis . . . . . xii

**Literaturverzeichnis** . . . . . 213

# A Verzeichnis regulärer Anhänge

Anhang 1: Liste der für den Item Nonresponse-Vektor verwendeten Variablen des ALLBUS 2006 . . . . . 183

Anhang 2: Items für Beispiel 1 des Methodenvergleichs bei Item Nonresponse . . . . . 188

Anhang 3: Items für Beispiel 2 des Methodenvergleichs bei Item Nonresponse . . . . . 189

Anhang 4: Items für Beispiel 3 des Methodenvergleichs bei Item Nonresponse . . . . . 190

Anhang 5: Sonstige visuelle Aufbereitung für Beispiel 1 des Methodenvergleichs bei Item Nonresponse . . . . . 191

Anhang 6: Sonstige visuelle Aufbereitung für Beispiel 2 des Methodenvergleichs bei Item Nonresponse . . . . . 197

Anhang 7: Sonstige visuelle Aufbereitung für Beispiel 3 des Methodenvergleichs bei Item Nonresponse . . . . . 201

Anhang 8: Änderung der erfassten Ausfallkategorien des ALLBUS von 1980-2008 . . . . . 205

Anhang 9: Item für den Anteilswert des Methodenvergleichs bei Unit Nonresponse . . . . . 206

Anhang 10: Item für den Mittelwert des Methodenvergleichs bei Unit Nonresponse . . . . . 206

Anhang 11: Items für das OLS-Modell des Methodenvergleichs bei Unit Nonresponse . . . . . 206

Anhang 12: Items für das Logitmodell des Methodenvergleichs bei Unit Nonresponse . . . . . 206

Anhang 13: Sonstige visuelle Aufbereitung für das OLS-Modell des Methodenvergleichs bei Unit Nonresponse . . . . . 207

Anhang 14: Sonstige visuelle Aufbereitung für das Logitmodell des Methodenvergleichs bei Unit Nonresponse . . . . . 209

# B Abbildungsverzeichnis

Abbildung 1: Kategorisierung im deutschen und englischen Sprachgebrauch . . . . . 4

Abbildung 2: Survey Lifecycle nach Groves et al. (2004) . . . . . 7

Abbildung 3: Übersicht Verfahren zum Umgang mit fehlenden Werte . . . . . 12

Abbildung 4: Survey Lifecycle und Item Nonresponse . . . . . 16

Abbildung 5: Schema Stimulus-Person-Reaktionsmodell . . . . . 19

Abbildung 6: Schema kognitiver Status und Item Nonresponse . . . . . 22

Abbildung 7: Antwortprozess nach Beatty und Herrmann (2002) . . . . . 24

Abbildung 8: Antwortprozess, Ausfallgrund und Ausfallmechanismus . . . . . 26

Abbildung 9: Erzeugung des Vektors mit Item Nonreponse . . . . . 31

Abbildung 10: Häufigkeitsverteilung und Boxplot für die Anzahl der Item Nonresponse . . . . . 32

Abbildung 11: Label von Stata Press: Ausschlag eines Pferdes . . . . . 33

Abbildung 12: Familie der Poissonmodelle nach Czado et al. (2007) . . . . . 37

Abbildung 13: Ausfallmuster von Beispiel 1 . . . . . 50

Abbildung 14: Ausfallmuster von Beispiel 2 . . . . . 54

Abbildung 15: Ausfallmuster von Beispiel 3 . . . . . 56

Abbildung 16: Verfahren zum Vergleich von Item Nonresponse-Korrekturmethoden . . . . . 59

Abbildung 17: Differenz der Konfidenzintervalllängen bei Beispiel 1 . . . . . 69

Abbildung 18: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Fremd im eigenen Land* . . . . . 70

Abbildung 19: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Fremd im eigenen Land* . . . . . 71

Abbildung 20: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Freie Meinungsäußerung* . . . . . 72

Abbildung 21: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Freie Meinungsäußerung* . . . . . 72

Abbildung 22: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Juden zuviel Einfluss* . . . . . 73

Abbildung 23: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Juden zuviel Einfluss* . . . . . 74

Abbildung 24: Differenz der Konfidenzintervalllängen bei Beispiel 2 . . . . . 78

Abbildung 25: CC Histogramm und Boxplot des Parameters zur Variablen *Alter*$^2$ . . . . . 79

Abbildung 26: MI Histogramm und Boxplot des Parameters zur Variablen *Alter*$^2$ . . . . . 79

Abbildung 27: CC Histogramm und Boxplot des Parameters zur Variablen *Arbeitslos* . . . . . 80

Abbildung 28: MI Histogramm und Boxplot des Parameters zur Variablen *Arbeitslos* . . . . . 80

Abbildung 29: CC Histogramm und Boxplot des Parameters zur Variablen *Vater niedrige Bildung* . . . . . 81

Abbildung 30: MI Histogramm und Boxplot des Parameters zur Variablen *Vater niedrige Bildung* . . . . . 82

Abbildung 31: CC Histogramm und Boxplot des Parameters zur Variablen *Vater Arbeiter* . . . . . 83

Abbildung 32: MI Histogramm und Boxplot des Parameters zur Variablen *Vater Arbeiter* . . . . . 83

Abbildung 33: CC Histogramm und Boxplot des Parameters zur Variablen *Ostdeutsch* . . . . . 84

Abbildung 34: MI Histogramm und Boxplot des Parameters zur Variablen *Ostdeutsch* . . . . . 84

Abbildung 35: CC Histogramm und Boxplot des Parameters zur Variablen *Politische Orientierung* . . . . . 85

Abbildung 36: MI Histogramm und Boxplot des Parameters zur Variablen *Politische Orientierung* . . . . . 86

Abbildung 37: Differenz der Konfidenzintervalllängen bei Beispiel 3 . . . . . 92

Abbildung 38: CC Histogramm und Boxplot des Achsenabschnitts . . . . . 93

Abbildung 39: MI Histogramm und Boxplot des Achsenabschnitts . . . . . 93

Abbildung 40: CC Histogramm und Boxplot des Parameters zur Variablen *Alter* . . . . . 94

Abbildung 41: MI Histogramm und Boxplot des Parameters zur Variablen *Alter* . . . . . 95

Abbildung 42: CC Histogramm und Boxplot des Parameters zur Variablen *Einkommen* . . . . . 96

Abbildung 43: MI Histogramm und Boxplot des Parameters zur Variablen *Einkommen* . . . . . 96

Abbildung 44: CC Histogramm und Boxplot des Parameters zur Variablen *Demokratiezufriedenheit* . . . . . 97

Abbildung 45: MI Histogramm und Boxplot des Parameters zur Variablen *Demokratiezufriedenheit* . . . . . 97

Abbildung 46: Survey Lifecycle und Unit Nonreponse . . . . . 106

Abbildung 47: Übersicht über Gründe von Unit Nonresponse nach Schnell (1997) mit Zuordnung zu Elementen des Survey Lifecycle . . . . . 108

Abbildung 48: Einordnung von Ausfallgründen . . . . . 109

Abbildung 49: Schema Kontaktwahrscheinlichkeit nach Groves und Couper (1998) . . . . . 112

Abbildung 50: Modell zur Verweigerung nach Groves und Couper (1998) . . . . . 117

Abbildung 51: Schaubild zur Leverage-Salience-Theorie nach Groves et al. (1998) . . . . . 119

Abbildung 52: Wertrationales Erklärungsmodell der Verweigerung nach Engel et al. (2004) . . . . . 121

Abbildung 53: Veränderung der Ausschöpfungsquote des ALLBUS 1980 - 2008 . . . . . 127

Abbildung 54: Ausfallgründe im ALLBUS 2008 . . . . . 137

Abbildung 55: Schema Verfahren zum Vergleich von Unit Nonresponse-Korrekturmethoden . . . . . 148

Abbildung 56: MI Histogramm des Anteilswerts zur Variablen *Wahl der SPD* . . . . . 159

Abbildung 57: Gewichtung 1 und 2: Histogramm des Anteilswerts zur Variablen *Wahl der SPD* . . . . . 159

Abbildung 58: Veränderung der Standardabweichung der geschätzten Werte des Anteilswerts unter MI mit sinkendem Ausfall . . . . . 161

Abbildung 59: MI Histogramm des Mittelwerts zur Variablen *Linksrechtsselbsteinstufung* . . . . . 161

Abbildung 60: Veränderung der Standardabweichung der geschätzten Werte des Mittelwerts unter MI mit sinkendem Ausfall . . . . . 162

Abbildung 61: Gewichtung 1 und 2: Histogramm des Mittelwerts zur Variablen *Linksrechtsselbsteinstufung* . . . . . 162

Abbildung 62: MI Histogramm des Achsenabschnitts der OLS-Regression . . . . . 164

Abbildung 63: Gewichtung 1 und 2: Histogramm des Achsenabschnitts der OLS-Regression . . . . . 165

Abbildung 64: Veränderung der Standardabweichung der geschätzten Werte des Achsenabschnitts unter MI mit sinkendem Ausfall . . . . . 166

Abbildung 65: MI Histogramm des Parameters der Variablen *Linksrechtsselbsteinstufung* . . . . . 167

Abbildung 66: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Linksrechtsselbsteinstufung* . . . . . 168

Abbildung 67: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Linksrechtsselbsteinstufung* unter MI mit sinkendem Ausfall . . . . . 168

Abbildung 68: MI Histogramm des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen* . . . . . 169

Abbildung 69: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen* unter MI mit sinkendem Ausfall . . . . . 170

Abbildung 70: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen* . . . . . 171

Abbildung 71: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen* unter MI mit sinkendem Ausfall . . . . . 173

Abbildung 72: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Wahrscheinlichkeit SPD zu wählen* unter MI mit sinkendem Ausfall . . . . . 173

Abbildung 73: MI Histogramm des Achsenabschnitts des Logitmodells . . . . . 174

Abbildung 74: Veränderung der Standardabweichung der geschätzten Werte des Achsenabschnitts unter MI mit sinkendem Ausfall . . . . . 175

Abbildung 75: Gewichtung 1 und 2: Histogramm des Achsenabschnitts des Logitmodells . . . . . 175

Abbildung 76: MI Histogramm des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen* . . . . . 177

Abbildung 77: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen* . . . . . 177

# C Tabellenverzeichnis

Tabelle 1: Verwendete ALLBUS-Erhebungen . . . . . 29

Tabelle 2: Operationalisierung und Codierung der unabhängigen Variablen . . . . . 43

Tabelle 3: Erklärungsmodelle für Item Nonresponse mit OLS-Regression,
Poissonregression, ZIP-Regression, ZIGP-Regression, negative Binomialregression,
Zero-Inflated negative Binomialregression . . . . . 44-45

Tabelle 4: Ausfall der Variablen in Beispiel 1 . . . . . 49

Tabelle 5: Verteilung des Anomieindex nach unterschiedlicher Konstruktion . . . . . 50

Tabelle 6: Replizierte Werte für Beispiel 1 . . . . . 51

Tabelle 7: Ausfall der Variablen in Beispiel 2 . . . . . 53

Tabelle 8: Replizierte Werte für Beispiel 2 . . . . . 54

Tabelle 9: Ausfall der Variablen in Beispiel 3 . . . . . 56

Tabelle 10: Replizierte Werte für Beispiel 3 (logistische Regression) . . . . . 57

Tabelle 11: Replizierte Werte für Beispiel 3
(Wahrscheinlichkeit der Wahlteilnahme) . . . . . 58

Tabelle 12: Übersicht vollständige Variablen . . . . . 60

Tabelle 13: Bildung von Typen für Beispiel 1 . . . . . 62

Tabelle 14: Bildung von Typen für Beispiel 2 . . . . . 62

Tabelle 15: Bildung von Typen für Beispiel 3 . . . . . 63

Tabelle 16: Übersicht über die Verwendung von Multipler Imputation in Veröffentlichungen mit den ALLBUS-Erhebungen . . . . . 65

Tabelle 17: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 1 . . . . . 68

Tabelle 18: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI . . . . . 74

Tabelle 19: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI . . . . . 75

Tabelle 20: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 2 . . . . . 76

Tabelle 21: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI . . . . . 87

Tabelle 22: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI . . . . . 87

Tabelle 23: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3 und Standardabweichung der Abweichung bei CC und MI . . . . . 88

Tabelle 24: MSE für ausgewählte Parameter . . . . . 88

Tabelle 25: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 3 . . . . . 90

Tabelle 26: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI (logistische Regression) . . . . . 98

Tabelle 27: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2
und Standardabweichung der Abweichung bei CC und MI
(logistische Regression) . . . . . 99

Tabelle 28: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3
und Standardabweichung der Abweichung bei CC und MI
(logistische Regression) . . . . . 99

Tabelle 29: MSE für ausgewählte Parameter . . . . . 100

Tabelle 30: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1
und Standardabweichung der Abweichung bei CC und MI
(Eintrittswahrscheinlichkeit) . . . . . 100

Tabelle 31: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2
und Standardabweichung der Abweichung bei CC und MI
(Eintrittswahrscheinlichkeit) . . . . . 101

Tabelle 32: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3
und Standardabweichung der Abweichung bei CC und MI
(Eintrittswahrscheinlichkeit) . . . . . 101

Tabelle 33: Berechnung der Ausschöpfungsquote bei ALLBUS-Erhebungen . . . . . 125

Tabelle 34: Ausschöpfungsquote, Feldzeit und Erhebungsinstitut
nach Landesteil . . . . . 128

Tabelle 35: Einflüsse auf die Entwicklung der Ausschöpfungsquote
(Beta-Koeffizienten) . . . . . 130

Tabelle 36: Veränderung des Anteils der Verweigerer, Nichterreichten
und Befragungsunfähigen nach Landesteilen . . . . . 131

Tabelle 37: Einflüsse auf die Entwicklung des Verweigerungsanteils
(Beta-Koeffizienten) . . . . . 132

Tabelle 38: Einflüsse auf die Entwicklung des Nichterreichbarenanteils (Beta-Koeffizienten) . . . . . 133

Tabelle 39: Einflüsse auf die Entwicklung des Befragungsunfähigenanteils (Beta-Koeffizienten) . . . . . 134

Tabelle 40: Berechnung der Anzahl auswertbarer Interviews für den ALLBUS 2008 . . . . . 138

Tabelle 41: Beschreibung der Variablen für das Erklärungsmodell für Unit Nonresponse . . . . . 140

Tabelle 42: Ergebnisse der Logitmodelle nach Ausfallgründen . . . . . 143

Tabelle 43: Bildung von Typen für die OLS-Regression . . . . . 152

Tabelle 44: Bildung von Typen für das Logitmodell . . . . . 153

Tabelle 45: Werte der Typen für den Anteilswert, Mittelwert, Parameter der OLS-Regression und des Logitmodells . . . . . 153

# 1 Fehlende Werte im Kontext sozialwissenschaftlicher Erhebungen

„Nonresponse is an epistemological problem with teeth. Nonresponse jeopardizes both the external and internal validity of our work.“ (Brehm 1993, S.16)

Man könnte diesem Zitat auch hinzufügen, dass Nonresponse ein Problem darstellt, dessen Zähne häufig nicht gezogen werden – in der Hoffnung, dass es nicht beißt. Tatsächlich kennen vermutlich die allermeisten empirisch Forschenden fehlende Werte. Empirisch arbeitende Sozial- und Wirtschaftswissenschaftler greifen in den meisten Fällen auf Daten zurück, die sie nicht selbst erhoben haben. Spätestens bei der Datenanalyse wird man mit dem Problem konfrontiert und entscheidet sich oft, das Problem zu ignorieren anstatt ein Korrekturverfahren, das die Statistik bereithält, anzuwenden. Die Möglichkeit, selbst eine Erhebung zu organisieren, haben die wenigsten Mitglieder der Forschungsgemeinschaft. Durch die Vielzahl von Vermeidungsstrategien, die die Methodenforschung vorschlägt, wird man hierbei fast eingeschüchtert. Die starke arbeitstechnische Trennung zwischen denjenigen, die Erhebungen organisieren und durchführen, denjenigen, die Korrekturmethoden testen und weiterentwickeln und schließlich denjenigen, die Erhebungsdaten zur Analyse von sozial- und wirtschaftswissenschaftlichen Datensätzen heranziehen, scheint das zarteste Fleisch für die Zähne des Problems Nonresponse sein. Die allzu starke Trennung der Arbeitsbereiche erscheint konktrapoduktiv für das Verständis fehlender Werte und deren nachhaltige Behandlung.In den sozial- und wirtschaftswissenschaftlichen Erhebungen existieren zudem noch einige Besonderheiten; den Kern einer Erhebung in den Wirtschafts- und Sozialwissenschaften bildet fast immer eine Art von sozialer Interaktion. In den jeweiligen Theoriekapiteln wird erklärt, welche Herausforderungen aus dem Problem der fehlenden Werte erwachsen. Da die Aussagen, die aus Erhebungen gewonnen werden, mittlerweile auch noch eine sehr hohe mediale Bedeutung bekommen können, ist ein umfassenes Verständnis der Missing Data Problematik zwingend. So erstreckt sich die Bedeutung des Surveys im öffentlichen Leben, also nicht nur auf die wissenschaftliche Sphäre verengt, nach Brehm zunächst auf die reine Gewinnung von Informationen, zur Unterstützung des wissenschaftlichen wie öffentlichen Diskurses sowie im konkreteren Fall auf die Planung von Projekten aller Art (Brehm 1993, S.16). Eine fehlerhafte Umfrage als Diskursinstrument und Informationsquelle macht sich daher umso mehr angreifbar, macht schließlich das Instrument an sich wertlos (Brehm 1993, S.4f). Geboten erscheint deshalb eine Erforschung der „Krankheit“ fehlender Werte im Kontext des gesamten Umfrageprozesses (Groves et al. 2004), mit dem Ziel, die teilweise schmalen Brücken zwischen Methodikern, Statistikern und letztlich den Datennutzern zu verbreitern.

Fehlende Werte lassen sich auf verschiedene Arten differenzieren. Diese Arbeit folgt im Aufbau der in der Literatur am häufigsten anzutreffenden Unterscheidung in Item und Unit Nonresponse. Zunächst sollen einige Definitionen zur den wichtigsten Begriffen der Arbeit vorgenommen und in eine Übersicht gebracht werden (2). Danach folgen parallel aufgebaut die Kapitel, die sich mit Item Nonresponse (3 und 4) und Unit Nonresponse (5 und 6) beschäftigen, ehe ein Fazit (7) die wichtigsten Ergebnisse zusammenfasst. Kapitel 3 und 4 bzw. 5 und 6 sind insofern parallel aufgebaut, als jeweils der theoretische Hintergrund und ein Überblick über den Forschungsstand gegeben werden. Darüber hinaus sollen anhand von ALLBUS Datensätzen beispielhaft fehlende Werte analysiert werden. Von zentraler Bedeutung ist jeweils ein Korrekturmethodenvergleich für ausgewählte Parameter.

# 2 Übersicht und Definitionen

Bevor eine Analyse und später die Korrektur fehlender Werte in den Fokus genommen werden kann, erscheint eine Definition von Schlüsselbegriffen sowie eine Übersicht zur Einordnung sinnvoll. Zuerst wird eine Übersicht der Synonyme für fehlende Werte und ihrer Verwendung in der Literatur (2.1) gegeben, ehe der Gegenstand definiert wird (2.2). Anschließend wird der Gegenstand in den Kontext des Survey Lifecycle und aller Erhebungsfehler eingeordnet (2.3). Schließlich behandelt der letzte Abschnitt des Kapitels kurz den Umgang mit fehlenden Werten und erklärt die Vor- und Nachteile des jeweiligen Vorgehens (2.4). Damit soll auch die Auswahl der zu vergleichenden Methoden vorbereitet werden.

## 2.1 Begriff

Grundsätzlich lässt sich von „Datenausfällen“ oder „fehlenden Werten“ (engl. Missing Data) sprechen, da diese Begriffe implizieren, dass etwas nicht vorhanden ist, das da sein müsste und dessen Existenz man angestrebt hat (Daten, Werte). „Datenausfall“ liegt vor, wenn beispielsweise ein Virus auf dem Rechner des Umfrageinstituts Informationen des Datensatzes löscht, aber auch wenn ein Befragter im Interview nicht antwortet. Fehlende Werte sind damit eigentlich der Überbegriff zu „Nonresponse“, der häufig im Fokus methodischen Erkenntnisinteresses steht (Groves et al. 2004, S.59), jedoch bereits im Namen den Ausfallgrund mit sich trägt und somit eher verengend wirkt. Trotzdem wird der Begriff Nonresponse gerade in der englischsprachigen Literatur überaus häufig gebraucht.
Nonresponse wird danach üblicherweise in Unit und Item Nonresponse unterschieden, wobei der Ausfall der Einheit für die Befragung als Unit, die Nichtbeantwortung einer Frage als Item Nonresponse bezeichnet wird. Bei spezifischen Erhebungskonzepten wie der Panelerhebung wird zudem der Begriff Wave Nonresponse oder Panelmortalität verwendet. Fragebogeninhärent ist auch der Begriff Missing by Design, wenn in der Folge von Filterfragen bewusst keine weiter Befragung stattfindet (Göthlich 2007). Datenausfallgründe liegen auch der Einteilung von McKnight et al. zugrunde. Die Klassifikation unterscheidet zwischen fehlenden Werten, die durch die Teilnehmer der Erhebung verursacht werden, aufgrund des Erhebungsdesigns oder aufgrund der Interaktion von Teilnehmern und Design entstanden sind (McKnight et al. 2007, S.5).


Abbildung 1: Kategorisierung im deutschen und englischen Sprachgebrauch

Der Begriff Nonresponse wurde bereits genannt. Daneben findet man auch den eher neutralen Begriff „incomplete data“ in der englischen Fachliteratur. Antiquiert erscheint „noncooperater“ als Synonym für Unit Nonresponse. Selten treten die Bezeichnungen „sampling mortality“, „incomplete samples“ oder „noninterview“ auf (Lessler und Kalsbeck 1992, S.107). Hinter vielen Bezeichnungen, die gelegentlich nur ein Autor verwendet, verbergen sich bestimmte Vorstellungen bzw. Konzepte zur Teilnahmebereitschaft und dem Antwortverhalten des Befragten.[1] In der deutschen Sprache dominieren die Begrifflichkeiten *neutrale* und *systematische* Ausfälle.[2] Ihre Verwendung impliziert bereits einen unterstellen Ausfallmechanismus, gleichzeitig bedürfen sie einer genaueren Definition, die sehr unterschiedlich ausfallen kann (Schnell 1997, S.23).[3] Die verwendeten Begriffe für das Phänomen der fehlenden Werte und dessen Spielarten sind zahlreich, die Definitionen der Begriffe sind geradezu babylonisch verwirrend.

[1] Häufig sind diese Begriffe Negationen der Einheiten, die erhoben werden: „Teilnehmer“(„Participant“), „cooperater“ oder eben auch „respondent“, vgl. Lessler und Kalsbeck (1992), S.107.

[2] Dies ist auch die offizielle Unterscheidung in der ALLBUS Dokumentation.

[3] Statistiker systematisieren Datenausfälle teilweise nur über deren Ignorierbarkeit, vgl. Spieß (2008), S.3ff.

## 2.2 Definitionen

Sehr selten wird in den vielen Veröffentlichungen zum Thema fehlender Werte überhaupt auf die Definition des Begriffs eingegangen. So findet sich bei Rässler beispielsweise einleitend die Erklärung, dass „unter Datenausfällen in Umfragen [...] also im folgenden das Phänomen der Nichtbeantwortung einzelner Fragen von Objekten verstanden [wird]“ (Rässler 2000, S.65).
Die Tatsache, dass viele Begriffe für das Phänomen existieren, zwingt letztlich zu Entscheidungen und sollte dann konsistent durchgehalten werden. So einfach aber beispielsweise die Unterteilung in Item und Unit Nonresponse ist, so umstritten und uneindeutig gestaltet sich die scharfe Abgrenzung des Gegenstands. Selbst bei gleicher Begrifflichkeit muss zudem nicht dasselbe gemeint sein. So definieren beispielsweise Kendall und Buckland (1960) allein Menschen (natürliche Personen) als Unit Nonresponse, das amerikanische Zensusbüro darüber hinaus alles, was als Einheit (Unit) aufgefasst werden kann. Auch beim Item Nonresponse lassen sich engere und ausgreifendere Definitionen finden. So legte das US-Censusbüro fest, dass Item Nonresponse „generally attributed to failure to obtain a response to a particular item“ sind (auch Kalton 1983). Es lässt sich aber auch die Position einnehmen, dass inkonsistente Antworten zu Item Nonresponse zählen. Ein weiterer Spielraum ergibt sich bei der Frage, ob Personen, die in der Stichprobe sind, aber nicht zur Zielgesamtheit gehören, als Nonrespondenten gezählt werden. Dies wird sehr unterschiedlich ausgelegt. Das Zensusbüro bejaht dies und zählt „ineligibles“ zu Unit Nonresponse. Einhelliger ist die Literatur in der Frage, ob bei „Undercoverage“ ein Fall von Nonresponse vorliegt. Kish (1965) und Kalton (1983) sehen in Undercoverage und Nonresponse zwei Arten von fehlenden Werten als Nonobservations. Auch Cochran sieht dies so: Noncoverage sei ein Typus von Nonresponse (Cochran 1977).[4]
Weniger kontrovers wird im wissenschaftlichen Diskurs um die Definition von Item Nonresponse gerungen. Strittig bleibt aber Grundlegendes, wenn es um konkrete Ausformungen geht: wie sind einzelne Kategorien wie „weiß nicht“ zu bewerten (Rubin et al. 1995, S.822ff; Schumann und Presser 1981, S.113f).
Formal gesehen sind es leere Zellen in einer Datenmatrix mit den Merkmalsträgern als Zeilen und den Items als Spalten, deren Umfang wir unter Umständen nicht genau kennen (im Falle von Unit Nonresponse). Betrachtet man nur die Datenmatrix der Nettostichprobe mit Item Nonresponse, dann lassen sich einige spezifische Ausfallmuster unterscheiden. Neben dem mono- und multivariaten Ausfall existieren monotone, disjunkte und allgemeine Ausfallmuster in den Daten (Little und Rubin 2002, S.4f).
Eine grundlegende – und wie wir sehen werden fruchtbare – Differenzierung der fehlenden Daten betrifft die Unterscheidung nach dem Ausfallmechanismus (oder Fehlendmechanismus) (Rubin 1976, 1987; Little und Rubin 2002, S.11f). Wenn der Ausfall rein zufälligen entsteht, wird dies als als Missing Completely At Random oder *MCAR* bezeichnet. Zweitens nennt man Missing At Random oder *MAR* denjenigen Ausfall, der durch andere (erhobene) Merkmale erklärt werden kann. Als letztes können fehlende Werte direkt von den nicht gegebenen Antworten abhängen. Dies wird als Missing Not At Random (kurz *NMAR*) oder nicht ignorierbar bezeichnet. Eine Anzahl von Korrekturmethoden setzt explizit (beispielsweise MAR für die Multiple Imputation) oder implizit (beispielsweise MCAR Complete Cases und Available Cases) einen bestimmten Ausfallmechanis-

[4] Lessler und Kalsbeck beschreiben die unterschiedliche Verwendung und die Diskussion um die Begriffe und ihre Abgenzung detailliert, vgl. Lessler und Kalsbeck (1992), S.107.

mus voraus. Die konkreten Ursachen der leeren Zellen können jedoch sehr unterschiedlich sein, gerade in umfangreichen und komplexen Erhebungen. Damit wird deutlich, dass nicht nur die Fragen nach dem Gegenstand „fehlende Werte“, sondern auch nach dessen Abgrenzung aus verschiedenen Blickwinkeln betrachtet werden und zu unterschiedlichen Konzepten führen können. Während der Methodiker eher auf die genaue Definition in der Erhebung abzielt, differenziert der Statistiker bereits in Hinblick auf die mögliche Korrekturmethoden nach Ausfallmustern und Ausfallmechanismus. Anders als frühere Untersuchungen soll hier das Phänomen Datenausfall deshalb möglichst im Rahmen des gesamten Erhebungsprozess betrachtet werden. Diese umfassende Perspektive erscheint angesichts der allgegenwärtigen Existenz und der Vielzahl der Quellen für das Problem der fehlenden Werte dringend geboten.

## 2.3 Kontext: Erhebung und Fehler

Ganz allgemein spricht man von einem Nonresponsefehler in Erhebungen. Doch was ist eine Erhebung und wo liegen generell Fehlerquellen bei der Durchführung einer Erhebung? Als Survey/Erhebung wird eine Forschungsstrategie bezeichnet, bei der in der Regel quantitative Informationen über eine Grundgesamtheit gesammelt werden (Scheuren 2004, S.9). Aus Kostengründen wird hierzu nicht die Grundgesamheit vollständig erhoben, sondern eine Stichprobe gezogen. Die Stichprobe befähigt dann zu inferenzstatistischen Aussagen mit Informationsgehalt über die Grundgesamtheit. Informationen sind Aussagen über die Realität. Realität im Kontext der Sozialwissenschaft meint wiederum Verhalten und Einstellungen, die das Verhalten beeinflussen.
Um einen Überblick und eine Systematisierung bezüglich einzelner Aspekte einer Erhebung zu erhalten, wird im Folgenden der sogenannte Survey Lifecycle verwendet. Auch für die weiteren Kapitel wird immer wieder auf den Survey Lifecycle zurückgegriffen, um abzuschätzen, wo welche Art von fehlenden Werten entsteht und warum dies so ist. Nach Groves et al. (2004) lässt sich für diesen Lebenszyklus einer Erhebung eine Prozess- und eine Designperspektive einnehmen, die im Schema als innerer (Prozess) und äußerer Kreis (Design) visualisiert sind (siehe Abbildung 2). Beide Kreise beginnen mit der Definition der Fragestellung, die aus dem Forschungsgegenstand entwickelt wird. Die rechte Seite symbolisiert die Repräsentationsfunktion der Erhebung, die linke Seite hingegen bildet die Messfunktion der Erhebung ab.
Wissenschaftler gehen bei der Fragestellung bereits mit fertigen oder zu diesem Zweck entwickelten, theoretischen Konstrukten in den Survey Lifecycle.[5] Das Konstrukt beeinflusst direkt die Messmethodik. Während von der anderen Seite die Wahl des Sampleframes, determiniert durch die Zielgesamtheit, die wiederum mit der Fragestellung zusammenhängt, in Wechselwirkung mit der Wahl des Erhebungsmodus tritt, beeinflusst der Sampleframe im weiteren die Stichprobenziehung.[6] Auf der anderen Seite erfährt die Messmethodik in der Konstruktion und (Pre-)Testung des Fragebogens eine prozessuale Konkretisierung. In der unteren Kreishälfte steht der Befragte mit

[5] Zum Begriff des Konstruktes zur Messung einer wissenschaftliche Fragestellung, vgl. Fowler et al. (2008), S.136).

[6] Sampleframe besteht aus „lists or procedures intended to identify all elements of a target population“ vgl. Groves et al. (2004), S.68

seiner Antwort dem Interviewer als Fragensteller gegenüber. Die Antwort des Befragten ist die Manifestation der Messfunktion. Letzeres als Element der Repräsentation der Gesamtheit, über die eine Aussage im Sinne der Fragestellung und des Untersuchungsgegenstandes gemacht werden soll. Der Befragte ist so gesehen das Produkt der realisierten Stichprobe, das gezogene und angetroffene und auskunftswillige bzw. kooperationsbereite Element aus der Auswahlgesamtheit.

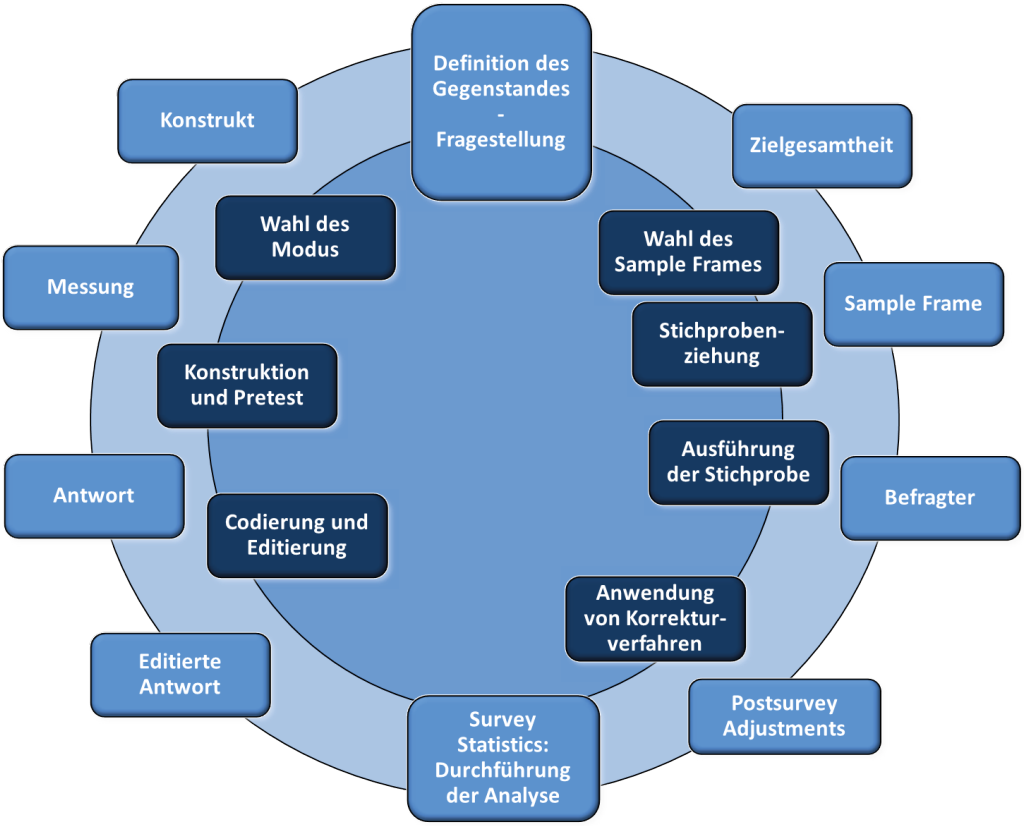



Abbildung 2: Survey Lifecycle nach Groves et al. (2004)

Der eigentliche Ort der Verschränkung von Messung und Repräsentation in der Manifestation des Interviews als Interaktion von Interviewer und Befragtem ist hier nicht explizit aufgeführt. In Kapitel 3 wird beschrieben, wie das Interview als Kristallisationspunkt von Item Nonresponse wirken kann, in Kapitel 5 gilt es die Kontaktaufnahme zum Befragten als ersten Abschnitt des Interviews auf den Zusammenhang mit Unit Nonresponse zu prüfen. Zurück zur linken Seite: nach der Äußerung der Antwort erfolgt durch Codierung und Editierung die Transformation der ursprünglichen Antwort zur editierten Antwort. Auf der rechten Kreisseite erscheint nun die nicht obligatorische, aber übliche Anwendung diverser Korrekturverfahren, die teilweise in direktem Zusammenhang mit der Editierung stehen. Das Gros der Datennutzer bekommt bis zu diesem Zeitpunkt von der Erhebung kaum etwas mit. Sogar die Postsurvey Adjustments werden häufig bereits vorgenommen, ehe der Datennutzer tatsächlich seine Analyse durchführt.

Jeder Fehler, der bei einer Erhebung auftritt, kann somit zu einem verzerrten Blick auf die Realität und damit zu falschen Schlussfolgerungen führen.[7] Der Nonresponsefehler hat sowohl auf die Mess- als auch auf die Repräsentationsfunktion Auswirkungen. Dieser Fehler ist allerdings nicht die einzige Quelle für Verzerrungen. In der Regel unterscheidet man zwischen Coveragefehler, Messfehler, Nonresponsefehler und Samplingfehler (Groves et al. 2004, S.48ff),[8] die den totalen Survey Error konstituieren (Lohr 2008, S.98), wobei der Nonresponsefehler für eine Variable $X$ definiert wird als:

$$\bar{x}_r - \bar{x}_s = \frac{m_s}{n_s}(\bar{x}_r - \bar{x}_m)$$

mit

$\bar{x}_s =$ Mittelwert der gesamten Stichprobe,
$\bar{x}_r =$ Mittelwert der Respondenten,
$\bar{x}_m =$ Mittelwert der Nichtrespondenten,
$m_s =$ Anzahl der Nichtrespondenten in der $s-$ten Stichprobe,
$n_s =$ Gesamtgröße der $s-$ten Stichprobe.

Allein mit dieser formalen Definition wird deutlich, dass nicht allein das Verhältnis von Respondenten und Nichtrespondenten, sondern die tatsächliche Verschiedenheit des Antwortverhaltens den Umfang der Nonresponseverzerrung beeinflusst. Es fällt zudem auf, dass die formale Definition des Nonresponsefehlers der des Coveragefehlers ähnelt (Groves et al. 2004, S.55):

$$\overline{X}_C - \overline{X} = \frac{U}{N}(\overline{X}_C - \overline{X}_U)$$

mit

$\overline{X} =$ Mittelwert der gesamten Zielgesamtheit,
$\overline{X}_C =$ Mittelwert der Gesamtheit im Sampleframe,
$\overline{X}_U =$ Mittelwert der Zielgesamtheit, die nicht im Sampleframe enthalten ist,

[7] Ein kleines Beispiel: In einer Gemeinde mit 14.000 Einwohnern wohnen 8.000 Personen mit normaler oder überdurchschnittlicher Begabung und 6.000 mit unterdurchschnittlicher Begabung – diese Verteilung ist natürlich nicht bekannt. Es soll der Anteil der unterdurchschnittliche Begabten Anhand einer Stichprobe geschätzt werden, in der jeder 20. Einwohner erhoben wird. So müsste die Stichprobe aus 400 normal oder überdurchschnittlich begabten Menschen und 300 unterdurchschnittlich begabten Menschen bestehen. Aus verschiedenen Gründen nehmen allerdings nur 500 Personen insgesamt an der Erhebung teil; was nicht bekannt ist, ist die Tatsache, dass sich die Höhe der Ausfälle nach Begabung verschieden verhält: während 80 % der ausgewälten normal und überdurchschnittlich Begabten an der Erhebung teilnehmen, sind es bei den unterdurchschnittlich Begabten nur 60 %. Dies würde uns zu einer Schätzung des Anteils von unterdurschnittlich Begabten von 36 % ($\frac{180}{500}$) anstatt von 43 % ($\frac{6.000}{14.000}$) führen. Dieses denkbar einfache Beispiel zeigt ebenso plausibel den Zusammenhang zwischen einem interessierenden Merkmal (Intelligenz) und dem Nonresponsefehler, vgl. Lynn (2008), S.36.

[8] Hinzu kommen Validitätsprobleme des theoretischen Konstrukts, Processing Error und Adjustment Error, vgl. Groves et al. (2004), S.48.

$N =$ Gesamtzahl der Merkmalsträger in der Zielgesamtheit,
$C =$ vom Sampleframe abgedeckte (covered) Merkmalsträger,
$U =$ nicht im Sampleframe enthaltene (not covered) Merkmalsträger.

In der Tat können beide formal getrennten Fehler in einer Erhebung zum selben Ergebnis führen: nämlich leere Zellen in einer Matrix (Lohr 2008, S.98). Dies zeigt auch, wie wenig hilfreich alleinige Konzentration auf die formale Definition sein kann, da die Identifizierung der Quelle fehlender Werte im Nachhinein nicht mehr möglich ist. Die formale Definition des Nonresponsefehlers bildet dennoch die Grundlage für weitere Überlegungen im Kontext der Diskussion um Erhebungsqualität und dessen Kriterien (De Leeuw et al. 2008).
Durch den Rahmen, den der Survey Lifecycle aufspannt, kann teilweise über solche formalen Definitionen hinaus die Komplexität einer Erhebung im Ganzen nachvollziehbar gemacht werden. Nachvollziehbar werden auch die Folgen von Designentscheidungen (Design) und deren Ausführung (Prozess) auf Item und Unit Nonresponse. So werden bestimmte Faktoren im Survey Lifecycle das Auftreten fehlender Werte begünstigen oder vermindern.

## 2.4 Umgang mit fehlenden Werten

Der Umgang mit fehlenden Werten läuft auf zwei Strategien hinaus. Die im Survey Lifecycle vorgelagerte Strategie sieht präventiv die Vermeidung als wichtigste Maßnahme. Die andere Strategie ist im Lifecycle der eigentlichen Datenerhebung nachgelagert; hierbei wird versucht, die durch die fehlenden Werte verursachte Verzerrung zu korrgieren (Postsurvey Adjustments) (De Leeuw et al. 2008, S.10).
Groves et al. nennen eine Reihe von „klassischen" Vermeidungsmaßnahmen, die sich auf Entscheidungen und Ausführungen im Survey Lifecycle beziehen. Sie setzen zum einen vor der Interviewdurchführung, zum anderen in der Feldphase an (Groves et al. 2004). Zunächst beziehen sich die Vermeidungsmaßnahmen auf die Verbesserung der Kontaktmöglichkeiten. Hierzu gehört beispielsweise die Verlängerung der Datenerhebung, die Optimierung der Kontaktanzahl und der Kontaktzeitpunkte, ein höherer Workload für die Interviewer oder die Verbesserung ihrer Fähigkeiten (Beobachtungsgabe etc.). Bei der Kontaktaufnahme ist auf das Interviewverhalten zu achten. Daneben gilt das Image des Auftragsgebers als teilnahmefördernd oder -verhindernd. In der Literatur lassen sich auch positive Effekte durch Ankündigungen der Befragung belegen. Ein weites Feld nimmt das Thema „incentives", also Geschenke und Anreize aller Art, in der Methodenforschung ein (Singer 2002; Singer et al. 1998). Generell gilt es bei der Kontaktaufnahme selbst Hürden abzubauen. So können die Rollen des Interviewten durchaus unterschiedlich ausgestaltet werden. Sehr komplex ist der Versuch eines optimalen Matchings zwischen Interviewer- und Respondententypen.
Auch bei zunächst nicht erfolgreichen Kontaktversuchen besteht die Möglichkeit nachzufassen, durch Schreiben zu überreden oder den Interviewer zu wechseln. Die Forschung im Bereich der „Konvertiten" nimmt mittlerweile breiten Raum ein (Stoop et al. 2010, S.161ff).[9] Mit den (nicht völlig erschöpfend) genannten Maßnahmen soll primär dem Auftreten von Unit Nonresponse entgegengesteuert werden. Eine wichtige und bei weitem noch nicht geklärte Frage betrifft die Auswirkungen der Vermeidungsstrategien für Unit Nonresponse: Verändert sich das Antwortverhalten dadurch oder wird der Nonresponsefehler in einer vielleicht verschlimmernden Weise verzerrt? (Groves et al. 2004, S.195).
Die Vermeidungskonzepte von Item Nonresponse beinhaltet dagegen eine breite Palette von Theorien bezüglich Befragten- und Antwortverhalten. Dieses interagiert mit Details wie der Fragebogenkonstruktion oder auch mit der Herangehensweise bei sensiblen Erkenntnisinteressen. Dillman betont dabei, dass die Vermeidung von Item Nonresponse die Hauptaufgabe der Fragebogenkonstruktion sein müsse (Dillman 2008, S.163).

Die Probleme der Korrekturverfahren, die nach der Feldphase ansetzen, werden an einer anderen Stelle sichtbar. Ihr Erfolg steht und fällt in dem Maße, in dem ihre theoretischen Annahmen im Einzelnen zutreffen und in der Praxis überhaupt anwendbar sind. Die folgende Aufzählung ist sicherlich nicht ganz vollständig, soll aber die Vor- und Nachteile skizzierend einen Überblick über die zahlreichen Korrekturmethoden geben (dazu Abbildung 3)[10].
Die Fallreduktion ist immer noch das wohl am meisten verwendete Verfahren, wobei sich schwer-

[9] Als Konvertiten werden Verweigerer bezeichnet, die nach mehreren Versuchen oder Überzeugungsarbeit doch noch zur Teilnahme an der Erhebung bewegt werden können, vgl. Stoop et al. (2010), S.161.

[10] Hier nicht aufgeführt sind Mixed Patterns-Modelle, vgl. Little und Rubin (2002), S.292ff.

lich von Korrekturverfahren sprechen lässt, da aufgrund der Annahmen keine Korrekturnotwendigkeit besteht.[11] Sie eliminiert Merkmalsträger mit fehlenden Werten. In der einfachsten Form werden nur noch *Complete Cases* verwendet. Diese Art der „Korrektur" birgt das Risiko eines hohen Datenverlustes bei multivariaten Analysen. Obwohl das Verfahren das Standardverfahren in nahezu allen Softwares darstellt, ist die damit verbundene Annahme MCAR häufig unrealistisch; die Schätzeigenschaft der Konstistenz ist unter MCAR allerdings in der Regel gegeben (Spieß 2008, S.14f). *Available Cases* reduzieren die Fälle der einzelnen Variablen, womit der Informationsverlust zwar verringert werden kann, jedoch die Fallbasis für die Analysen jeweils unterschiedlich ist (Bankhofer 1995, S.91 ff; Rässler 2000, S.67; Göthlich 2007, S.123f), was beispielsweise dazu führt, dass ein Wert für den Korrelationskoeffizient außerhalb des Intervalls [-1;1] liegen kann (Bankhofer 1995, S.94).[12]

Nahezu nur für Unit Nonresponse angewandt werden *Gewichtungsverfahren* (Gabler et al. 1994). Bezüglich der Nonresponsekorrektur werden nach einem Soll-/Istvergleich die Verteilungen durch die Gewichtungsvariablen angepasst. Da Gewichtung als das Verfahren schlechthin für die Vermeidung von Unit Nonresponse-Verzerrung in der Praxis gilt, wird es in späteren Abschnitten genauer ausgeführt.

Eine weitere Gruppe besteht aus *Sample Selection Modellen* (SSM), ihr berühmtester Vertreter ist das Heckman-Modell (Heckman 1976; Rässler 2000, S.68). Diese Modelle benötigen aber vergleichsweise viele Informationen über den Ausfallmechanismus. Nach zunächst euphorischer Nutzung der SSM macht sich für die konkrete Anwendung doch Skepsis breit (Schnell 1997, S.248), da die notwendigen Informationen in der Regel nur im geringen Maße vorhanden sind.[13]

Weitere Korrekturmethoden lassen sich zusammenfassen als *Ergänzungs- oder Imputationsverfahren*.[14] Die Art der Ergänzung variiert dabei enorm. Von einfachen Median- oder Mittelwertergänzungen bis hin zu den Hot-Deck-Verfahren oder modellbasierten Ergänzungsmethoden. Bei letzteren Methoden handelt es sich beispielsweise um auf der Likelihood der unvollständigen Daten beruhende oder bayesianisch motivierte Methoden – wie die Single oder Multiple Imputation.[15]

[11] Deshalb gehören eliminierende Verfahren zum „naive approach" (Little und Schenker 1994).

[12] Auch merkmalseliminierende Verfahren sind denkbar, allerdings sehr unpraktisch und unökonomisch, vgl. Bankhofer (1995), S.98.

[13] Spieß skizziert ein Beispiel für ein Selektionsmodell, vgl. Spieß (2008), S.28ff.

[14] von lat. *imputare* – einschneiden, pfropfen.

[15] Hier nicht weiter angesprochen werden Methoden wie *Experteneinschätzung*, *Imputation mittels Zufallszahlen*, *multivariate Imputationsmethoden* – mit Ausnahme der Regressionsmethode – und *Imputation des Verhältnisschätzers*, vgl. Bankhofer (1995), S.104ff).

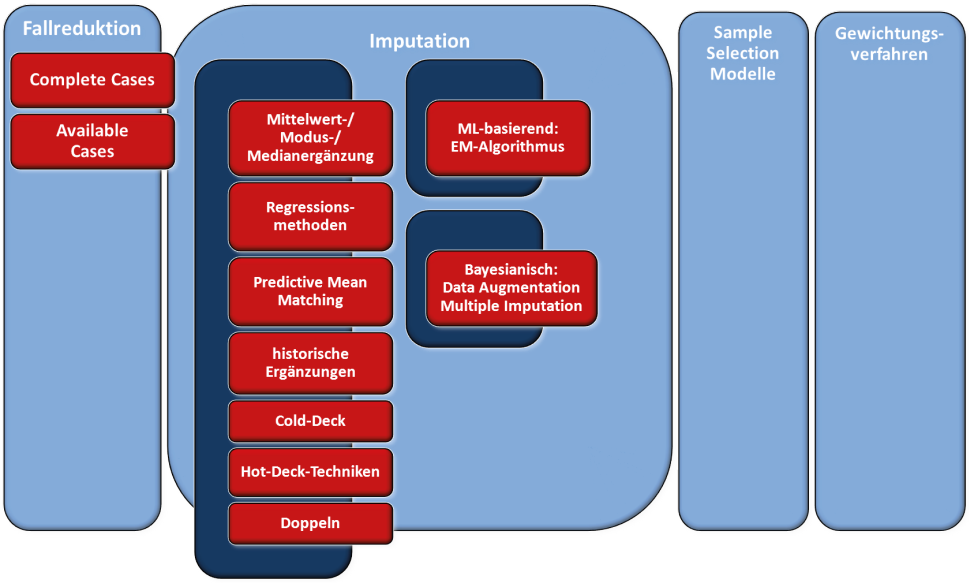



Abbildung 3: Übersicht Verfahren zum Umgang mit fehlenden Werten

Bei *Ergänzungen durch Lageparameter* werden die fehlenden Werte durch die Werte des arithmetischen Mittels, des Medians oder des Modus der beobachteten Fälle ersetzt. Die Ersetzung lässt sich beliebig nach Teilgruppen differenzieren und führt prinzipiell zu konsistenten Schätzungen unter MCAR (Spieß 2008, S.15). Das Verfahren ist zwar äußerst simpel, häufig in der Software auch implementiert, jedoch verweist die Literatur ausführlich auf einen großen Nachteil. Die Varianz wird unter Umständen deutlich unterschätzt (Bankhofer 1995, S.106; Rässler 2000, S.71).

Bei *Hot-Deck-Techniken* werden zuerst Imputationsklassen gebildet, die Antwortende und Nichtantwortende enthalten. Mit Hilfe verschiedener Mechanismen wird nun innerhalb der Gruppe für die fehlenden Werte der Wert aus den Antwortenden zugewiesen, wobei verschiedene Varianten der Auswahl dieser Spender möglich sind. Eine spezielle Methode, einen geeigneten Ersatzwert zu finden, ist das Nearest-Neighbour-Verfahren. Durch eine Distanzmatrix wird derjenige Antwortende ausgewählt, der dem Nichtantwortenden für eine definierte Anzahl von Variablen am ähnlichsten ist. Da hier eine Art Matching stattfindet, wird das Verfahren auch als „statistical matching“ bezeichnet (sehr ausführlich bei Bankhofer 1995, S.120ff; Rässler 2000, S.71). Das sogenannte *Doppeln* lässt sich als „degenerierte Form des Hot-Deck-Verfahrens oder als Gewichtungsmethode“ (Rässler 2000, S.71) auffassen. Das Verfahren fehlende Werte durch einen beobachteten Wert aus vorhergehenden Erhebungen zu ergänzen, wird als *Historische Ergänzung* bezeichnet. Bei regelmäßigen Erhebungen bietet sich diese Form der Ergänzung an, die auch Trendkorrekturen erlaubt (Göthlich 2007, S.126f).
Eine antiquierte Form der Hot-Deck-Techniken, bei der die Spender für die fehlenden Werten aus alten Datensätzen stammt, heißt *Cold-Deck-Techniken* (Göthlich 2007, S.126f). Vorteil dieser Korrekturmethoden ist zunächst, dass nur in den Daten existierende Werte imputiert werden (Spieß 2008, S.19).

Bei entsprechender Datenlage und Zusammenhängen lassen sich *Regressionsgleichungen* (seltener Logit- oder Probitmodelle) aufstellen, bei denen vollständige Variablen als Regressoren dienen. Die fehlenden Werte lassen sich dann mit dem Modell vorhersagen, wobei das Problem auftreten kann, dass gerundet werden muss, da vorhergesagte Werte keine gültigen Ausprägungen sind. Gerade in sozial- und wirtschaftswissenschaftlichen Erhebungen ist dies häufiger der Fall, da der Großteil der Variablen diskret ist. Um Variationsverluste auszugleichen, werden stochastische Störterme addiert.[16]. Der Ausfallmechanismus kann MAR sein.
*Predictive Mean Matching* führt den Ansatz der Regressionsmethode weiter (Rubin 1986, Little 1988). Anstatt zu runden, orientiert man sich am Ergebnis des Regressionsmodells und sucht dann unter den Antworten denjenigen Wert, der am ähnlichsten ist, heraus. Dadurch wird sichergestellt, dass nur tatsächlich auftretende Werte imputiert werden (Göthlich 2007, S.125; Koller-Meinfelder 2010, S.31ff).
*Maximum-Likelihoodbasierte Methode* ist beispielsweise die Schätzung unvollständige Daten mit dem EM-Algorithmus. Da der Datenausfall die ML-Schätzung (genauer die Maximierung der Loglikelihood) stark verkompliziert, werden im *E*xpectation-Schritt die fehlenden Daten als Erwartungswerte für Variablen mit fehlenden Werten zum unbekannten Parameter ergänzt.[17] Anschließend wird im *M*aximiziation-Schritt ein neuer Wert aus den ML-Schätzer des Parameters aus den beobachteten und den im E-Schritt ergänzten Daten errechnet. Das Verfahren iteriert bis eine definierte Abweichung unterschritten wird (Dempster et al. 1977). Als stochastische Variante des EM-Alghorithmus gilt Data Augmentation (Schafer 1997a, S.37ff). Geläufiger ist die Bezeichnung *Markov-Chain-Monte-Carlo-Methoden (MCMC)*.[18] Hier wird die Imputation durch einen Zufallsterm ergänzt (Göthlich 2007, S.128; Allison 2002). Als Voraussetzung für die Anwendung der parameterschätzenden Verfahren gelten MAR und die Normalverteilungsannahme. Teilweise wird auch ein spezielles Muster der fehlenden Werte vorausgesetzt. Zudem geht die Theorie dieser Verfahren zunächst von stetigen Merkmalen aus (Bankhofer 1995, S.166f).
*Multiple Imputation* stellt dem Nutzer mehrere vervollständigte Datensätze zur Verfügung (Rubin 1978; Longford 2008, S.143). In der Zahl unterschiedlich ergänzter Datensätze spiegelt sich die Unsicherheit, die durch die Imputation entsteht, wider (dieser Vorteil entfällt bei der Single Imputation). Zu den grundlegenden Bedingungen zählt, dass der Ausfall MAR ist. Zwar vereinfacht ein monotones Ausfallmuster die Imputation, jedoch stellen beliebige Ausfallmuster heute für aktuelle Software kein Problem mehr da (Stichwort: Chained equitations).[19]

[16] Ganz ausführlich, auch zu ihrer historischen Entwicklung vgl. Bankhofer (1995), S.126ff. Deshalb zählt Nordholt konventionelle Imputationen mit zusätzlichen Residuen zur Gruppe der „stochastic imputation methods“, vgl. Nordholt (1998), S.160.

[17] Bei Rässler (2000), S.73ff wird die Maximierung der Likelihood unter Datenausfall sehr anschaulich erklärt.

[18] Schafer (1997a), S.68 spricht hier allerdings von einer zu laxen Definition des MCMC-Begriffs.

[19] Beispiel von Muster werden in Abschnitt 4.2.1, 4.2.2 und 4.2.3 dargestellt.

## 2.5 Zwischenfazit

Das erste inhaltliche Kapitel sollte deutlich machen, dass der Forschungsgegenstand aufgrund der Vielzahl von Definitionen und der unterschiedlichen Perspektiven von hoher Komplexität ist. Die Ubiquität fehlender Werte trifft gerade in den Sozialwissenschaften auf unterschiedliche Konzepte der Definition und der Behandlung. Während eine große Surveymethodenliteraturauswahl sich mit Vermeidungsstrategien befasst, konzentrieren sich Statistiker in erster Linie auf Korrekturmethoden. Die Einbettung des Problems in dem gesamten Erhebungsprozess, wie einführend durch den Survey Lifecycle skizziert wurde, erfordert jedoch eine Berücksichtigung beider Perspektiven. Die Breite der statistischen und methodischen Literatur verdeckt auch die Tatsache, dass die meisten Erhebungen in teils bewusster, teils unbewusster Ignoranz gegenüber den Forschungsergebnissen bezüglich Missing Data durchgeführt werden. Wie in den praktischen Teilen noch zu sehen sein wird, wird vor allem in den Sozialwissenschaften in der Erhebungspraxis allenfalls eine geringe Bandbreite von Vermeidungsstrategien eingesetzt. Die Anwendung von Korrekturverfahren orientiert sich sehr häufig an einem naiven Vorgehen.

# 3 Item Nonresponse: Theorie und Determinanten

## 3.1 Einleitung

Es gibt nur wenige Ansätze zu einer umfassende Theorie für Item Nonresponse, die tatsächlich den gesamten Erhebungsprozess berücksichtigten. Häufig handelt es sich um Theorien, die eine bestimmte Komponente des Erhebungsprozesses – z.B. die Fragestellung, die Stichprobenziehung oder das Interview – bezüglich Item Nonresponse analysiert.[20] Da einzelne Elemente einer Erhebung an sich bereits komplex sind, sind viele vorhandene Erkenntnisse in ihrer Aussagekraft eingeschränkt, andere stehen isoliert. Abschnitt 3.2 fasst die theoretischen Erkenntnisse bezüglich der Entstehung von Item Nonresponse zusammen, 3.3 arbeitet die Determinanten unter Vewendung des ALLBUS 2006 heraus.

## 3.2 Theorie zur Entstehung von Item Nonresponse

Die folgenden Abschnitte nähern sich sukzessiv dem theoretischen Kern von Item Nonresponse. Als großer Rahmen dient zunächst der Survey Lifecycle, mit dessen Hilfe Item Nonresponse begünstigende und mindernde Faktoren analysiert werden (3.2.1). Item Nonresponse manifestiert sich während des Interviews, das als soziale Interaktion in Abschnitt 3.2.2 betrachtet wird. Daran knüpf sich eine Theorie des Antwortens an (angelehnt z.B. an Tourangeau et al. 2000, S.23ff). Diese Theorie findet eine Konkretisierung im Entscheidungsprozess nach Beatty und Herrmann (2002), in dem das Auftreten von Item Nonresponse als spezifisches Antwortverhalten erklärt wird (3.2.3). Mit Hilfe dieser Konkretisierung lassen sich für große Stichprobenerhebungen sodann Schlüsse für die Auswahl der weiteren Korrekturmethoden ziehen (3.2.4).

[20] Bei McKnight et al. (2007) wird das Konzept einer umfassenden Berücksichtigung der Erhebung angedacht. Auch Bankhofer (1995) hat Ansätze für eine umfassende Perspektive.

### 3.2.1 Übersicht: Item Nonresponse begünstigende und mindernde Faktoren

Einzelne Teile des Survey Lifecycle können als Schlüssel zum Verständnis des Item Nonresponse dienen, der gesamte Lifecycle sogar zur Einordnung des Problems. Diese Überlegungen erleichtern darüberhinaus den Umgang mit Item Nonresponse, indem sie Hinweise zur Auswahl der geeigneten Korrekturmethode geben können.
Im Folgenden soll nun anhand der Prozess- und Designperspektive eine Übersicht über die Wirkung und Wechselwirkung einzelner Parameter einer Erhebung und deren Einfluss auf Item Nonresponse gegeben werden. Die jeweiligen Stellen im Survey Lifecycle, die sich als wichtige Faktoren in der Literatur herausgestellt haben, sind farblich hervorgehoben.

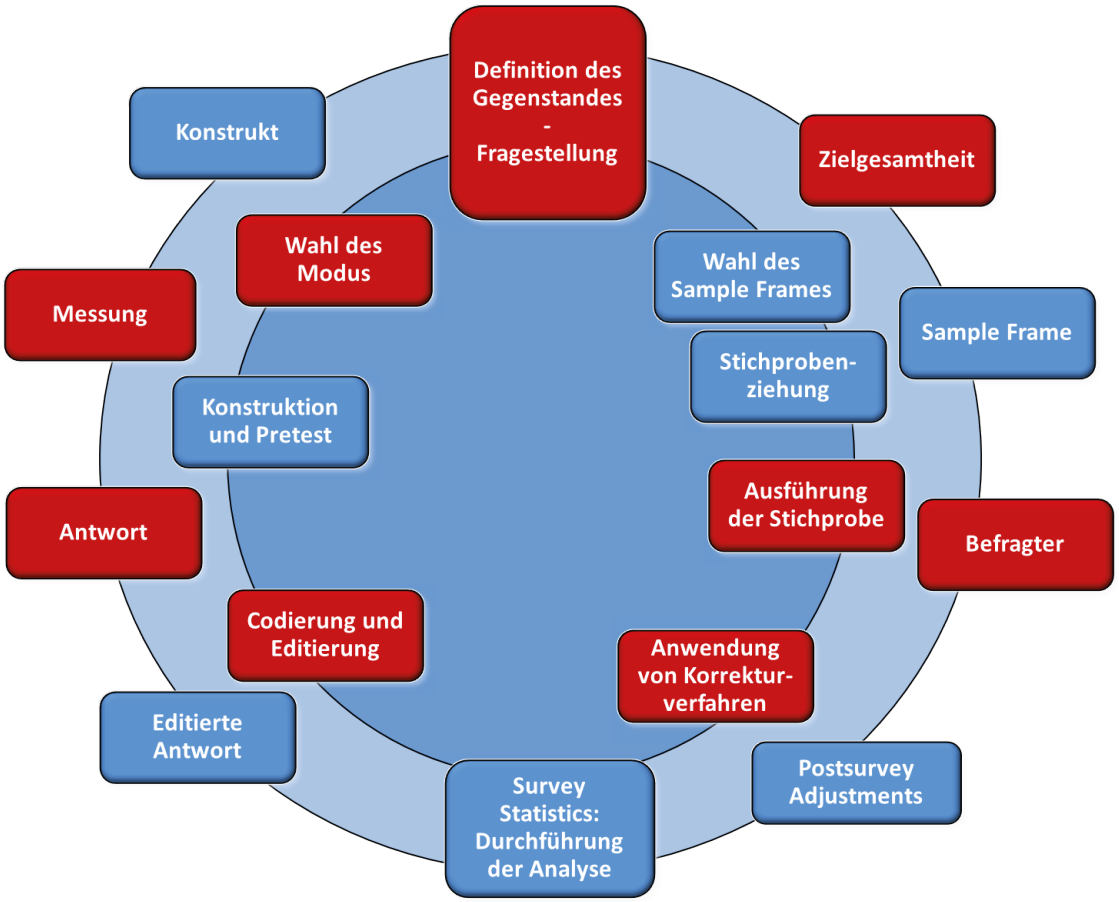



Abbildung 4: Survey Lifecycle und Item Nonresponse

So wie die **Fragestellung** den weiteren Erhebungsprozess konstituiert, determiniert sie auch das spätere Auftreten fehlender Werte. Betroffen sind neben Wissensfragen vor allem sensible Fragestellungen.[21] Eine Abgrenzung von sensiblen und nicht sensiblen Themen lässt sich nicht scharf

[21] Wenn zum Beispiel Datums- oder Zeitangaben abgefragt werden, vgl. Tourangeau (2000), S.100ff; Fowler and

durchführen. In der Regel gelten Fragen zu „sexual behavior, drugs and alcohol abuse, criminal offences and fraud, ethical problems, and attitudes involving abortion, euthanasia and suicide, as well as charity, politics, medical compliance, psychological problems“ (Lensvelt-Mulders 2008, S.462) als sensibel. Prinzipiell wird in der Literatur vermutet, dass sensible Themen zu steigenden Nonresponse führen – und zwar sowohl zu steigenden Item als auch zu steigenden Unit Nonresponse-Anteilen (Catania et al. 1990; Lensvelt-Mulders 2008, S.464). Findet in einer großen Erhebung eine Befragung größtenteils über nicht sensible Themen statt, werden nur bei den sensiblen Fragen höhere Item Nonresponse-Anteile zu beobachten sein. Die empirischen Ergebnisse sind hier allerdings nicht eindeutig. So ermitteln beispielsweise Tourangeau et al. keinen dramatisch höheren Ausfall bei Fragen zur sexuellen Orientierung und Verhaltensweisen (Tourangeau et al. 2000, S.263f). Anders als bei Unit Nonresponse, die aufgrund sensibler monothematischer Erhebungen real ansteigen können, werden viele Befragte im Interview nicht unbedingt Item Nonresponse generieren, sondern machen eine falsche Angabe. Das heißt, dass unter Umständen kein wesentlich erhöhter Item Nonresponse-Anteil zu beobachten ist, sondern eher mit einer verzerrten Messung zu rechnen ist. Damit würde das Nonresponse Problem zum Messproblem – ein Trade-off, der häufig nicht umgangen werden kann (Tourangeau et al. 2000, S.264f). Der Zusammenhang fällt dabei wohl je nach Erhebungsmodus unterschiedlich stark aus.

Zusammen mit dem Erkenntnisinteresse, das in der Fragestellung zum Audruck kommt, ist die zweite zentrale Frage die nach der **Zielgesamtheit**: Über wen sollen Informationen gesammelt werden? Die Zielgesamtheit kann je nach Fragestellung durchaus vermehrt Personen mit Merkmalsausprägungen aufweisen, von denen aus der Literatur bekannt ist, dass sie zu einer höheren Eintrittswahrscheinlichkeit von Item Nonresponse führen. Es lässt sich beispielsweise bei Personen mit niedriger Bildung oder schlechterem Sprachverständnis (Zuwanderer) in Studien ein höheres Niveau an Item Nonresponse beobachten.[22]

Die Möglichkeit, dass das Messkonzept oder dessen Umsetzung, also die **Messung** selbst, einen Einfluss auf die Höhe des Item Nonresponse haben kann, ist evident. So wird beispielsweise die Höhe des Item Nonresponse-Anteils davon abhängen, ob das Messkonzept die „weiß nicht“- oder „keine Angabe“- Kategorien zulässt (Borg und Staufenbiel 2007; allgemein zu Antwortalternativen: Schwarz et al. 2008, S.20; Tourangeau et al. 2000, S.299). Zudem kann die Komplexität der Messung, die sich in einem schwierigen Wording, komplexer Filterführung oder ähnlichen Designfaktoren manifestiert, Item Nonresponse-Auftreten beeinflussen (Schwarz et al. 2008, S.19f).

Dass auch der **Erhebungsmodus** einen Einfluss auf Item Nonresponse haben kann, ist zumindest plausibel (De Leeuw 2008, S.123ff; Tourangeau et al. 2000, S.306). Der Zusammenhang gestaltet sich aber komplexer: Je mehr Unpersönlichkeit der Modus generiert, desto unproblematischer kann die Beantwortung sensibler Fragestellungen sein. Danach lässt sich generell nach Erhebungen mit Interviews und „self administered“ Erhebungen unterscheiden, wobei das Telephoninterview unpersönlicher als das persönliche Interview ist. Bei Telefoninterviews entfällt zumindest der Komplex nonverbaler Kommunikation. Der Interviewer kann bezüglich der Entstehung von Item Nonresponse eine ambivalente Stellung einnehmen. Er kann das Risiko erhöhen oder vermindern, dass beispielsweise aufgrund sozialer Erwünschtheit Item Nonresponse entsteht.[23] Viele Aufgaben des Interviewers können allerdings auch fehlende Werte verhindern; de Leeuw (2008), S.115 fasst

Consenza (2008), S.143ff; zu sensible Fragestellungen vgl. Schuman und Presser (1981), S.203ff.

[22] Einige Ergebnisse aus der Unit Nonresponse-Forschung lassen sich hier übertragen, es gibt aber auch einzelne Beiträge speziell für Item Nonresponse, vgl. Thiessen und Blasius (1998), S.240f.

[23] Allgemein zum Phänomen sozialer Erwünschtheit, vgl. Stocké und Hunkler (2004).

diese Aufgaben zusammen: „they have to motivate respondents, to deliver and when necessary clarify questions, to answer respondent's queries, and to probe after inadequate answer"Ėmpirische Analysen deuten deshalb darauf hin, dass zwar die Item Nonresponse-Quote durch das persönliche Interview geringer wird, jedoch – wie sich am Beispiel sensibler Fragestellungen zeigen lässt – die Beantwortungsqualität schlechter ist als bei postalischen Erhebungen (De Leeuw 2008, S.123f).
Daneben verlangen unterschiedliche Erhebungsmodi von den Befragten unterschiedliche Fähigkeiten. Dies reicht von der generellen Lese- und Schreibfähigkeit, der Fähigkeit zuzuhören, der Konzentrationsfähigkeit, der Gedächtnisleistung bis hin zur Tippfähigkeit und dem Umgang mit Tastaturen (z.B. bei computergestützten Ausfüllung durch den Befragten selbst) (De Leeuw 2008, S.122f; Lynn 1998, S.9f; Martin et al. 1993, S.654). Zusammenhänge zwischen diesen Faktoren und Item Nonresponse erscheinen plausibel, sind bisher allerdings kaum methodisch aufgearbeitet.[24]
Da die eigentliche Manifestation des Item Nonresponse erst in der **Interviewsituation** erfolgt, bedarf es einer ausführlicheren Darstellung dieser komplexen sozialen Interaktion. Ob Item Nonresponse tatsächlich entsteht oder nicht, wie einzelne vorgelagerte Faktoren des Erhebungsprozess die Entstehung von Item Nonresponse faktisch begünstigen oder minimieren, entscheidet sich letztlich während des Interviews. Deshalb findet in den folgenden Abschnitten (3.2.2 und 3.2.3) eine detaillierter Auseinandersetzung mit der Interviewsituation statt.
Nach der eigentlichen Datenerhebung werden die **Antworten editiert**. Neben der Digitalisierung (falls noch nicht geschehen) gilt es, die Daten in gängige Softwareformate zu bringen. Außerdem werden häufig die Antworten auf ihre Konsistenz geprüft. So werden Antworten, die außerhalb gewisser Bandbreiten liegen, entweder korrigiert, wenn es plausible Anhaltspunkte dafür gibt, oder im Zweifel als Item Nonresponse definiert. Im Zuge der Editierung identifiziert man häufig Ausreißer und definiert diese unter Umständen als Item Nonresponse, ehe die Datensätze den Nutzern zugänglich gemacht werden (Groves et. al 2004, S.44). Intensive Editierung enthält durchaus subjektive Züge, was die Definition von Ausreißern und Konsistenzen angeht. Qualitativ hochwertigen Datensätzen ist deshalb eine Dokumentation der Editierung beigefügt. Bei entsprechender Qualität der Erhebung sollte die Editierung allerdings eine vernachlässigbare Quelle von Item Nonresponse sein (Schmith 2002, S.34ff).
Korrekturmethoden nach der eigentlichen Erhebung erzeugen an sich keine Item Nonresponse. Ihr Einsatz entscheidet aber darüber, ob und wieviel Item Nonresponse im Datensatz zu finden ist, ehe die Datennutzer auf ihn zugreifen können. Dies begrenzt dann die später noch anwendbaren Korrekturmethoden der jeweiligen Nutzer (Enders 2010; Longford 2008, S.135ff), die ebenfalls dokumentiert werden müssen.

[24]Moderierende Merkmale für Erhebungen seien nach Tourangeau et al. Unpersönlichkeit, Legitimität, Wichtigkeit der Studie und kognitive Hürden. Diese drei besitzen wohl Einfluss auf die Höhe der Item Nonresponse, wobei die empirischen Befunde auf keine allzugroße empirische Evidenz hindeuten, vgl. Tourangeau et al. (2000), S.304ff.

### 3.2.2 Interview: Interaktion von Interviewer und Befragten

„Wer kommuniziert – interagiert gleichzeitig. Wer interagiert – kommuniziert gleichzeitig."
Paul Watzlawick

Die weitere Ausführungen und die Anwendung spezifischer Theorien beziehen sich auf den Typus der stark strukturierten Befragung, wie sie auch bei den später verwendeten Beispieldatensätzen durchgeführt wurde. Es handelt sich zudem um persönliche oder face-to-face Interviews.[25] Der Interviewer kann unterschiedlich stark involviert sein. Es gibt ein breites Spektrum, das von starker Involvierung wie beim persönlichen Einzelinterview bis hin zu schwacher Involvierung bei postalischen Interviews reicht, jedoch lässt sich in allen Fällen davon ausgehen, dass es sich um einen „sozialen Vorgang" handelt: das Interview ist eine soziale Situation. „Von sozialer Situation ist selbst dann zu sprechen, wenn jemand für sich allein auf einen schriftlichen Fragebogen Antwort gibt[...]" (Atteslander 2010, S.112). Das bedeutet, dass auch weniger strukturierte und unpersönlichere Befragungen die im Folgenden dargestellten Merkmale und Prozesse abgestuft aufweisen.
Fasst man das Interview als ein *Stimulus-Reaktionsmodell* auf, spielen alle möglichen Reize wie Interviewer, Zeitpunkt und Interviewort, der Fragebogen, aber auch die Erwartungen und Normen des Befragten selbst nur als allgemeine Störfaktoren eine Rolle. Dieses Modell geht davon aus, dass ein direkter, ausschließlicher und zwingender Zusammenhang zwischen einem Stimulus (Frage) und einer bestimmten Reaktion (Antwort) besteht. Demnach ist allein der Fragebogen (als Summe der Stimuli) der Schlüssel zu einer Befragung – eine nicht selten vertretene Position unter Methodikern (Atteslander 2010, S.112).

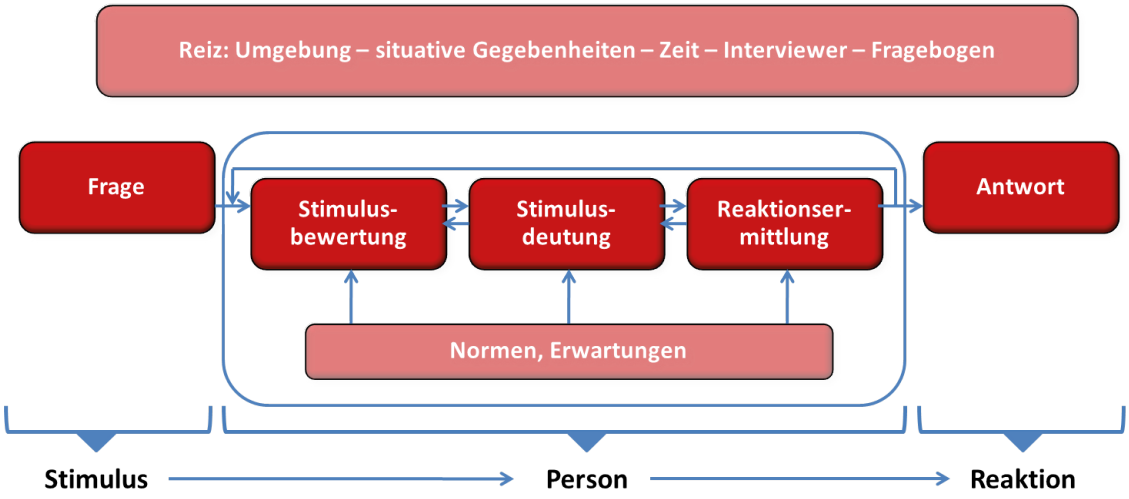



Abbildung 5: Schema Stimulus-Person-Reaktionsmodell

[25] „[...] face-to-face interview in the context of survey research can be defined as a face-to-face interaction between two persons in which one person (interviewer) asks questions by means of a questionnaire and the other person (respondent) answers these questions." (Loosveldt 2008, S.201).

Um das Modell realitätsgerechter zu machen, muss das Stimulus-Reaktionsmodell durch die Einbeziehung der Person des Befragten zum *Stimulus-Person-Reaktionsmodell* erweitert werden (siehe Abbildung 5). Der Stimulus, den der Befragte empfängt, wird bewertet, gedeutet und daraufhin abgeschätzt, was die mögliche Reaktion für den Befragten bedeutet. Nach Atteslander gelten bei der kognitiven, emotionalen und rationalen Verarbeitung des Fragestimulus auch Normen des Befragten als wichtiger Faktor. Diese Normen lassen sich wiederum in gesamtgesellschaftliche, gruppenspezifische und interviewspezifische Normen unterteilen (Atteslander 2010, auch schon bei Atteslander und Kneubühler 1975).[26] Latent wirken sich auf den Befragten und den in ihm ablaufenden Prozessen der Stimulusverarbeitung eine Fülle von Reizen aus: Umgebungsmerkmale, situative Gegebenheiten, Fragebogen und als wichtigstes die Person des Interviewers.
Zu den **Umgebungsmerkmalen** zählt neben atmosphärisch situativen Merkmalen das Vorhandensein dritter Personen während des Interviews.[27] Die Anwesenheit von Dritten kann nach Reuband vom Part des passiven Zuhörers bis hin zum Adressaten oder Katalysatoren des Interviews reichen (Reuband 1984, S.117). Leider werden Studien bezüglicher Dritter selten auf den Spezialfall von Item Nonresponse analysiert, sondern versuchen, Inkonsistenzen oder falsche Antwortangaben nachzuweisen. Jedoch kann man davon ausgehen, dass die Anwesenheit von weiteren Personen während des Interviews auch die Verweigerung gerade bei sensiblen Items beeinflusst (Haunberger 2006, S.39f).

Die Auswirkungen der **Interviewereigenschaften** stellen den wohl bedeutendsten Bestandteil der sozialen Interaktion Interview dar. Gerade die Erwartungen des Befragten beziehen sich auf die Person des Interviewers und des Interviews (Esser 1984, Reinecke 1991) und beeinflussen damit die Reaktion auf den Stimulus. Das persönliche Interview gilt in seiner Gesamtheit als Kommunikation, die sich aber aufgrund der Rollenaufteilung in Frager und Befragten von der alltäglichen Kommunikation unterscheidet (Loosveldt 2008) und sogar als einschüchternde Ausfragesituation wahrgenommen werden kann (Krebs und Schüssler 1987). Der Interviewer besitzt sowohl sichtbare als auch unsichtbare Merkmale, die der Befragte wahrnehmen kann (Reinecke 1991, S.27ff). Nach dem Eisberg (Watzlawick et al. 2011) lässt sich davon ausgehen, dass diese Merkmale – großteils unbewusst – Signale aussenden, die vom Befragten wiederum bewusst oder unbewusst perzipiert werden können. Es darf nicht vergessen werden, dass Interviewer und Befragter Reizempfänger und Reizsender gleichermaßen sein können (Esser 1984). Aus dieser Tatsache erwächst der Interaktion Interview eine dynamische Komponente.[28] In der zeitlichen Entwicklung wirkt die soziale Interaktion auf die (unsichtbaren) Merkmale, die auf die Interviewsituation selbst Einfluss nehmen (Steinerts 1984). Deshalb teilt Esser (1984) den Interviewprozess in drei Einzelaspekte ein: Erstens die Reaktion des Befragten in Abhängigkeit vom Fragestimulus und der Beeinflussung durch den Interviewer. Zweitens die Stimulierung, Beeinflussung und Reaktion des Interviewers, die einerseits der Befragte anderseits der **Fragebogen** (Fragedesign) auslöst. Schließlich die Vercodungshandlung, falls diese allein beim Interviewer liegt. Auch diese Handlung unterliegt der

[26] Da die Prozesse im Befragten nicht empirisch, sondern allenfalls theoretisch erfasst werden können, soll an dieser Stelle auf die ausführlichen Erklärungsmodelle bei Reinecke (1991), S.35ff verwiesen werden.

[27] Dabei können Umgebungsmerkmale im Interview niemals vollständig empirisch erfasst werden, vgl. Esser (1984).

[28] Man kann sich dies so vorstellen, dass Abbildung 5 den einmaligen Ablauf von Stimulus-Person-Reaktion veranschaulich. Da ein Interview in der Regel aus mehreren oder sehr vielen Stimuli besteht, durchläuft die Interaktion mit sich veränderten Parametern bis zum Interviewende immer wieder den in Abbildung 5 dargestellten Prozess.

Beeinflussung der Befragtenreaktion, weiteren situativen und schließlich den eigenen Merkmalen des Interviewers. Dieser Einfluss kann – wie oben – genannt nur im schriftlichen Interview (Mail, Post) im großen Umfang vermindert werden, da es für die Interviewerperson unmöglich ist, nur auf den Fragestimulus bezogen zu kommunizieren (Watzlawick et al. 2011).
Eine frühe Studie zur empirischen Evidenz stammt von 1972: In einer multivariaten Analyse untersuchten Schanz und Schmidt (1972) den Einfluss von Unmgebungsmerkmalen, Interviewereigenschaften und Reaktionen des Befragten. Hierbei griffen sie auf eine große Bandbreite von Variablen zurück, die bereits in der älteren Literatur als vermutete Reizfaktoren galten:[29] Alter, Geschlecht, Schulbildung und Status des Interviewers, die Anwesenheit Dritter, alsdann als interagierende Variablen Alter, Geschlecht und Bildung des Interviewers mit den entsprechenden Merkmalen des Befragten, die Interaktion sozialer Distanz zwischen Interviewer und Interviewtem (Schanz und Schmidt 1984, S.74f.). Gerade die sichtbaren soziodemografischen Merkmale des Interviewer könnten die Reaktion des Befragten bei solchen Themen beeinflussen, bei denen eines der Merkmale einen Hinweisreiz darstellt.
Leider werden bei vielen Studien nur die generellen Auswirkungen von Reizen auf das generelle Antwortverhalten untersucht, während die spezifischen Auswirkungen auf Item Nonresponse ausgeblendet werden. Im Zentrum steht zumeist die Frage, inwieweit die Messung durch soziale Erwünschtheit oder die inhaltsunabhängige Zustimmungstendenz der Befragten verzerrt wird (Reinecke 1991, S.23ff). Haunberger (2006) hat beispielsweise einfach messbare Merkmale untersucht, allerdings dabei nicht nur nach Verzerrung (also „falscher“ Beantwortung) sondern auch nach den Konsequenzen für die Enstehung von Item Nonresponse gefragt (Haunberger 2006, S.28ff).[30] Sie identifiziert in erster Linie über die Frage nach dem Netto- bzw. Haushaltseinkommen sensible Themen als Stimuli mit erhöhter Verweigerungsrate. Hauptsächlich erklärende Variablen sind dabei vor allem Interaktionseffekte wie der Umstand, dass Interviewer und Befragter zur gleichen Bildungsgruppe gehören (Haunberger 2006, S.43).

Da Item Nonresponse keine zwingende, sondern nur eine von mehreren Reaktionen (Antworten) auf den Fragestimulus darstellt, muss in Abschnitt 3.2.3 die Frage geklärt werden, unter welchen Umständen das Ergebnis des Antwortprozesses auf Item Nonresponse verengt wird.

[29] Interviewereffekt bedeutet, dass sichtbare oder unsichtbare Merkmale des Interviewers zu einem Antwortverhalten führen, das nicht das „wahren“ Antwortverhalten des Befragten widerspiegelt, vgl. Esser (1984).

[30] Als Reize wurden in der Analyse berücksichtigt: Anwesenheit Dritter, Interviewlänge, Bildung, Alter und Geschlecht des Befragten und des Interviewers sowie Interaktionsterm aus dem Bildungsniveau, vgl. Haunberger (2006), S.39.

### 3.2.3 Theorie zu Entscheidungsprozessen bei Item Nonresponse

Beatty und Herrmann (2002) formulieren einen theoretischen Rahmen, der den Entscheidungsprozess der befragten Person gleichsam im letzten Stück des Stimulus-Personen-Reaktionsmodells nachzeichnet und wiederum von drei Faktoren bestimmt wird. Der erste Faktor umfasst den kognitiven Status (KS), der als eigener theoretischer Ansatz noch explizit ausgeführt wird. Zweitens entscheidet der Befragte über die Angemessenheit seiner Antwort und schließlich unterliegt die tatsächliche Antwort der kommunikativen Absicht des Befragten (Beatty und Herrmann 2002, S.72). Diese Faktoren werden sowohl von den persönlichen Eigenschaften des Befragten beeinflusst, als auch von den im Abschnitt 3.2.2 beschriebenen Normen und Erwartungen bestimmt, mit der der Befragte in das Interview geht, sowie den Erwartungen und Empfindungen, die sich in der Interaktion mit dem Interviewer im Interview herausbilden. Beatty und Herrmann greifen dabei einerseits auf Erkenntnisse der psychologischen Gedächtnis- und Lernforschung zurück (z.B. Hasher und Griffin 1978; Reder 1988; für die Survey Methodologie vgl. Tourangeau et al. 2000, S.62ff) andererseits beziehen sie sich implizit auf das Stimulus-Person-Reaktionsmodell.
Der Stimulus konfrontiert den Befragten mit zwei grundlegenden Fragen: ‚Kann ich auf den Stimulus antworten?' und ‚Will ich auf den Stimulus antworten?' Bei der ersten Frage spielt die Aktivierung relevanten Wissens die entscheidende Rolle. Bei der zweiten Frage ist die Motivation zur Beantwortung entscheidend. Der Befragte kann aus verschiedenen Motiven die Antwort verweigern, weil er nicht auf eine sensible Frage antworten möchte, weil er die Anstrengung scheut sich z.B. zu erinnern oder weil der Befragte einen vermeintlichen Konflikt und daraus resultierende Unannehmlichkeiten bei der Beantwortung fürchtet.
Kern des ersten Teils ist der kognitive Status des Befragten, der die Aktivierbarkeit des notwendigen Wissens beschreibt, auf den der Stimulus, die Frage, trifft. Beatty und Herrmann (2002) teilen deshalb zunächst das Wissen des Befragten in vier kognitive Status ein, die im Schema dargestellt sind (Beatty und Herrmann 2002, S.73ff).

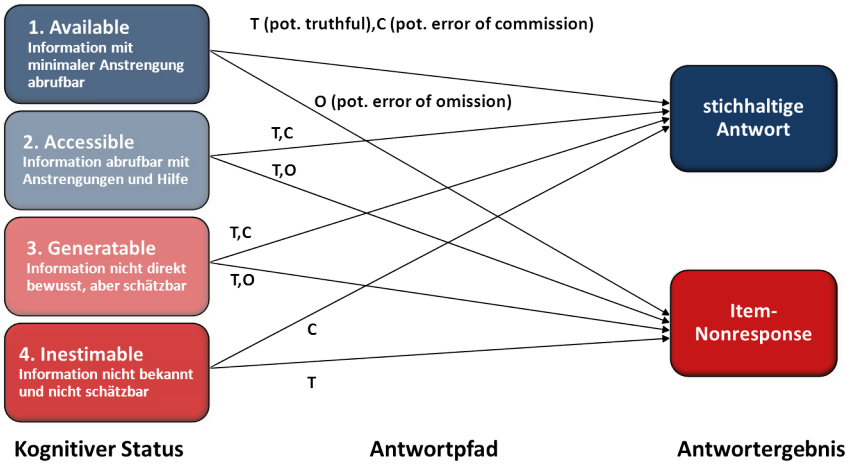



Abbildung 6: Schema kognitiver Status und Item Nonresponse

Wenngleich Wissen eine (latent) stetige Variable darstellt, werden der Übersichtlichkeit halber vier Levels kognitiver Status gebildet. Die Levels lassen sich in eine Reihenfolge bringen, in der das Wissen für die Beantwortung des Fragestimulus verfügbar ist: vorhanden, zugänglich, generierbar und nicht (oder kaum) schätzbar. Im ersten Status ist die Information, die zur Beantwortung der Frage benötigt wird, vorhanden und mühelos abrufbar. Gibt der Befragte nun eine Antwort, kann diese nach menschlichem Ermessen immer falsch sein oder richtig. Verweigert der Befragte die Antwort aufgrund der anderen Entscheidungsfaktoren, hat es bewusst gegen den eigentlichen kognitiven Status reagiert (zumindest mit hoher Wahrscheinlichkeit). Im zweiten Status lässt sich die für die Antwort benötigte Information mit etwas Mühe abrufen. Es ist eine abgeschwächte Form des ersten Status. Beim dritten Status wird der Aufwand für eine Fragebeantwortung noch einmal deutlich höher und kann unter Umständen nur noch geschätzt werden. Im letzten kognitiven Status besitzt der Befragte das notwendige Wissen schlichtweg nicht. Die Konsequenz ist die Entscheidung für eine Antwortverweigerung. Dies ist jedoch keinesfalls zwingend. Der Befragte kann nämlich eine Antwort geben, die richtig oder falsch ist. Mit zunehmender kognitiver Herausforderung sind zunächst die Stärke der Erinnerung oder die allgemeinen intellektuellen Fähigkeiten des Befragten sowie der Willen, diese zu benutzen, entscheidend. Außerdem muss der Befragte zum Schluss kommen, dass seine Reaktion (Antwort) adäquat für den Fragestimulus zu sein scheint.
Beatty und Herrmann richten ihr Hauptaugenmerk auf die mittleren Status, bei denen mit einem gewissen Aufwand, die Information generiert oder geschätzt werden kann. Dabei entscheidet der Faktor der adäquaten Urteilsbildung unter Umständen über die Entscheidung einer vielleicht nicht richtigen Antwort oder über die Antwortverweigerung, die auch falsch sein kann, wenn die Information doch genierbar oder schätzbar gewesen wäre. Hier kommen wir zur zweiten Frage. Wenn die befragte Person das notwendige Wissen besitzt oder generieren kann, steuert das kommunikative Ziel des Befragten die Beantwortung der Frage.
Das nachfolgende Schema (Abbildung 7) verdeutlicht im Prozess vom Stimulus bis zur Entscheidung, ob Item Nonresponse generiert wird oder nicht. Dabei werden noch einmal die Erklärungsfaktoren für das Zustandekommen von Item Nonresponse zusammengefasst: Unverständnis bezüglich der Fragestellung, niedrige Motivation, Zurückhaltung der Informationen, obwohl sie vorhanden sind, und schließlich die Überzeugung, dass Item Nonresponse die einzig adäquate Antwortform ist.[31]

[31] Ähnliche Faktoren auch bei Krosnick: er betont allerdings noch gesondert die eventuell einschüchternde Wirkung der Interviewsituation, vgl. Krosnick (2002).

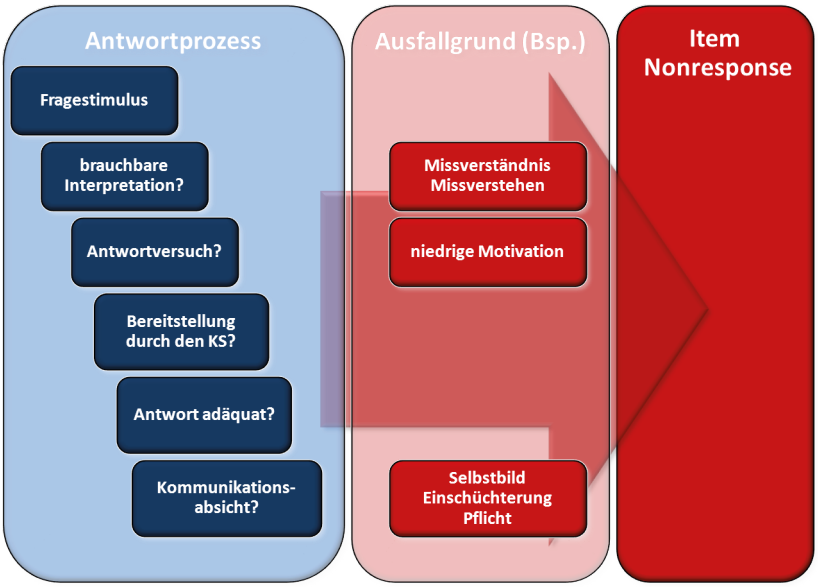



Abbildung 7: Antwortprozess nach Beatty und Herrmann (2002)

Ist die Motivation sowieso so gering, dass unter Umständen der Befragte die Hürden zur Beantwortung gar nicht nehmen will, wird er zügig zur Item Nonresponse-Kategorie kommen. Determinanten für die Antwortmotivation können beispielsweise Zentralität und Interesse am Fragethema sein (Krosnick 2002, S.94). Ist die Motivation nicht das Problem und die Antwort wäre kognitiv verfügbar, jedoch beispielsweise aus Gründen der sozialen Erwünschtheit nicht adäquat, steht als Reaktion am Ende ebenfalls eine Verweigerung. In diesem Fall ist das Streben des Befragten im Interview nach Selbstdarstellung (bezüglich seiner selbst und des Interviewers) und Anerkennung (von Seiten des Interviewers für den Befragten) Ziel der Nichtantwort (beispielsweise Esser 1984; Reinecke 1991; Büschges et al. 1995; Krosnick 2002).[32] Bei Befragten, die eine verminderte Kommunikationsabsicht haben, können allerdings sowohl Antwortverweigerungen oder die Vorgabe, die Antwort nicht zu wissen, als auch die bewusste Angabe einer falschen Antwort in Frage kommen (Reinecke 1991).[33]
Die eben skizzierten Theorien erfüllen keinen Selbstzweck. Trotz ihrer starken Vereinfachung des sehr komplexen Gegenstandes, können sie auch interessante Anstöße geben, wie mit Item Nonresponse umgegangen werden kann.

[32] Auch das Extrem der Einschüchterung durch die Fragesituation ist denkbar, vgl. Krosnick (2002), S.98.

[33] Beatty und Herrmann (2002), S.73 sprechen bei ersterem entsprechend der Theorie des kognitiven Status von „error of omission“, im zweiten Fall von „error of commission“.

### 3.2.4 Konsequenzen für den Umgang mit Item Nonresponse

Aus den obigen theoretischen Überlegungen ergeben sich zunächst Konsequenzen für die Vermeidung von Item Nonresponse. So fordern Beatty und Herrmann eine möglichst simple Formulierung der Fragen im Fragebogen, um Verständnishürden abzubauen. Die Zerstreuung von Befürchtungen, die Betonung von Vertraulichkeit und Privatheit sind genauso Aufgabe des Interviewers, wie eine aktive Rolle bei der Klärung von Verständnisproblemen. Sie weisen also der Person des Interviewers eine sehr hohe Bedeutung zu, positiv auf den Entscheidungsprozess einzuwirken (Beatty und Herrmann 2002, S.84f). Krosnick betont zudem, wie wichtig die Existenz von „weiß nicht"- und „keine Angabe"-Kategorien im Fragebogen ist (Krosnick 2002, 99f).[34] Die große Mehrheit in der Wissenschaftsgemeinschaft besitzt die Möglichkeiten der Vermeidung nicht und muss sich mit den Gegebenheiten des Datensatzes auseinandersetzen. Auch hier können die theoretischen Überlegungen Hilfe leisten.
Der Mehrzahl der Korrekturverfahren liegen Annahmen über den Ausfallmechanismus zu Grunde, wenn auch zum Teil nur implizit. Bei der Verwendung von Available Cases (AC) oder Complete Cases (CC) geht man beispielsweise implizit von einem MCAR aus. Dagegen lässt sich auch unter der Annahme von MAR mit Multiple Imputation arbeiten. Schwieriger zu handhaben sind NMAR-Ausfälle. Durch den theoretischen Rahmen von Beatty und Herrmann lässt sich zumindest abschätzen, wie realistisch NMAR-Ausfälle sind. Dabei muss betont werden, dass das Konzept des Ausfallmechanismus zunächst theoretisch ist. In der Praxis ist es nahezu undenkbar, dass beispielsweise ein Ausfallmechanismus nur MCAR generiert. Vielmehr könnten MCAR, MAR und NMAR in der Realität großer Bevölkerungserhebungen als Kontinuum verstanden werden.[35] Das bedeutet auch, dass MAR kaum vollständig in empirischen Datensätzen gegeben ist. Es kommt in der Praxis vielmehr darauf an, dass der Ausfall hinreichend MAR ist, um z.B. mit Hilfe der Multiplen Imputation „gut genug" schätzen zu können – wenn also andere Variablen den fehlenden Wert in gewissen Umfang erklären können; das Vorhandensein dieser Variablen im konkreten Datensatz vorausgesetzt.
In der folgenden Grafik (Abbildung 8) wird für einen Fragestimulus das Entscheidungsmodell dargestellt. Um eine Brücke zu den Ausfallmechanismen zu schlagen, muss jeweils nach den Gründen und der Motivation gefragt werden, die zur Entstehung von Item Nonresponse führen. Teilweise lässt sich recht gut abschätzen, welcher Ausfallmechanismus wahrscheinlich hinter dem jeweiligen Item Nonresponse im Rahmen des Entscheidungsprozesses steht.

[34] In der Methodenforschung dreht sich ein Teil der Diskussion, ob der Fragebogen „weiß nicht"- und „keine Angabe"-Kategorien sichtbar beinhalten soll, um die von Converse (1970) entfachte Problematik der Nichteinstellung und deren Messung. Krosnick (2002) sieht beispielsweise die Funktion der Item Nonresponse in der Unterscheidung substantieller Antworten von Nichteinstellungen.

[35] Es ist plausibel anzunehmen, dass für jedes Merkmal für verschiedene Individuen der Ausfallmechanismus anders sein wird.

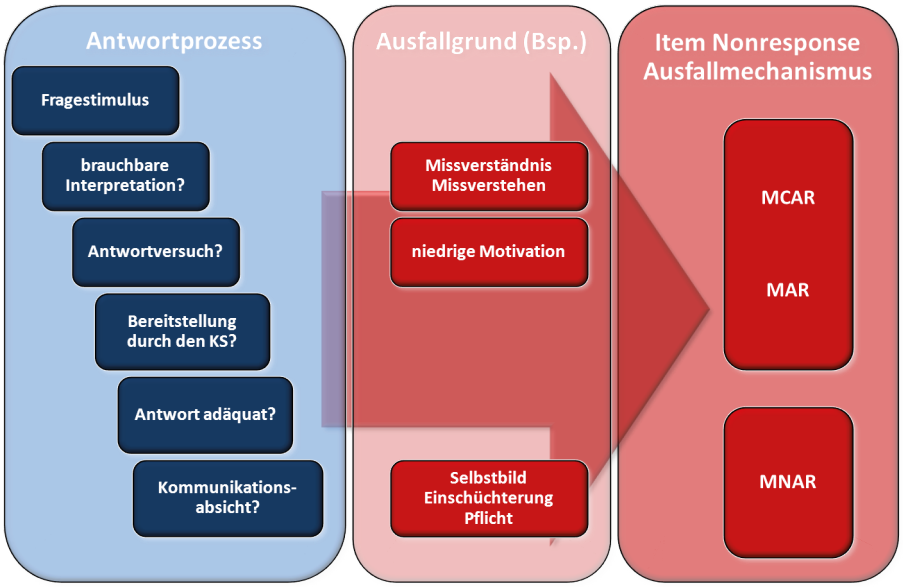



Abbildung 8: Antwortprozess, Ausfallgrund und Ausfallmechanismus

Der schnellste Weg zum Item Nonresponse führt nach der Fragestellung über den kognitiven Status der Nichtverfügbarkeit einer adäquaten Interpretation der Frage (auch: Jobe und Herrmann 1996). Schwarz et al. (2008) verdeutlichen dieses Problem des Nicht-Verständnisses: „Language comprehension is not about words per se, but about speaker meaning.“ (Schwarz et al. 2008, S.19) Tatsächlich spielt die Fragebogenkonstruktion an dieser Stelle eine entscheidende Rolle. Zwei Faktoren stechen hierbei besonders hervor. Die Formulierung der einzelnen Fragestellung kann zunächst ausschlaggebend dafür sein, ob der Befragte den Fragestimulus in der vom Fragebogenersteller gemeinten Art und Weise versteht und ob er die Frage überhaupt erfasst. Zweitens ist der Kontext der Frage wichtig, da der Befragte Informationen zum Verständnis und zur Beantwortung der Frage aus den Fragenkontext bezieht (Schwarz et al. 2008, S.21).
Hat der Befragte den Sinn der Fragestellung verstanden, muss er als weiteres den Versuch unternehmen zu antworten. Wegen niedriger Motivation kann dies ausbleiben und der Befragte zieht es aus Bequemlichkeit vor, Item Nonresponse zu generieren. Auch hier scheint der Grund für den Datenausfall nicht in der Antwort zu liegen, sondern in der aus unterschiedlichen Gründen beeinflussten Motivation des Befragten. Diese Gründe wurden in dem vorhergehenden Kapitel ausgeführt. Das bedeutet, dass auch an dieser Stelle des Entscheidungsprozesses NMAR eher nicht wahrscheinlich ist, sondern wiederum MCAR oder MAR unterstellt werden kann.
Die Hürde, die es bis zur Antwort zu überwinden gilt, wird, wie im vorhergenden Abschnitt 3.2.3 erläutert, vom kognitiven Status beeinflusst. Wenn der Befragte die Information zur Beantwortung nicht zur Verfügung hat und diese nicht einmal schätzen kann, wäre die logische Konsequenz, die Frage nicht zu beantworten. Wenn der Befragte dieses Nichtwissen mitteilen will, folgt Item Nonresponse. Die Alternative wäre eine falsche gültige Antwort zu geben, also in der Terminologie von Beatty und Herrmann einen „error of commission“ zu begehen (Beatty und Herrmann 2002, S.73). Da die Antwortausprägung, die der Befragte nicht kennt, nicht verschwiegen werden kann, lässt sich NMAR nahezu ausschließen; das bedeutet, die Unkenntnis der Antwort schützt vor „Item Non-

response“ aufgrund von NMAR. Vollkommen anders ist der Fall gelagert, wenn die Informationen zur Antwort leicht verfügbar sind. Entschließt sich hier der Befragte zur Antwortverweigerung oder gibt Nichtwissen vor, könnte NMAR vorliegen. Dies ist nicht zwangsläufig, sondern nur der Fall, wenn der Befragte aufgrund der spezifischen Antwortausprägung zum Schluss kommt, diese nicht mitzuteilen. Es gibt in der Literatur und empirischen Forschung viele Ergebnisse, die auf diesen Umstand hindeuten. Ein Beispiel, das auch für die später verwendeten Datensätze zutrifft, ist die Frage nach dem Einkommen. Die Frage wird in der Regel wohl verstanden, die Antwort lässt sich normalerweise recht einfach generieren, doch aufgrund von Furcht, Scham oder sozialer Erwünschtheit wird darauf nicht geantwortet.[36]

Weist der Befragte einen ambivalenten kognitiven Status auf, d.h. er muss mehr oder weniger Mühe für die Informationsbeschaffung aufwenden, wird nach dem Modell wiederum gefiltert, ob die Antwort auf Basis der verfügbaren Informationen als adäquat erachtet wird. Die Faktoren für diese Abwägung könnten zu einem Ausfall nach NMAR führen, da die Zurückhaltung der Antwort von ihr selbst abhängt. Die Frage ist, unter welchen Einflüssen der Filter zur Beurteilung der Antwortadäquatheit steht. An dieser Stelle scheint nur zu konkreten Fragen eine eindeutigere Aussage über den Ausfallmechnanismus möglich zu sein.

Für die vorangegangenen, aus dem Entscheidungsprozess abgeleiteten Überlegungen gilt generell, dass die Chance eines NMAR-Ausfalls mit dem Umfang der Erhebung (Anzahl der Variablen) eher sinkt. Auch wenn die Theorie Hinweise liefert, dass es sich bei der Mehrzahl der Item Nonresponse um MAR handelt, darf eine eingehende Prüfung der verwendeten Analysevariablen von Seiten des Datennutzers nicht unterbleiben. Ein kurzes Beispiel macht dies plausibel: Gerade wurde als Beispiel NMAR-Ausfälle die Einkommensvariable angeführt. Selbst wenn keine Variable ein Konstrukt wie Scham oder Soziale Erwünschtheit messen kann, kommt es z.B. für Multiple Imputation nur darauf an, ob im gesamten Datensatz Merkmale vorhanden sind, die die Einkommenshöhe erklären können; es zählt allein die Korrelation nicht unbedingt die Kausalität. Und hier stehen für dieses Beispiel in fast allen großen Erhebungen Merkmale zur Erklärung des Einkommens und zur Imputation der fehlenden Werte in dieser Variable zur Verfügung: Alter, Bildung, sozio-ökonomischer Status oder immer häufiger auch Wohnumgebung oder Zustand des Wohngebäudes. Auch hier ist der Datennutzer wiederum auf eine sorfältige und möglichst ausführliche Datendokumentation angewiesen. Dies führt bereits zum praktischen Teil des Item Nonresponse-Kapitels.

[36]Der Ausfall muss genau mit der Ausprägung der eigentlich richtigen Antwort zusammenhängen; im Falle des Einkommens würde die beispielsweise Kausalkette folgendermaßen lauten: niedriges Einkommen $->$ Scham $->$ Item Nonresponse.

## 3.3 Determinanten

Die Frage nach dem Umgang mit fehlenden Werten soll hier, soweit es möglich ist, mit realen Datensätzen oder mit Datensätzen, die auf realen Datensätzen basieren, beantwortet werden. Die grundlegenden Datensätze für die weiteren Analysen und Vergleiche von Korrekturverfahren bei Item Nonresponse sind der ALLBUS 2002, der ALLBUS 2004 und der ALLBUS 2006 (Wasmer et al. 2007; Haarmann et al. 2006; Blohm et al. 2003). Der ALLBUS ist eine der wichtigsten und umfangreichsten sozialwissenschaftlichen Erhebungen im deutschsprachigen Raum und findet in verschiedenen Disziplinen Verwendung (Koch und Wasmer 2004; Terwey 2000). Ein Teil des Themenspektrums variiert, ein Teil ist wiederkehrend.[37]
Die Analyse der Item Nonresponse erfolgt nach einigen allgemeinen Betrachtungen auf Grundlage einer modifzierten Poissonregression. Zuvor wird ausgeführt, warum sich Item Nonresponse als Zähldaten interpretieren lassen und warum eine einfache Poissonregression, oder gar eine normale OLS-Schätzung, für eine Analyse ungeeignet erscheinen. Neben dem explorativen Charakter des Modells, das vor allem Interviewereinfluss und situative wie persönliche Merkmale des Befragten enthält, sollen hier Variablen gefunden werden, die für Item Nonresponse Erklärungspotential besitzen. Diese Informationen werden in den späteren Vergleichen von Korrekturmethoden herangezogen. Basis der Vergleiche bilden statistische Analysen folgender Veröffentlichungen:

- Hennig, E. (2009) „Einen Schlussstrich unter die nationalsozialistische Vergangenheit ziehen“ Zur politischen Soziologie eines historischen Deutungsmusters, Einsicht, 2, S.42-49.

- Schnabel, C. und J. Wagner (2005) Who Are the Workers Who Never Joined a Union? Empirical Evidence from Germany, IZA, Paper 1698.

- Schäfer, A. (2009) Alles halb so schlimm? Warum eine sinkende Wahlbeteiligung der Demokratie schadet, MPIfG Jahrbuch, S.5-10.

Hennig (2009) untersucht mit Hilfe bivariater Analysen und Kontrastgruppen die Einstellungen der Gruppe von Befragten, die für die Beendigung der Diskussion über die nationalsozialistische Vergangenheit in Deutschland sind. Hierfür benutzt er den ALLBUS 2006. Schnabel und Wagner (2005) verwenden für ihre Analyse von Arbeitnehmern, die niemals gewerkschaftlich organisiert waren, den ALLBUS 2002 und den European Social Survey 2002/2003, da in diesem Datensatz noch einige andere Untersuchungsvariablen zur Verfügung standen. Das Zentrum der Analyse bildet ein multivariates Probitmodell. Anhand der Daten des ALLBUS 2004 geht Schäfer (2009) der Frage nach, ob eine sinkende Wahlbeteiligung der Demokratie schadet. Dabei schätzt er Wahrscheinlichkeiten für die Teilnahme an verschiedenen Partizipationsformen auf Grundlage eines multivariaten Logitmodells für Individuen mit definierten sozio-ökonomischen Merkmalen. Nach einer kurzen Vorstellung der Aufsätze und der geschätzten Parameter wird jeweils die zentrale Analyse oder der wichtigste Teil der Analysen exakt repliziert.[38]

[37] Allgemeine Informationen über Zweck und Aufbau des ALLBUS befinden sich auf der GESIS Homepage: http://www.gesis.org/allbus/allgemeine-informationen/c5425, abgerufen am 22.8.2011.

[38] Die Replikation wurde mit freundlicher Unterstützung der genannten Autoren durchgeführt.

### 3.3.1 Analyse von Item Nonresponse

Datengrundlage sind speziell die ALLBUS 2002, 2004 und 2006. Die Datensätze enthalten folgende Anzahl von Variablen und Befragten:

| | ALLBUS 2002 | ALLBUS 2004 | ALLBUS 2006 |
|---|---|---|---|
| Erhebungszeitraum | Feb.-Aug. 2002 | März-Juli 2004 | März-Aug. 2006 |
| Befragte | 2.820 | 2.946 | 3.421 |
| Variablen | 722 | 895 | 743 |

Tabelle 1: Verwendete ALLBUS-Erhebungen

Diese ALLBUS besitzen in der Ausführung nahezu identische **Survey Lifecycles**. Betrachtet man die kritischen Einfallspunkte für Item Nonresponse, lässt sich für diese Erhebungen zusammenfassen: Alle ALLBUS enthalten sensible **Fragestellungen**, die neben als unproblematisch eingestuften Items im Fragebogen verstreut liegen. Die ausgewählten Veröffentlichungen beinhalten ebenfalls eine mehr oder weniger große Anzahl sensibler Items (im Detail hierzu Abschnitt 4.2.1 bis 4.2.3). Da die **Zielgesamtheit** alle erwachsenen Personen (Deutsche und Ausländer) in Privathaushalten umfasst, ergibt sich keine Problematisierung der Zielgesamtheit im Vergleich mit anderen großen Bevölkerungsstichproben (Koch und Wasmer 2004, S.29ff). Angemerkt sei aber, dass die Einstufung der Sprachfähigkeit subjektiv von den Interviewern durchgeführt wurde. Für die replizierten Analysen haben sie teilweise keine Bedeutung, weil bestimmte Items sowieso als Missing by Design definiert werden. ALLBUS 2002, 2004 und 2006 wurden mit dem gleichen **Erhebungsmodus** durchgeführt, nämlich als Computer Assisted Personal Interviews (CAPI) bzw. Computer Assisted Selfadministered Interviews (CASI) durchgeführt (Koch und Wasmer 2004, S.28f).[39] In der **Interviewsituation** unterlagen alle drei Erhebungen denselben Einflüssen und Kriterien, auch stehen für alle drei dieselben Analysevariablen für die spätere Analyse der Item Nonresponse-Faktoren zur Verfügung.[40] Die Datensätze wurden jeweils **editiert** und sind regulär als Nettostichprobe[41] für die Datennutzer zugänglich, wobei der allergrößte Teil der Item Nonresponse systemfehlend definiert wird.[42] Es gibt keine Anwendung von **Korrekturmethoden** für

[39] Der Einflusskomplex Messung soll jeweils für die replizierten Analyse im Einzelnen diskutiert werden.

[40] Mit der Feldarbeit wurden jedoch unterschiedliche Institute beauftragt: 2002 infas, 2004 und 2006 TNS-Infratest.

[41] Zum Begriff der Nettostichprobe siehe Abschnitt 5.2.

[42] Inwieweit von der Editierung Item Nonresponse betroffen ist, wird nicht detailliert dokumentiert.

den ALLBUS. Damit kann der jeweiligen Nutzer selbst über die Möglichkeit entscheiden, ob und wie er Item Nonresponse korrigieren will.[43]

Zunächst soll ein genauerer Blick auf die Antwortkategorien geworfen werden, die schon als Missing definiert wurden. Es spiegelt sich darin die Vielzahl von unterschiedlichen Item Nonresponse-Definitionen, wie sie in Abschnitt 2.1 und 2.2 diskutiert wurden. Üblicherweise sind folgende Ausprägungen als Item Nonresponse definiert, ehe der Datennutzer eigene Definitionen trifft:

1. „keine Angabe", „verweigert": das bedeutet, dass der Befragte die Antwort verweigert hat oder keine Angaben machen wollte.

2. „weiß nicht": zusätzlich ist bei den meisten, jedoch nicht bei allen Variablen, diese Kategorie ausgewiesen. Der Befragte weiß keine Antwort oder gibt vor, keine zu wissen.

3. „trifft nicht zu": im strengen Sinn ist diese Ausprägung kein Item Nonresponse, da sie als Folge von Filtersetzung anzusehen ist (Missing by Design). Dem Befragten wurde diese Frage niemals gestellt. Beispielsweise werden nicht-deutschen Staatsbürgern einige Items zu politischen Einstellungen nicht vorgelegt.

4. Fragespezifische Residualkategorien: „Keines der Gebiete", „nicht Deutscher" usw.: auch hier lässt sich bei einer Vielzahl von Fällen darüber streiten, ob tatsächlich Item Nonresponse vorliegt. Beispielsweise ist die Ausprägung „nicht Deutscher" eine allgemein zulässige, informationstragende Ausprägung. Sie wurde bei der Editierung nur als fehlender Wert definiert und in der Regel mit „Trifft nicht zu" (TNZ) bezeichnet.

Schließt man nun die Kategorien TNZ aus und editiert mit genügend Sensibilität alle Item Nonresponse-Kategorien stringent, so ergibt sich eine eindeutige Reihenfolge, die auch Longford benutzt (Longford 2000, S.75), im Umfang der Ausfälle nach Itemarten. Sogenannte nicht-reaktive Daten (z.B. Ost-West, deutsche Staatsangehörigkeit, Zustand des Wohngebäudes, Geschlecht, vom Interviewer ausgefüllte Variablen über Bereitschaft, Zuverlässigkeit und Erreichbarkeit) weisen keine oder nur sehr wenige fehlende Werte auf. Geringfügig höher liegt der Wert der Ausfälle bei den meisten soziodemografischen Variablen (z.B. Alter mit 0,3 %, allgemeiner Schulabschluss mit 0,2 %, Gesundheitszustand mit 0,2 %, Anzahl der im Haushalt lebenden Personen mit 0,7 %). Dabei bildet die Fragen nach dem Einkommen (bis zu einem Drittel) und beispielsweise der früheren Gewerkschaftsmitgliedschaft mit 12,3 % Ausnahmen.[44]
Einstellungs- und Meinungsvariablen weisen im Durchschnitt die höchste Item Nonresponse-Rate auf. Innerhalb der Einstellungs- und Meinungsvariablen weicht die Höhe der Item Nonresponse-Rate nach Fragethema beträchtlich voneinander ab:

[43] Alle Angaben sind aus den jeweiligen Methodenreporten entnommen, vgl. Wasmer et al. (2007); Haarmann et al. (2006); Blohm et al. (2003).

[44] Da die Erhebungen ALLBUS 2002, 2004 und 2006 methodisch nahezu identisch abgelaufen sind, wird für die Modellschätzung stellvertretend auf den ALLBUS 2006 zurückgegriffen und die weitere Item Nonresponse-Analyse auf diesen Datensatz beschränkt.

- Abtreibung: 4,6 % - 6,6 %
- Einstellungen gegenüber Ausländern: 1,3 % - 6,2 %
- Antisemitismus: 8,7 % - 13,7 %
- Beschreibung der eigenen Persönlichkeit: 1,1 % - 2,6 %
- höchster Ausfall: Sonntagsfrage (23,0 %), Linksrechtsselbsteinstufung (10,0 %)

Der ALLBUS 2006 besitzt ohne ISSP-Fragebogen 743 Variablen.[45] Nach Abzug der Ableitungen oder Derivate von Variablen sowie der nicht-reaktiven Variablen bleiben 226 Variablen übrig (Anhang 1). Diese ausgewählten Items weisen jedoch nicht alle fehlende Werte auf. Zur Konstruktion einer Item Nonresponse-Variablen wurden die 226 Variablen umkodiert und daraus eine Indikatormatrix erstellt, in der bei fehlendem Wert („weiß nicht“, „keine Angabe“) der Zelle der Wert 1, bei allen anderen Angaben der Wert 0 (auch: „trifft nicht zu“, da dies kein fehlender Wert ist, der vom Befragten verursacht wird) zugewiesen wurde. Jede Zelle dieser Indikatorenmatrix gibt Auskunft über den Ort (welche Variable im Interview) und über den Merkmalsträger, der im Sinne der Definition einen fehlenden Wert aufweist (Little und Rubin 2002; Loosveldt et al. 1998). An dieser Stelle findet eine Verdichtung der Daten statt. Aus der Matrix wird ein Vektor, dessen Elemente nur noch die Anzahl der Item Nonresponse für den i-ten Befragten angibt. Die Reihenfolge der Ereignisse Item Nonrepsonse ist damit nicht mehr bestimmbar und lässt sich dann nur durch eine ebensolche Verdichtung in die andere Dimension als Zeitreihe analysieren. Abbildung 9 veranschaulicht das Verfahren:

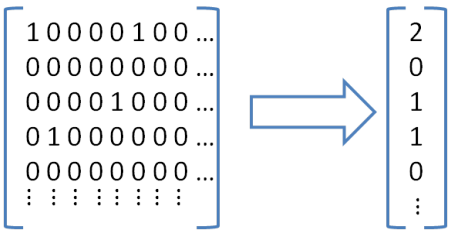



Abbildung 9: Erzeugung des Vektors mit Item Nonreponse

Der Vektor „Anzahl der Item Nonresponse“ lässt sich weiterhin zu einer Häufigkeitsverteilung verdichten und in einem Bloxplot beschreiben:

Das Stabdiagramm zeigt deutlich eine Verteilung, wie sie typisch sein dürfte für Item Nonresponse in einer großen Bevölkerungsstichprobe. Mehr als ein Drittel der im ALLBUS 2006 befragten Personen weist keinerlei Item Nonresponse auf, dies ist auch der Modus der Verteilung. Die Verteilung ist damit rechtsschief und linkssteil. Knapp zwei Drittel der Befragten haben mindestens

[45] Das International Social Survey Panel ist eine internationale Erhebung, die dem ALLBUS angefügt ist.

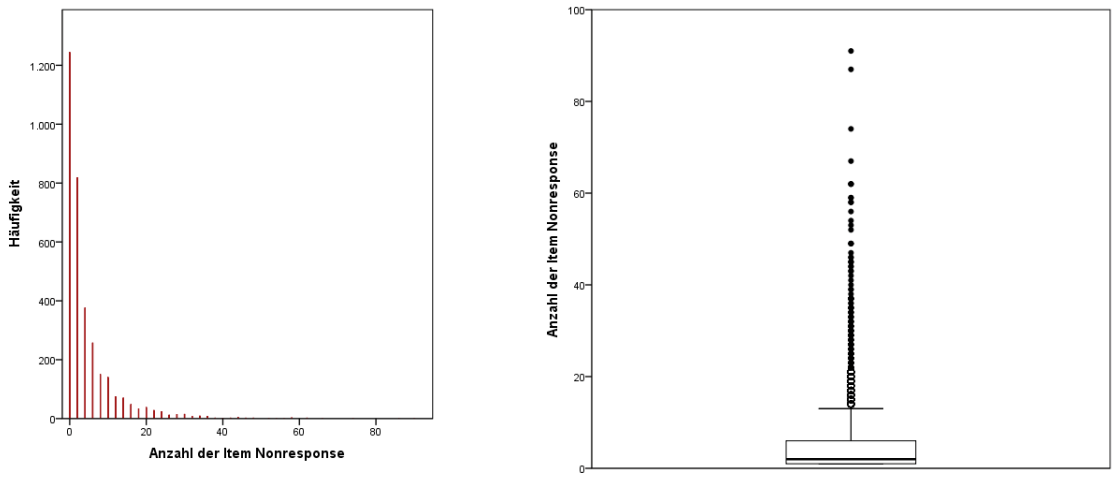



Abbildung 10: Häufigkeitsverteilung und Boxplot für die Anzahl der Item Nonresponse

eine Frage nicht beantwortet. 39 % haben auf eine bis fünf Fragen nicht geantwortet. Etwa 12 % haben zwischen sechs und 20 Fragen unbeantwortet gelassen. Immerhin gut 12 % haben mehr als zehn Fragen offen gelassen oder verweigert, darauf zu antworten. Mit großem Abstand und 171 fehlenden Antworten gibt es einen Ausreißer (der jedoch für die Analyse zusammen mit einem weiteren ausgeschlossen wurde). Die restlichen Befragten haben alle unter 100 Item Nonresponse, was allerdings bei 226 Fragen eine Nichtbeantwortungsquote von knapp 45 % bedeutet. Im Boxplot sticht ebenfalls die Schiefe der Verteilung hervor: Im Mittel hat jeder Befragte etwa 4,48 Item Nonresponse im Datensatz hinterlassen, der Median liegt hingegen bei 1. Die Streuung erscheint mit einer Varianz von 62,77 enorm.
Damit stellt die Item Nonresponse-Verteilung für die weiteren multivariaten Analysen durchaus eine Herausforderung dar.

### 3.3.2 Item Nonresponse als Zähldaten

Ein kurzer Blick auf die oben gezeigte diskrete Verteilung führt zur Annahme, dass Item Nonresponse in der so aggregierten Form als Zähldaten interpretiert werden können (Loosveldt et al. 1998).
Charakteristisch für Zähldaten ist das Auftreten von (seltenen) zufälligen Ereignissen in einem räumlich-zeitlichen Kontinuum (Raabe-Hesketh und Skrondal 2008, S.373ff; Winkelmann 2008, S.7). Ein solches Ereignis kann das Ausschlagen eines Pferdes und die damit verbundene Verletzung eines Soldaten sein (Abbildung 11), wie es Borkiewicz 1898 untersucht hat (Raabe-Hesketh und Skrondal 2008, S.373), oder eben das Auftreten von Item Nonresponse.

Abbildung 11: Label von Stata Press: Ausschlag eines Pferdes

Damit verbunden ist die Vermutung, dass der sich daraus ergebenden Verteilung ein stochastischer Prozess, „ein count process“ innerhalb eines (stetigen) Zeitraums, zu Grunde liegt (Winkelmann 2008, S.8).[46] Dabei kann das Interview als jenes räumlich-zeitliche Kontinuum aufgefasst werden. Die Zähldaten sind damit das Ergebnis dieses Prozesses. So können wir eine Zufallsvariable folgendermaßen definieren:

$X_i$= Anzahl von auftretenden Item Nonresponse im i-ten Interview

Zwei Verteilungen werden in der Praxis zur Modellierung von Count-Data am häufigsten verwendet: die Poissonverteilung und die Negative Binomialverteilung (Hilbe 2007, S.8f).
Die Poissonverteilung besitzt nur einen Parameter, so dass die Zufallsvariable $X_i$ einer Poissonverteilung mit $\lambda$ folgen würde. Entsprechend böte sich eine Poissonregression für ein Modell zur Erklärung von Item Nonresponse als Standardverfahren bei Count Data an.[47] Folgende drei Annahmen werden für die *univariate Poissonregression* (Poi) getroffen:

**Annahme 1**

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \text{mit } y = 0,1,2,\ldots \tag{1}$$

wobei $f(y|\lambda)$ die bedingte Wahrscheinlichkeitsfunktion von $y$ gegeben $\lambda$ ist. Bedingung ist, dass $\lambda > 0$ ist.

[46] Für weitere Eigenschaften siehe ebenfalls Winkelmann (2008), S.8ff; dort findet sich auch die Diskussion zu Alternativen wie der Negativen Binomialverteilung.

[47] Winkelmann merkt an, dass Nicht-Count-Data keine Notwendigkeit für die Poissonregression darstellen, vgl. Winkelmann (2008), S.65.

**Annahme 2**

$$\lambda = \exp(x'\beta) \tag{2}$$

wobei $\beta$ die Parameter darstellen und $x'$ die unabhängigen Variablen. In der Regel beinhalten sie auch eine Konstante.

**Annahme 3**

Man geht dann von Beobachtungspaaren $(x_i, y_i)$ mit $i = 1, \ldots, n$, die unabhängig verteilt sind, aus.

Annahmen 1 und 2 lassen sich zu einer bedingten Wahrscheinlichkeitsfunktion zusammenfassen:

$$f(y|\lambda) = \frac{\exp(-\exp(x'\beta))\exp(yx'\beta)}{y!} \text{mit } y = 0, 1, 2, \ldots \tag{3}$$

So hat die Poissonverteilung lediglich einen Parameter für den bedingten Erwartungswert und die Varianz:

$$E(Y_i|X = x_i) = \lambda = \exp(x'\beta) \tag{4}$$

$$Var(Y_i|X = x_i) = \lambda = \exp(x'\beta) \tag{5}$$

Das heißt, dass der Erwartungswert (4) gleich der Streuung (5) ist.[48] Zusammen mit der 3. Annahme lässt sich eine Maximum-Likelihood-Schätzung durchführen. Winkelmann betont zudem, dass selbst im Falle einer fehlenden oder unplausiblen stochastischen Unabhängigkeit der Ereignisse die Poissonregression eine robuste Schätzung der Parameter erlaubt.[49] Hilbe (2007) nennt insgesamt 7 Verletzungen von Verteilungsannahmen, die sowohl bei Verwendung einer Poissonverteilung als auch einer negativen Binomialverteilung in der Praxis auftreten können. So können

[48] Equidisperson, vgl. Winkelmann (2008), S.64.

[49] Winkelmann zeigt auch die Probleme mit anderen Regressionsmodellen, die bei Zähldaten teilweise verzerrt, in der Regel aber nicht effizient sind, vgl. Winkelmann (2008)

überhaupt keine Nullen gezählt werden, ein Überschuss an Nullen kann vorkommen, die Daten weisen eine klare Mischverteilung auf, die Daten sind zensiert oder gestutzt, die Datenstruktur entstammt einem Panel, sind geclustert oder liegen im Längsschnitt vor, oder die Treffer basieren auf dem Wert einer anderen Variablen (Hilbe 2007, S.11ff).
Hinzu kommt noch folgendes Problem: normalerweise wird für Poissonverteilungen angenommen, das Eintreten eines Ereignisses geschehe für alle unabhängigen Beobachtungen innerhalb eines gleichen Zeitraumes (auch als Period at Risk bezeichnet, Winkelmann 2008, S.74).[50] Nun lässt sich fragen, ob das Interview im Zuge des ALLBUS als zeitliches Kontinuum interpretierbar ist. Die Anzahl der gestellten Stimuli ist tatsächlich für alle Befragte ähnlich.[51] Die Zeit, die dabei aber verstreicht ist sehr unterschiedlich. Der ALLBUS 2006 beinhaltet sowohl ein Interview der Länge 20 Minuten als auch ein Interview der Länge 180 Minuten; das heißt, die Ausdehnung der Zeitperiode ist nicht $t$ konstant, sondern für jeden Interviewer mehr oder weniger unterschiedlich. Für die Lösung des Problems kann beispielsweise die Annahme der Proportionalität eingeführt werden; dann wird der bedingte Erwartungswert wird mit einem „Offset“ versehen (Winkelmann 2008, S.78; Raabe-Hesketh und Skrondal 2008, S.376):

$$E(y|x) = t\exp(x'\beta) = \exp(x'\beta + \log t) \tag{6}$$

Ein spezifisches Problem der Poissonverteilung erwächst aus der Equidispersion (Winkelmann 2008, S.45ff). Die empirischen Daten weichen davon sehr oft ab. Die negative Binomialverteilung besitzt dagegen einen Mittelwertparameter und einen Parameter für die Anzahl der Treffer (Hilbe 2007, S.10) und gilt in der Literatur häufig als erste Alternative zur Poissonverteilung.
McCullagh und Nelder (McCullagh und Nelder 1989) bezeichnen die Vergrößerung der Varianz gegenüber dem Erwartungswert als Overdisperion oder Extra-Poisson-Variation (auch Winkelmann 2008, S.110).[52] Die Parameterschätzung der Poissonverteilung $\lambda_i$ wird insuffizient (Böhning et al. 1999, S.17; Pickery et al. 1998, S.34f). Eine Ursache von Overdispersion ist häufig nicht beobachtete Heterogenität der Grundgesamtheit, beispielsweise weil Variablen nicht beobachtet werden oder latent wirken (Winkelmann 2008, S.103ff).[53] Für die Schätzung bedeutet das größere Standardfehler und damit insignifikante Parameter.

Mit Blick auf die Visualisierung der Item Nonresponse-Verteilung in Abschnitt 3.3.1, lässt sich ein deutlicher Überschuss an Nullen erkennen (Winkelmann 2008, S.109). Rechnerisch ist die Abweichung zwischen der vorliegenden Verteilung und einer angepassten Poissonverteilung erheblich. Während bei eine Poissonverteilung mit $\lambda_i = 4,48$ 39 Personen keine Item Nonresponse besitzen dürften, sind es bei der konkreten Item Nonresponse-Verteilung jedoch 1.245. Wie oben bereits diskutiert wurde, stellt eine Null-Inflation sowohl für die Poisson- als auch für die negative

[50] Auch für die negative Binomialverteilung wird dies angenommen, vgl. Hilbe (2007), S.9.

[51] Ausnahmen bilden bei einer derart komplexen Erhebung die Filterfragen, die jedoch bei der Auswahl der Variablen für den Item Nonresponse-Vektor nach Möglichkeit vermieden wurden.

[52] Underdispersion tritt dagegen auf, wenn die Varianz kleiner als der Mittelwert ist, was jedoch in der Praxis wesentlich seltener der Fall ist, vgl. Hilbe (2007), S.177.

[53] Eine andere Ursache ist eine Variation des zeitlichen Kontinuums oder die Tatsache, dass die Ereignisse nicht konstant auftreten, vgl. Barron (1992), S.185f.

Binomialverteilung eine Annahmeverletzung dar. Beide Verteilungen können aber entsprechend erweitert werden.[54]
Eine Mischverteilung löst dieses Problem dadurch, dass es geteilt wird. So nutzen Böhning et al. (1999) eine Mischung aus einer Zwei-Massen-Verteilung gegeben $\omega - 1$ für den Wert 0 sowie gegeben $\omega$ für die anderen Werte mit dem Parameter $\lambda$.[55]

$$f(y_i|\lambda,\omega) = \begin{array}{ll} \omega + (1-\omega)\exp(-\lambda), & \text{wenn } y = 0 \\ (1-\omega)\exp^{-\lambda}\frac{\lambda^y}{y!}, & \text{wenn } y > 0 \end{array} . \tag{7}$$

wobei dann Erwartungswert und Varianz wie folgt berechnet werden:

$$E(Y_i|X = x_i) = \lambda \tag{8}$$

$$Var(Y_i|X = x_i) = \lambda + \lambda^2 \tag{9}$$

Diese häufig für medizinische Analysen verwendete Verteilung (7) wird Zero-Inflated Poissonmodell (ZIP) genannt (Böhning et al. 1999).[56] Eine Erweiterung und Systematisierung stellen Czado et al. (2007) vor. Hier wird nicht nur der Fall von Nullüberschuss durch den Parameter $\omega$, sondern auch ein Parameter $\varphi$ für die Überdispersion berücksichtigt:[57]

$$f(y_i|\lambda,\omega,\varphi) = \begin{array}{ll} [\omega + (1-\omega)exp(-\lambda)] & \text{wenn } y = 0 \\ + \left[(1-\omega)\frac{\lambda(\lambda+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}\exp^{-}\frac{1}{\varphi}(\lambda+(\varphi-1)y)\right] & \text{wenn } y > 0 \end{array} \tag{10}$$

Leicht zu erkennen ist, dass sich diese Dichte zur Verteilung der Poi (1) reduziert, wenn $\varphi = 1$ und $\omega = 0$, zur Generalized Poissonregression (GP), wenn $\omega = 0$ und schließlich zur ZIP-Regression, wenn $\varphi = 1$. Die Systematisierung der gesamten Klasse lässt sich folgendermaßen veranschaulichen:

[54] Insgesamt zählt Hilbe (2007) 22 Modellvariationen für negative Binomialverteilungen auf, die jeweils den Umgang mit einer oder mehreren Annahmenverletzungen erleichtern, vgl. Hilbe (2007), S.78.

[55] Dies ist der nicht-parametrische Fall für die Schätzung der Heterogenität und damit der einfachste Fall, vgl. Böhning et al. (1999), S.18.

[56] Für die Eigenschaften der ML-Schätzung für ZIP-Regression vgl. Lambert (1992), S.6f; und die Vorteile gegenüber der Negativen Binomialverteilung, S.10f.

[57] Zu den Restriktionen folgender Wahrscheinlichkeitsdichte siehe Czado et al. (2007), S.3.

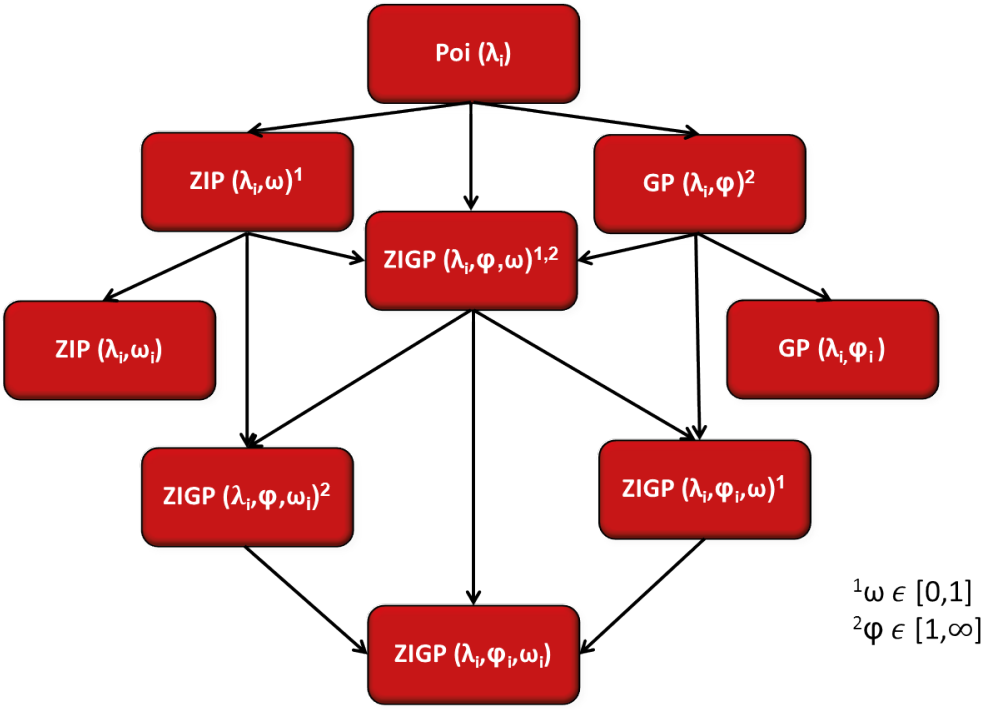



Abbildung 12: Familie der Poissonmodelle nach Czado et al. (2007)

Erwartungswert und Varianz für die Zero-Inflated Generalized Poissonregression (ZIGP) sind:

$$E(Y_i|X = x_i) = (1-\omega)\lambda_i \tag{11}$$

$$Var(Y_i|X = x_i) = (1-\omega)\lambda_i(\varphi^2 + \lambda_i\omega) \tag{12}$$

Abbildung 12 verdeutlicht, dass sich die Poissonregression (Poi) in die Zero-Inflated (ZIP) und die Generalized Poissonregression (GP) erweitern lässt; genau betrachtet ist die Poissonregression der eigentliche Spezialfall. Hier wird, wie oben bereits erwähnt, nicht nur der Parameter $\lambda$, sondern auch $\omega$ und $\varphi$ geschätzt. In der GP kann die Overdispersion durch einen zusätzlichen Paramter abgebildet werden. Ebenso lässt sich der Nullenüberschuss durch einen weiteren Parameter ausdrücken. Die ZIGP vereint beides bzw. ermöglicht die Regression auf Erwartungwert, Nullen und die Overdispersion (Consul 1989).[58]
Schaut man sich die Verteilung der Item Nonresponse für die ausgewählten Items des ALLBUS 2006 im Hinblick auf die Parameter an, so wurde bereits das Phänomen der Null-Inflation und der Überdispersion konstatiert, deren Auftreten im Übrigen auch bedingt sein kann (Hilbe 2007, S.62).

[58] In Abbildung 12 sind jeweils auch die Varianten dargestellt, die die Regression auf $\omega$ und $\varphi$ zulassen.

Im folgenden Abschnitt 3.3.3 soll die Fülle der Varianten, Zähldaten zu modellieren, ausgenutzt werden. Dabei werden geschätzt

- eine OLS-Regression (nicht logtransformiert),
- eine ZIP$(\lambda_i, \omega)$,
- eine ZIGP$(\lambda_i, \omega, \varphi)$ sowie
- eine Negative Binomialregression (NB),[59]
- eine Zero-Inflated Negative Binomialregression (ZNB).[60]

Auf die Familie der Negativen Binomialregressionen soll nicht mehr weiter eingegangen werden, da der Fokus in dieser Arbeit auf Item Nonresponse als Zähldaten liegt.

[59] Für Wahrscheinlichkeitsdichte, Erwartungswert und Varianz siehe Hilbe (2007), S.79ff.
[60] Für Wahrscheinlichkeitsdichte, Erwartungswert und Varianz siehe Hilbe (2007), S.160ff.

### 3.3.3 Erklärungsmodell für Item Nonresponse im ALLBUS 2006

Aufgrund des ähnlichen Aufbaus und der geringen Abweichung zwischen den ALLBUS 2002, 2004 und 2006 wurde der ALLBUS 2006 für die multivariate Analyse ausgewählt. Der Vektor, in dem jeder Befragte die Anzahl der Item Nonresponse zugewiesen bekommt, dient als abhängige Variable und wurde bereits im vorgehenden Kapitel beschrieben. Zusammen mit der Theorie für Item Nonresponse, wie sie in Abschnitt 3.2 beschrieben wurde, lässt sich eine Vielzahl von möglichen Faktoren, die Einfluss auf die Anzahl der Item Nonresponse haben könnten, identifizieren und durch entsprechende Variablen aus dem ALLBUS 2006 operationalisieren. Da das Interview eine komplexe soziale Interviewsituation darstellt, sollte nicht nur auf potentielle Merkmale der Interviewer geachtet werden, sondern auch auf Variablen der Befragten sowie auf interagierende Merkmale zurückgegriffen und die Situation bzw. Umgebungsfaktoren soweit wie möglich abgebildet werden. In der Literatur findet sich eine Vielzahl von Variablen, die zur Analyse von Item Nonresponse benutzt wurden. Die Prämisse für die Auswahl der Regressoren besteht darin, dass die Items selbst möglichst keine Item Nonresponse aufweisen. Im ALLBUS 2006 ist für die in Frage kommenden Merkmale diese Prämisse fast ausnahmslos erfüllt. Zunächst werden kurz die Variablen zur Operationalisierung vorgestellt. Aus Gründen der Übersichtlichkeit werden die Hypothesen gleich im Anschluss an die Vorstellung angefügt. Die Einführung der Regressoren erfolgt in der Reihenfolge: **Befragtenmerkmale**, Items zu **Umgebung und Interviewsituation**, **Interviewermerkmale** und **Interviewer-Befragter-Merkmale**.[61]

Wie aus der Theorie nach Beatty und Herrmann hervorgeht, stellen kognitive Fähigkeiten und Zustände Faktoren für das Auftreten von Item Nonresponse dar (Alwin und Krosnick 1991; bei Loosveldt 1998 getrennt nach Fähigkeit und Motivation; Schuhman und Presser 1981; Thiessen und Blasius 1998). Mit dem Alter des Befragten lässt sich von einer sinkenden Fähigkeit z.B. des Erinnerungsvermögens ausgehen (Loosveldt 1998). Als weiteres persönliches Befragtenmerkmal wird Bildung herangezogen. Der höchste allgemeine Bildungsabschluss operationalisiert die kognitiven Fähigkeiten – wenngleich auch nur in engen Grenzen (Krosnick 2002; Loosveldt und Loosveldt 1997; Beatty und Herrmann 2002; Thiessen und Blasius 1998). Eine weitere Variable, die zur Operationalisierung physischer und psychischer Belastbarkeit dienen kann, ist der allgemeine Gesundheitszustand (Loosveldt 1998).

[61] Soweit es durch die empirischen Daten möglich war, wurden Merkmale entsprechend der Begriffe aus dem Stimulus-Personen-Responsemodell und der Theorie des Kognitiven Status gewählt. Die zur Analyse ausgewählten Variablen decken zum großen Teil die als Reize bezeichnen Variablen ab; ergänzt werden diese Prädiktoren mit Befragtenmerkmalen.

*Hypothese 1*: *Bildung* gilt als „proxy indicators of motivation, cognitive ability, and communicative skills“ (Loosveldt 1997, S.387). Je höher das *Alter* und je niedriger das *Bildungsniveau*, desto höher die Anzahl der Item Nonresponse.

*Hypothese 2*: Zudem ist es vorstellbar, dass ein schlechter *Gesundheitszustand* sich sowohl auf die Fähigkeiten zur Antwort als auch auf die allgemeine Motivation negativ auswirkt und somit die Wahrscheinlichkeit von Item Nonresponse erhöht.
Die Antworten auf das Item *Gesundheitszustand* leiden zwar kaum unter Item Nonresponse, das körperliche und geistige Befinden zählt jedoch zu den sensiblen Fragestellungen, weshalb mit sozial erwünschten Antworten gerechnet werden muss (Lensveldt-Mulders 2008, S.262).

Für die Hypothese nach dem Einfluss des *Befragtengeschlechts* sind weder die theoretischen Fragmente in der Literatur noch die empirischen Hinweise eindeutig.

*Hypothese 3*: Nach Thiessen und Blasius (1998) antworten Frauen häufiger als Männer nicht auf Fragen.

Die vom Interviewer vermerkte Variable *Zuverlässigkeit des Antwortverhaltens* (Beatty und Herrmann 2002) soll Hinweise auf mögliche Unsicherheiten und Zögern des Befragten geben: In der Antwortgeschwindigkeit könnte auf die „Richtigkeit“ des Antwortprozess geschlossen werden (Krosnick 2002; Norman 1982). Speziell als Proxyvariable kann das Merkmal *Antwortbereitschaft* verwendet werden. Auch dieses Merkmal wird vom Interviewer erhoben und hält zumindest den subjektiven Eindruck des Interviewers von der Motivationslage des Befragten fest (Beatty und Herrmann 2002).

*Hypothese 4*: Höhere *Zuverlässigkeit* und höhere *Antwortbereitschaft* während des Interviews wirken sich negativ auf Item Nonresponse aus.

Mit der Aufteilung des Erhebungsgebietes in *Ost- und Westdeutschland* stellt die Bundesrepublik einen kulturellen Sonderfall dar. Verweisend auf Johnson et al. (2002) lassen sich auch für das Auftreten von Item Nonresponse kulturelle Unterschiede im Antwortverhalten vermuten. Gerade im Falle der beiden Landesteile könnte von einer deutlich unterschiedlichen Kultur des Berichtens ausgegangen werden. Den kulturellen Sozialisationsunterschieden wird durch die Aufschlüsselung nach Ost- und Westdeutschland Rechnung getragen.

*Hypothese 5*: Die investigativen Züge des DDR-Systems in *Ostdeutschland* vor 1990 dürfte die allgemeine Antwortbereitschaft senken (neben der konkreten Antwortbereitschaft (Beatty und Herrmann 2002)) und sich in zunehmendem Auftreten von Item Nonresponse niederschlagen.

Die Variablen *Bereitschaft* und die *Erreichbarkeit* des Befragten werden zumeist im Kontext der Analyse von Unit Nonresponse gebraucht (Pötschke und Müller 2006). Beide können aber auch als Faktoren für Item Nonresponse in Frage kommen. Die *allgemeine Bereitschaft*, an der Befragung teilzunehmen, lässt sich durchaus als Indikator für die Motivation im Interview heranziehen. Die *Erreichbarkeit* kann zumindest indirekt über den Stressfaktor des Befragten eine Aussage machen (Stoop 2007).

*Hypothese 6*:Wer gut zu erreichen ist, könnte weniger unter Stress leiden und ist wohl eher geneigt, die Fragen im Interview sorgsam zu beantworten.

Als potentielle Proxyvariable für Einkommen sind der *Zustand des Wohngebäudes* und die *Einschätzung der Wohnumgebung* zu nennen.[62]

*Hypothese 7*: Die beiden Variablen *Zustand des Wohngebäudes* und die *Einschätzung der Wohnumgebung* müssten als Indikatoren des sozio-ökonomischen Status bei steigenden Ausprägungen zu geringerer Item Nonresponse-Wahrscheinlichkeit führen.

Ein weiteres Merkmal, das bei Unit Nonresponse-Analysen zum Tragen kommt, ist die *Größe der Gemeinde*, in der der Befragte wohnt. Auch für die Anzahl der Item Nonresponse könnte dies entscheidend sein, wenn der Befragte normalerweise in einem urbanen, anonymen Umfeld wohnt.

*Hypothese 8*: Mit zunehmender *Gemeindegröße* wächst die Anzahl der Item Nonresponse aufgrund der Sorge um die Weitergabe von Informationen an Fremde und den Schutz des Persönlichen in einer zunehmend anonymen Umgebung.

In der Literatur richtet sich auch immer wieder das Augenmerk auf die *Anwesenheit* oder sogar den *Eingriff Dritter* in das Interview, wobei die Effekte durchaus widersprüchlich sind (Haunberger 2006 vs. Reuband 1984). Die Situation des Interviews kann aus dem ALLBUS 2006 mit den Variablen *Anwesenheit Dritter* und dem möglichen *Eingriff Dritter* teilweise operationalisiert werden. Diese Variablen geben Auskunft ob ein Dritter während des Interviews anwesend war und in welchem Grad die anwesende Person ins Interview eingegriffen hat.

[62]Die Einkommensvariable findet sich zwar in der Regel bei Analysen zu Unit Nonresponse, dennoch erscheint ein abweichendes Antwortverhalten bezüglich Item Nonresponse bei sozial schlechter Gestellten als denkbar, vgl. Goyder (1987), S.84f.

*Hypothese 9*:*Anwesenheit* und erst recht der *Eingriff Dritter* in das Interview dürfte auf das Auftreten von Item Nonresponse wirken.

Zu den Variablen, die die Situation des Interviews beschreiben, zählt in diesem Zusammenhang auch die *Erhebungsmethode*, die bei der persönlichen Befragung eine Rolle bezüglich der Privatheit spielt (Dillman et al. 2002; Reuband 1984). Für den ALLBUS 2006 wurde sowohl CAPI als auch CASI verwandt.

*Hypothese 10*: Im Vergleich zu *CAPI* stellt *CASI* eine *Erhebungsmethode* dar, die mit größerer Privatheit einhergeht. Dies macht Item Nonresponse unwahrscheinlicher.

An Interviewermerkmalen stehen folgende Variablen zur Verfügung: das *Bildungsniveau des Interviewers*, die Dauer seiner Tätigkeit als Interviewer und außerdem das *Alter* und das *Geschlecht* des Interviewers. Die Dauer der Tätigkeit für das Befragungsinstitut soll die *Erfahrung* des Interviewers operationalisieren (Hox et al. 1998; Haunberger 2006; Schanz und Schmidt 1984; Esser 1984; Pickery und Loosveldt 1998; Loosveldt et al. 2007).

*Hypothese 11*: Mit zunehmendem *Alter* und zunehmender *Erfahrung* des Interviewers wird das Auftreten von Item Nonresponse unwahrscheinlicher. Ebenso lässt sich erwarten, dass ein höheres *Bildungsniveau* des Interviewers Item Nonresponse reduziert.

Mit diesen Variablen und ihren vermutlichen Wirkungen auf das Auftreten von Item Nonresponse wird die bereits ausführlich besprochene Wirkung des Interviewers auf den Befragten zum Ausdruck gebracht. Da das Interview allerdings als soziale Interaktion aufgefasst werden kann, werden dem Modell noch einige *Interaktionsvariablen* hinzugefügt, um ihre Wirkung auf Item Nonresponse zu testen. Dies erscheint bei folgenden Konstellationen sinnvoll: Geschlecht-I.-Geschlecht-B., Bildung-I.-Bildung-B., Alter-I.-Alter-B..

*Hypothese 12*: Wie in der Unit Nonresponse-Forschung lässt sich nach Koch (2002) von einem *Homogenitätsprinzip* ausgehen. Das heißt, dass gleiche Ausprägungen beim Interviewer und Befragten die Anzahl der Item Nonresponse während des Interviews senken.

Die im Erklärungsmodell verwendeten Variablen sollen in Tabelle 2 noch einmal kurz dargestellt werden.[63] Dort sind auch die Codierungen der einzelnen Variablen verzeichnet.

[63] Die 12 Hypothesen stellen die Alternativhypothesen zur $H_0$-Hypthese „Variable hat keine Einfluss auf die Anzahl der Item Nonresponse“.

| Persönliche Merkmale | Kategorien | Ausprägungen |
|---|---|---|
| Alter | 4 Kategorien dichot. | s. Modelle |
| Geschlecht | dichotom | „Frau“=1 |
| Gesundheitszust. | 5 Ausprägungen | „sehr gut“=1, „schlecht“=5 |
| Bildung | 5 Ausprägungen | „kein Ab.„=1, „Hochschulab.„=5 |
| Wohnumgebung | 5 Ausprägungen | „sehr gut“=1, „schlecht“=5 |
| Einsch. des Wohngeb. | 3 Ausprägungen | „sehr gut“=1, „stark renovierungsbed.“=3 |
| Zuverlässigkeit | dichotom | „unzuverlässig“=1 |
| Antwortbereitschaft | dichotom | „nicht gut“=1 |
| allg. Teilnahmebereitschaft | dichotom | „gut“=1 |
| Erreichbarkeit | dichotom | „gut“=1 |
| Umgebungsmerkmale | | |
| Erhebungsgebiet | dichotom | „Ostdt.“=1 |
| Gemeindegröße | 10 Ausprägungen | „<1.999 Einw.“=1;„>499.999 Einw.“=10 |
| Interviewsituation | | |
| Anwesenheit Dritter | s. Eingriff | - |
| Eingr. Dritter | 4 Ausprägungen | „allein“=1; „n. allein und st. Eingr.“=4 |
| Erhebungsmodus | dichotom | „CAPI“=1 |
| Interviewermerkmale | | |
| Alter | 3 Kategorien dichot. | s. Modelle |
| Geschlecht | dichotom | „Frau“=1 |
| Bildung | 4 Ausprägungen | „kein Ab.“=1, „Hochschulab.“=4 |
| Erfahrung | 4 Kategorien dichot. | s. Modell |
| Interagierende Variablen | | |
| Alter-I.-Alter-B. | 4 Kategorien dichot. | s. Modell |
| Geschlecht-I.-Geschlecht-B. | dichotom | „ungleich“=1 |
| Bild.-I.-Bild.-B. | 3 Ausprägungen | „B. höhere Bild.“=1, „I. höhere Bild.“=3 |

Tabelle 2: Operationalisierung und Codierung der unabhängigen Variablen

Die Modelle wurden mit der Statistiksoftware Latent Gold und SPSS geschätzt. Die Schätzung erfolgt sowohl bei den Poissonregressionen als auch bei den Modellen mit Negativer Binomialverteilung mit Maximum Likelihood.[64]

[64] Für die Details einer Maximum Likelihood Schätzung bei Poissonregressionen und Negativer Binomialverteilung vgl. Hilbe (2007), S.19ff; Winkelmann (2008), S.77ff; für die Details der Schätzung bei ZIP und ZIGP-Regressionen vgl. Czado et al. (2007), S.4f; Winkelmann (2008), S.189ff.

| | OLS | Poi | ZIP | ZIGP | neg. Bin. | ZNB |
|---|---|---|---|---|---|---|
| Persönliche Merkmale | | | | | | |
| - 29 Jahre | $-1{,}4999^{ns}$ | $-0{,}0355^{ns}$ | $0{,}0685^{ns}$ | $-0{,}1649^{ns}$ | $-0{,}5983^{***}$ | $-0{,}0746^{ns}$ |
| 30 - 44 Jahre | $0{,}7955^{*}$ | $-0{,}0403^{ns}$ | $0{,}0637^{ns}$ | $-0{,}1833^{**}$ | $-0{,}5379^{***}$ | $-0{,}0835^{ns}$ |
| 45 - 59 Jahre | $-1{,}2945^{ns}$ | $-0{,}0018^{ns}$ | $-0{,}1022^{**}$ | $-0{,}1590^{***}$ | $-0{,}3897^{***}$ | $-0{,}0039^{ns}$ |
| 60 - | - | - | - | - | - | - |
| Geschlecht | $1{,}0231^{***}$ | $0{,}0792^{***}$ | $0{,}0792^{***}$ | $0{,}0687^{*}$ | $0{,}2477^{***}$ | $0{,}1693^{***}$ |
| Gesundheitszustand | $0{,}0350^{ns}$ | $0{,}0068^{ns}$ | $0{,}0068^{ns}$ | $0{,}0016^{ns}$ | $0{,}0154^{ns}$ | $0{,}0043^{ns}$ |
| Bildung | $-0{,}6163^{**}$ | $-0{,}0771^{***}$ | $-0{,}0771^{***}$ | $-0{,}0667^{**}$ | $-0{,}1386^{***}$ | $-0{,}1484^{***}$ |
| Wohnumgebung | $0{,}9055^{***}$ | $0{,}1058^{***}$ | $0{,}1058^{***}$ | $0{,}0820^{***}$ | $0{,}1995^{***}$ | $0{,}1821^{***}$ |
| Einschätzung des Wohngebäudes | $-0{,}1298^{ns}$ | $-0{,}0493^{**}$ | $-0{,}0493^{**}$ | $-0{,}0431^{ns}$ | $-0{,}0373^{*}$ | $-0{,}0736^{***}$ |
| Zuverlässigkeit | $10{,}5176^{***}$ | $0{,}4705^{***}$ | $0{,}4705^{***}$ | $0{,}5085^{***}$ | $1{,}0444^{***}$ | $0{,}9344^{***}$ |
| Antwortbereitschaft | $2{,}6812^{***}$ | $0{,}1838^{***}$ | $0{,}1838^{***}$ | $0{,}1681^{***}$ | $0{,}4896^{***}$ | $0{,}3113^{***}$ |
| allgemeine Teilnahmebereitschaft | $0{,}2558^{ns}$ | $0{,}0558^{***}$ | $0{,}0558^{**}$ | $0{,}0538^{ns}$ | $0{,}0815^{***}$ | $0{,}0871^{***}$ |
| Erreichbarkeit | $-0{,}2667^{ns}$ | $-0{,}0743^{***}$ | $-0{,}0743^{***}$ | $-0{,}0675^{*}$ | $-0{,}0688^{***}$ | $-0{,}1140^{***}$ |
| Umgebungsmerkmale | | | | | | |
| Erhebungsgebiet | $-1{,}0118^{**}$ | $-0{,}0892^{***}$ | $-0{,}0892^{***}$ | $-0{,}0726^{*}$ | $-0{,}2899^{***}$ | $-0{,}1881^{***}$ |
| Gemeindegröße | $0{,}2250^{***}$ | $0{,}0165^{***}$ | $0{,}0165^{***}$ | $0{,}0184^{***}$ | $0{,}0533^{***}$ | $0{,}0311^{***}$ |
| Interviewsituation | | | | | | |
| Eingriff | $0{,}5363^{**}$ | $0{,}0287^{**}$ | $0{,}0287^{**}$ | $0{,}0395^{*}$ | $0{,}0967^{***}$ | $0{,}0654^{***}$ |
| Erhebungsmodus | $-0{,}4444^{ns}$ | $-0{,}0579^{***}$ | $-0{,}0579^{***}$ | $-0{,}0590^{*}$ | $-0{,}1243^{***}$ | $-0{,}1022^{***}$ |

| Interviewermerkmale | OLS | Poi | ZIP | ZIGP | neg. Bin. | ZNB |
|---|---|---|---|---|---|---|
| - 44 Jahre | $0{,}5303^{ns}$ | $0{,}3168^{***}$ | $0{,}1987^{***}$ | $0{,}3122^{***}$ | $0{,}5858^{***}$ | $0{,}3554^{***}$ |
| 45 - 59 Jahre | $0{,}7548^{*}$ | $0{,}2714^{***}$ | $0{,}1532^{***}$ | $0{,}2880^{***}$ | $0{,}6048^{***}$ | $0{,}2908^{***}$ |
| 60 Jahre - | - | - | - | - | - | - |
| Geschlecht | $-0{,}0174^{ns}$ | $-0{,}0313^{ns}$ | $-0{,}0313^{ns}$ | $-0{,}0178^{ns}$ | $0{,}0104^{ns}$ | $-0{,}0645^{**}$ |
| Bildung | $-0{,}5017^{*}$ | $-0{,}0497^{**}$ | $-0{,}0497^{**}$ | $-0{,}0479^{ns}$ | $-0{,}1540^{***}$ | $-0{,}0961^{***}$ |
| - 10 Jahre tätig | $0{,}6397^{ns}$ | $0{,}4451^{***}$ | $0{,}4539^{***}$ | $-0{,}0354^{ns}$ | $0{,}6358^{***}$ | $0{,}1037^{**}$ |
| 11 - 20 Jahre tätig | $-1{,}4891^{**}$ | $-0{,}4965^{***}$ | $-0{,}4025^{***}$ | $-0{,}0745^{ns}$ | $-0{,}8537^{***}$ | $0{,}0198^{ns}$ |
| 21 - 30 Jahre tätig | $-2{,}6160^{***}$ | $-0{,}5756^{***}$ | $-0{,}3234^{***}$ | $-0{,}1519^{*}$ | $-1{,}0896^{***}$ | $-0{,}2215^{***}$ |
| 31 Jahre tätig - | - | - | - | - | - | - |
| Interagierende Variablen | | | | | | |
| Interviewer wesentlich älter | $-1{,}7452^{*}$ | $-0{,}1277^{*}$ | $0{,}0443^{ns}$ | $-0{,}1226^{ns}$ | $-0{,}3126^{***}$ | $-0{,}1116^{ns}$ |
| Interviewer etwas älter | $-1{,}4627^{*}$ | $-0{,}0529^{ns}$ | $-0{,}1191^{*}$ | $-0{,}0807^{ns}$ | $-0{,}2324^{***}$ | $-0{,}0958^{*}$ |
| etwas gleichaltrig | $-1{,}2192^{**}$ | $-0{,}0125^{ns}$ | $-0{,}1595^{***}$ | $-0{,}0465^{ns}$ | $-0{,}1977^{***}$ | $-0{,}0608^{ns}$ |
| Interviewer jünger | - | - | - | - | - | - |
| Geschlecht-I.-Geschlecht-B. | $-0{,}1886^{ns}$ | $-0{,}0202^{ns}$ | $-0{,}0202^{ns}$ | $-0{,}0148^{ns}$ | $-0{,}0176^{ns}$ | $-0{,}0170^{ns}$ |
| Bildung-I.-Bildung-B. | $0{,}7580^{ns}$ | $-0{,}0958^{**}$ | $-0{,}0958^{**}$ | $-0{,}0646^{ns}$ | $-0{,}1173^{***}$ | $-0{,}1926^{***}$ |
| Log-Likelihood | - | -5916,82 | -5917,32 | -5219,20 | -15733,65 | -12693,74 |

Tabelle 3: Erklärungsmodelle für Item Nonresponse mit OLS-Regression, Poissonregression, ZIP-Regression, ZIGP-Regression, Negative Binomialregression, Zero-Inflated Negative Binomialregression[65]

[65] Signifikanzniveaus im Folgenden: $***: p \leq 0{,}001$; $**: p \leq 0{,}01$; $*: p \leq 0{,}05$; ns=nicht signifikant.

Um die Vielzahl der Ergebnisse zu systematisieren, soll zunächst ein Blick auf die Determinanten geworfen werden, die in allen Modellen einen signifikanten Beitrag zur Erklärung der Anzahl der Item Nonresponse erbringen. Dies sind die Regressoren von *Geschlecht* (H3), *Bildung* (H1), *Wohnumgebung* (H7), *Zuverlässigkeit der Antwort* (H4) und *Antwortbereitschaft* (H4) des Befragten. Darüber hinaus sind bei den Variablen zum Befragtenumfeld *Erhebungsgebiet* (H5) und *Gemeindegröße* (H8) stets signifikant sowie beim Interviewumfeld *Eingriff* und bei den Interviewermerkmalen Interviewer mit einer *Erfahrung zwischen 21 und 30 Jahren*. Bis auf eine Ausnahme bestätigen diese Koeffizienten die in den Hypothesen geäußerten Erwartungen, dass diese oder jene Determinante auf die Anzahl der Item Nonresponse senkend und vermehrend wirken. Eine Ausnahme stellt die Variable *Erhebungsgebiet* (H5) dar. Sie zeigt den entgegengesetzten Einfluss – nämlich eine verminderte Wirkung auf das Eintreten von Item Nonresponse, wenn der Befragte in Ostdeutschland wohnt.
Zwei Variablen besitzen über alle Modelle keinen signifikanten Einfluss, nämlich der *Gesundheitszustand* (H2) des Befragte sowie die Interaktionsvariable *Geschlecht-I.-Geschlecht-B.* (H12). Eventuell findet die Wirkung des letztgenannten Merkmals bereits bei der Kontaktaufnahme statt. In Abschnitt 5.3.2 soll hierauf eine Antwort gegeben werden.
Alle anderen Determinanten besitzen in mindestens einem Modell Signifikanz. Bezieht man an dieser Stelle die Modellgüten (Log-Likelihood) ein, zeigt sich interesanterweise, dass nicht einmal die OLS-Regression das schlechteste Modell ist, sondern die Negative Binomialregression. Die besten Log-Likelihoods weisen die Modelle der Poissonregression auf, wobei zwischen der ZIP-Regression und der Poissonregression kaum ein Unterschied besteht, die Einbeziehung der Overdispersion allerdings verbessert das Modell noch einmal merklich. Beim ZIGP-Modell ist zusätzlich eine Reihe von Variablen signifikant: die zwei mittleren *Altersgruppen* (H1) zeigen einen signifkanten negativen Einfluss auf das Eintreten von Item Nonresponse, der sich aber mit dem Älterwerden abschwächt. Auch zeigt in diesem Modell der *Erhebungsmodus* einen zumindest schwach signifikanten leichten Einfluss, und zwar entsprechend der Hypothesenformulierung (H10). Bei den Interviewermerkmalen schwächt sich mit dem *Alter* der positive Einfluss ab. Damit bestätigt sich oben formulierte Hypothese H11. Auch die größere *Erfahrung* des Interviewers wirkt sich in diesem Modell signifkant aus (H11). Durch die Modellierung der Overdispersion wird den Interaktionsvariablen die Erklärungskraft vollständig genommen – ein Phänomen, das auch bei der Negativen Binomialregression eintritt, wenn man die letzte Spalte betrachtet. Dies verwundert nicht, empfiehlt doch beispielsweise Hilbe (2007) als Alternative zur Modellierung der Overdispersion, eine Mehrebenenanalyse zu verwenden.
Insgesamt hat sich die modifizierte Poissonregression für ein Erklärungsmodell von Item Nonresponse bewährt. Ein großer Teil der Hypothesen konnte empirisch bestätigt werden, wenngleich für die Interviewer- und die Interaktionsvariablen eher schwache und häufig insignifikante Einflüsse geschätzt wurden.
Nachdem nun ausführlich auf die Analyse von Item Nonresponse eingegangen worden ist, soll ein Korrekturmethodenvergleich für Item Nonresponse vorgenommen werden.

# 4 Item Nonresponse: Korrekturmethoden im Vergleich

## 4.1 Einleitung

Bei allen nun vorgestellten Beispielen haben sich die Autoren entschieden, für ihre Analysen ausschließlich Complete Cases (CC) zu verwenden, und damit gegen eine aktive Korrekturmethode für den teilweise massiven Datenausfall. Im folgenden Methodenvergleich wurde im Unterschied zu den allermeisten älteren Veröffentlichungen darauf geachtet, nicht auf rein simulierte Daten zu setzen (Bankhofer 1995, S.189; Göthlich 2007, S.130), sondern tatsächlich auf realen Daten aufbauend einen Datensatz mit gleichen Strukturen zu erzeugen, der objektiv die gleiche Ausgangslage schafft, und als Indikatoren bereits publizierte Modelle zu berechnen. Letztlich können so die Vorteile von realen Daten und einer Simulation kombiniert werden. Sicherlich nachteilig im Vergleich zu einer Simulationsstudie ist die Einschränkung der Untersuchungsmöglichkeiten: sowohl Datenausfall als auch Stichprobengrößen und zu schätzende Parameter sind vorgegeben. Da dieses Vorgehen ganz neuartig ist, wird es in Abschnitt 4.3 eingehend erklärt. Nachdem das durchaus aufwendige Verfahren zum Methodenvergleich beschrieben wurde, richtet sich das Augenmerk auf die Frage, ob die Verfahren der Multiple Imputation zu besseren Ergebnissen führen als die von den Autoren gewählte Methode, nur CC zu verwenden.

## 4.2 Ausgewählte Beispiele

Zum Vergleich der Leistungsfähigkeit verschiedener Korrekturverfahren werden im Folgenden drei Publikationen kurz vorgestellt, ausgewählte Parameterschätzungen repliziert und schließlich die Auswirkung der fehlenden Werte und deren Korrektur evaluiert.

- Beispiel 1: Hennig, E. (2009) „Einen Schlussstrich unter die nationalsozialistische Vergangenheit ziehen“. Zur politischen Soziologie eines historischen Deutungsmusters, Einsicht 2, S.42-49.

- Beispiel 2: Schnabel, C. und J. Wagner (2005) Who Are the Workers Who Never Joined a Union? Empirical Evidence from Germany, IZA, Paper 1698.

- Beispiel 3: Schäfer, A. (2009) Alles halb so schlimm? Warum eine sinkende Wahlbeteiligung der Demokratie schadet, MPIfG Jahrbuch, S.5-10.

Die Reihenfolge der Publikationen ist nach der Komplexität der zu schätzenden Parameter geordnet: während bei Eike Hennig die Schätzung bedingter Anteilswerte im Vordergrund steht, berechnen Schnabel und Wagner ein multivariates Probitmodell. Armin Schäfer schätzt auf Basis eines Logitmodells individuelle Wahrscheinlichkeiten. Die replizierten und analysierten Parameterschätzungen werden nun in dieser Reihenfolge kurz vorgestellt. Da der spätere Methodenvergleich nicht auf einer Simulation im normalen Sinn fußt, sondern diese als Vorlagen in den Rahmen einer Quasisimulation übernommen werden, wird auch kurz auf die möglichen allgemeinen methodischen Einschränkungen der Artikel eingegangen.

### 4.2.1 Beispiel 1: Anteilswerte

Die Veröffentlichung *„Einen Schlussstrich unter die nationalsozialistische Vergangenheit ziehen". Zur politischen Soziologie eines historischen Deutungsmusters* beschäftigt sich mit der Analyse jener Bevölkerungsgruppe, die unter die öffentliche Diskussion um die NS-Vergangenheit einen Schlussstrich ziehen möchte. Der Autor schätzt zu Beginn diese Gruppe als uneinheitlich ein. Die Fragestellung lautet „Wer ist Träger, welche Einstellungen treffen sich im Schlussstrich?" (S.46). Zwar wird der ALLBUS 2006 als Sekundärdatengrundlage genutzt, doch selektiert die Fragestellung eindeutig Variablen, die durchgehend in dieser Erhebung als sensibel gelten. Aus der Literatur wird eine Reihe von persönlichen und sozio-ökonomischen Merkmalen einerseits und andererseits einige Einstellungen, die im Zusammenhang mit einer Schlussstrichbefürwortung stehen könnten, herausgearbeitet. Die Auswahl für die Replikation und den späteren Methodenvergleich fiel auf die Anteilswerte von bedingten Verteilungen bei folgenden Merkmalen:

- Anomie,
- Ausländerfeindlichkeit,
- freie Meinungsäußerung,
- Inflationsangst,
- Antisemitismus.

Die Hälfte der Items besteht aus der sogenannten Antisemitismus Itembatterie, die im ALLBUS erst zweimal abgefragt wurden. Bereits diese Items tangieren eindeutig den Bereich sensibler Themen. Die Einordnung der Wichtigkeit von Inflationsbekämpfung und freier Meinungsäußerung ist Bestandteil des Ingelhart-Index zu postmaterialistischen Einstellungen (Inglehart 1977; Inglehart und Abramson 1999). Als Item für Ausländerfeindlichkeit wird eines von mehreren Items ausgewählt: *Fremd im eigenen Land wegen Ausländern*. Anomie wurde als Index aus drei Items gebildet: „Lageverschlechterung für einfache Leute", „keine Kinder mehr in diese Welt setzen",

„Mehrheit uninteressiert an Mitmenschen“. Diese Variablen sind jeweils dichotom mit „stimme zu“ =1 bzw. „stimme nicht zu“ =0. Die Ausfälle der einzelnen Variablen sind sehr verschieden und weisen – anders als bei Einstellungsvariablen zu erwarten – nur in einem Fall den Wert Null auf:[66]

| **Item** | **Ausfall in %** | **Bezeichnung** |
|---|---|---|
| Schlussstrich | 2,2 | Schlussstrich |
| Anomie | 0,0 | Anomie |
| Fremd im eigenen Land wg. Ausländern | 0,7 | Fremd |
| Inflationsbekämpfung | 3,2 | Inflation |
| Meinungsäußerung | 3,4 | Meinung |
| Juden haben zu viel Einfluss | 11,7 | Antisem3 |
| Keine Scham über dt. Untaten an Juden | 5,2 | Antisem2 |
| Juden nutzen dt. Vergangenheit aus | 10,2 | Antisem1 |
| Juden an Verfolgung nicht unschuldig | 12,4 | Antisem4 |
| n = 3.193 | | |

Tabelle 4: Ausfall der Variablen in Beispiel 1

Die Gesamtzahl von n = 3.193, auf die sich alle prozentualen Ausfallangaben beziehen, erhält man nach Abzug der nicht-deutschen Befragten, da die thematische Fragestellung des Aufsatzes nahelegt, dass die „Schlussstrichziehung“ eine Diskussion der deutschen Öffentlichkeit sei. Den höchsten Ausfall weisen erwartungsgemäß die Antisemitismusvariablen auf, von denen bis auf das Scham-Item alle Merkmale mehr als 10 % Item Nonresponse zeigen.[67] Auffallend erscheint die geringe Missing-Rate im Anomieindex. Dies liegt an der Konstruktion des Index als additiver Index. An einem Beispiel soll kurz verdeutlich werden, wie dies vonstatten geht:

| | **NA** | **0** | **1** | **2** | **3** | **∑** |
|---|---|---|---|---|---|---|
| Anomie additiv | 0 | 187 | 663 | 1.265 | 1.058 | 3.193 |
| Anomie CC | 250 | 146 | 555 | 1.184 | 1.058 | 3.193 |

Tabelle 5: Verteilung des Anomieindex nach unterschiedlicher Konstruktion

[66] Die Bezeichnungen in Tabelle 4 finden sich dann entsprechend in Abbildung 13.
[67] Im Anhang 2 ist der Wortlaut der Items.

Die obere Zeile zeigt die Verteilung mit dem additiven Index, bei dessen Bildung für fehlende Werte null imputiert wird. Die untere Zeile weist entsprechend die eigentlich fehlenden Werte auf; das heißt, dass 250 Befragte auf eine, zwei oder alle drei Fragen, die den Anomieindex bilden, keine Antwort gegeben haben. Danach würde die Variable Anomie immerhin unter einem Ausfall von n=250 Nichtantwortenden oder 7,8 % leiden. In der Praxis ist die Methode der additiven Indexbildung weit verbreitet, ohne dass die Konsequenzen berücksichtigt würden.

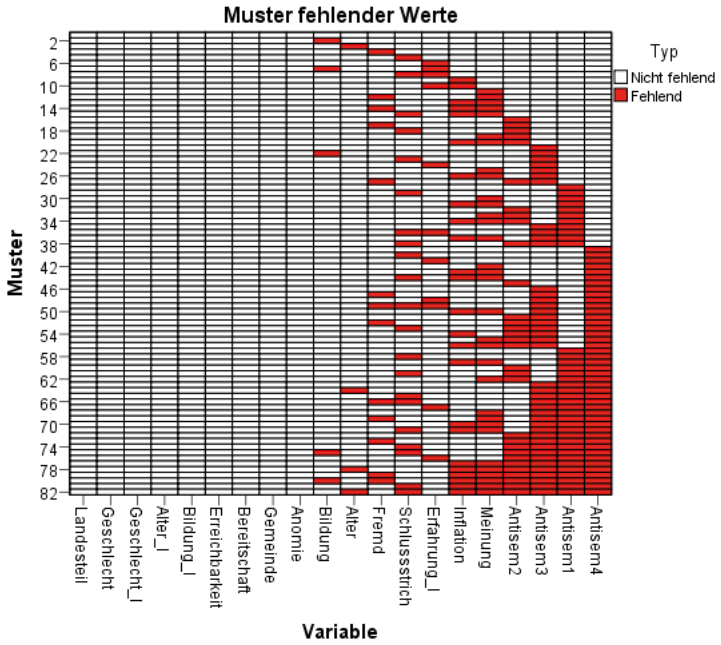



Abbildung 13: Ausfallmuster von Beispiel 1

Die grafische Aufbereitung der fehlenden Werte (Abbildung 13) zeigt ein Mosaik, das keine Hinweise auf einen nur annähernd monotonen Datenausfall gibt.[68] Für den späteren Vergleich wurden insgesamt 16 Parameter zur Schätzung ausgewählt. Zum einen wurde der Anteil der Personen, die als „Schlussstrichzieher“ anomische, fremdenfeindliche oder andere einschlägige Tendenzen aufweisen (linke Spalte: 1. bedingter Anteilswert mit Basis „Schlussstrichzieher“), geschätzt. Daneben befindet sich der Anteil der „Schlussstrichzieher“ von allen Personen, die eine anomische, fremdenfeindliche oder andere einschlägige Tendenz aufweisen (rechte Spalte: 2. bedingter Anteilswert mit Basis anomische, fremdenfeindliche Tendenzen). Ab welcher Ausprägung von einer Tendenz gesprochen wird, hat der Autor für jede Variable einzeln definiert. Die Replikation ergab folgende Werte:

[68] Die gezeigte Grafik, die auch für die beiden anderen Datensätze erstellt wurde, wurde in SPSS berechnet und gezeichnet. Als Merkmale sind jeweils die Analysevariablen sowie – der Übersichtlichkeit wegen – nur ein Teil der Variablen, die für spätere Korrekturmethoden verwendet werden, einbezogen. Die Variablen sind nach Ausfallmuster und Ausfallumfang sortiert. Bei monotonen Ausfällen dürfte es keine Überschneidung in den Zeilen geben.

| Schlussstrich bzw. | **1. bed. Anteil** | **2. bed. Anteil** | **nach CC gültige Fälle** |
|---|---|---|---|
| Fremd | *0,40* | 0,77 | n = 3.105 |
| Inflation | *0,43* | 0,72 | n = 3.032 |
| Meinung | 0,37 | 0,55 | n = 3.025 |
| Judeneinfluss | 0,38 | *0,78* | n = 2.788 |
| Keine Scham | 0,16 | 0,80 | n = 2.984 |
| Juden nutzen aus | 0,44 | 0,80 | n = 2.830 |
| Juden nicht unschuldig | 0,23 | 0,70 | n = 2.764 |
| Anomie | *0,39* | 0,75 | n = 3.124 |

Tabelle 6: Replizierte Werte für Beispiel 1

Die kursiven Zahlen weichen von den publizierten ab, da dort Berechnungsfehler aufgetreten sind.[69] Die 16 bedingten Anteilswerte sind die Parameter, die später noch einmal unter verschiedenen Korrekturmethoden geschätzt werden. Dabei ist es aus statistischer Sicht durchaus reizvoll, dass sowohl symmetrische als auch sehr schiefe Verteilungen vorliegen.
Da es sich um die Replikation von Analysen handelt, wird auch die umstrittene additive Indexbildung mit der Nullen-Imputation im Weiteren verwendet.

[69] Es wird weiter mit den richtigen Werten gearbeitet.

### 4.2.2 Beispiel 2: Multivariates Probitmodell

Der Artikel von Schnabel und Wagner beschäftigt sich mit der Frage, welche Arbeitnehmer niemals in einer Gewerkschaft organisiert waren. Die Autoren sprechen von der empirischen Suche nach „Resistenzfaktoren“ gegenüber Gewerkschaften. Hierfür wird unter anderem mit Variablen aus dem ALLBUS 2002 eine Analyse durchgeführt. Zum Vergleich wurden auch Daten des European Social Surveys herangezogen. Der ALLBUS 2002 enthält dabei eine große Zahl von Merkmalen, die Schnabel und Wagner aus der Literatur als sinnvolle Determinanten der Gewerkschaftsmitgliedschaft identifizieren. Die abhängige Variable ist eine Konstruktion aus zwei Variablen:

- früher Gewerkschaftsmitglied,
- jetzt Gewerkschaftsmitglied.

Die unabhängigen Variablen sind zunächst persönliche Merkmale und sozio-ökonomische Variaben, die sich auf das Erwerbsleben beziehen:

- Geschlecht,
- Alter,
- Landesteil,
- jemals arbeitslos,
- Bildung,
- Bildung von Vater und Mutter,
- Statusgruppe der Eltern,
- Vollzeitbeschäftigung,
- Mitglied des öffentlichen Dienstes.

Einzige Einstellungsvariable ist die Linksrechtsselbsteinstufung des Befragten auf einer Skala von 1 bis 11:

- Politische Orientierung (Linksrechtsselbsteinstufung).

Ein Teil der Variablen weist keine fehlenden Werte auf, aber gerade die Fragen nach dem Bildungsniveau und der beruflichen Stellung der Elternteile enthalten teilweise zweistellige Item Nonresponse-Raten. Dies ist nicht verwunderlich, da Erinnerungslücken oder Unsicherheit der Befragten bezüglich der Informationen über die Elterngeneration durchaus nicht unplausibel sind (Tourangeau et al. 2000; Willis et al. 1999). In der Übersicht stellen sich die für die Analyse aufbereiteten und umcodierten Variablen mit ihren Ausfällen in % dar:

| Item | Ausfall in % | Bezeichnung |
|---|---|---|
| niemals Gewerkschaftsmitglied | 0,4 | Gewerkschaftsmgl |
| Geschlecht | 0,0 | Geschlecht |
| Alter | 0,0 | Alter |
| $Alter^2$ | 0,0 | - |
| jemals arbeitslos | 0,0 | Arbeitslosigkeit |
| niedrige Bildung | 0,8 | Bildung |
| Universitätsabschluss | 0,3 | Uniabschluss |
| niedrige Bildung Vater | 9,6 | Bildung_Vater |
| niedrige Bildung Mutter | 4,7 | BildungMutter |
| Vater Arbeiter | 13,6 | Arbeiter_Vater |
| Arbeiter | 0,0 | Arbeiter |
| Ostdeutsch | 0,0 | Landesteil |
| politische Orientierung | 5,0 | PolitischeOrient |
| Vollzeitbeschäftigung | 0,0 | Vollbeschäftigung |
| Öffentlicher Dienst | 0,2 | Öffent_Dienst |
| n=1.162 | | |

Tabelle 7: Ausfall der Variablen in Beispiel 2

Die Ausfälle sind prozentual zu einer Stichprobengröße von n=1.162 Befragten.[70] Von der ursprünglichen ALLBUS-Gesamtstichprobe von n= 2.820 werden aufgrund der Fragestellung alle unter 18-Jährigen und über 65-Jährigen sowie alle Nichtdeutschen aus der Stichprobe genommen. Ob der Vater Arbeiter war und welches Bildungsniveau dieser hat oder hatte, konnten immerhin knapp 14 % bzw. knapp 10 % der Befragten nicht beantworten. Gerade die meisten sozio-ökonomischen Variablen besitzen überhaupt keine fehlenden Werte. Die Bildung der Mutter und die Frage nach der *politischen Orientierung* (Linksrechtsselbsteinstufung) haben etwa 5 % Item Nonresponse. Die Selbsteinstufung benötigt eine Vorstellung von der politischen Linksrechtsdimension und die intellektuelle Fähigkeit sich darin selbst einzuordnen. Zudem lassen sich die immerhin 5 % Ausfälle (ohne über 65-Jährige) damit erklären, dass eine politische Grundeinstellung als sensible Frage betrachtet wird. Das Merkmal *politische Orientierung* ist neben der Altersvariable zudem das einzige Merkmal, das nicht dichotom ist oder für die Modellierung dichotomisiert wurde.

[70] Bei Fragen zur eigenen Vergangenheit und der der Eltern konnte neben „keine Angabe“ die Ausweichkategorie „weiß nicht“ angegeben werden; bei allen anderen Variablen stand nur die Kategorie „keine Angabe“ zur Verfügung.

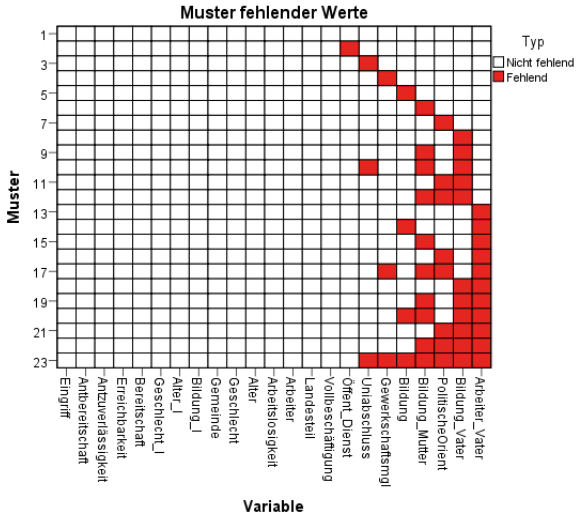



Abbildung 14: Ausfallmuster von Beispiel 2

Für den Methodenvergleich wurde das Probitmodell mit den folgenden Parameterschätzungen ausgewählt und ohne Gewichtung (Ost-West-Gewichtung) repliziert.[71] Im multivariaten Raum reduziert sich die Stichprobengröße um fast ein Drittel auf n=888.

| | **gew. Werte (original)** | **replizierte Werte** |
|---|---|---|
| Achsenabschnitt | $4{,}583^{***}$ | $4{,}596^{***}$ |
| Geschlecht | $-0{,}039^{ns}$ | $-0{,}035^{ns}$ |
| Alter | $-0{,}182^{***}$ | $-0{,}184^{***}$ |
| $\text{Alter}^2$ | $0{,}002^{***}$ | $0{,}002^{***}$ |
| jemals arbeitslos | $-0{,}133^{ns}$ | $-0{,}132^{ns}$ |
| niedrige Bildung | $0{,}149^{ns}$ | $0{,}149^{ns}$ |
| Universitätsabschluss | $0{,}210^{*}$ | $0{,}217^{*}$ |
| niedrige Bildung Vater | $0{,}176^{ns}$ | $0{,}178^{ns}$ |
| niedrige Bildung Mutter | $0{,}046^{ns}$ | $0{,}051^{ns}$ |
| Vater Arbeiter | $-0{,}248^{**}$ | $-0{,}246^{**}$ |
| Arbeiter | $-0{,}317^{**}$ | $-0{,}317^{**}$ |
| Ostdeutsch | $-0{,}892^{***}$ | $-0{,}889^{***}$ |
| politische Orientierung | $0{,}101^{***}$ | $0{,}102^{***}$ |
| Vollzeitbeschäftigung | $-0{,}301^{**}$ | $-0{,}305^{**}$ |
| Öffentlicher Dienst | $-0{,}296^{***}$ | $-0{,}301^{***}$ |
| | n=888 | n=888 |

Tabelle 8: Replizierte Werte für Beispiel 2

[71] Die Gewichtung wurde unterlassen, um den späteren Methodenvergleich nicht unnötig kompliziert zu machen.

### 4.2.3 Beispiel 3: Multivariates Logitmodell und individuelle Eintrittswahrscheinlichkeiten

Schäfer analysiert die Verzerrung verschiedener politischer Partizipationsformen.[72] Theoretische Grundlagen sind die sogenannten Normalisierungs-, Zufriedenheits- und Substitutionsthese. Studien zeigten, dass eine Wahl umso verzerrter ist, je niedriger die Wahlbeteiligung ist. Die abhängige Variable ist unter anderem die Wahlbeteiligung, die von verschiedenen Faktoren determiniert wird. In der Literatur finden sich Hinweise auf das *Alter*, die *Sozialisation* in Westdeutschland, das *Haushaltseinkommen* und das *politische Interesse* (S.7). Zudem werden für die Analyse noch die einschlägige Variable *Demokratiezufriedenheit* sowie ein *Partizipationsindex*, der aus folgenden Variablen gebildet wurde, herangezogen:

- Unterschriften sammeln,
- kritischer Konsum,
- Demonstrationsteilnahme,
- Teilnahme an politischer Versammlung,
- Kontakt zu Politikern,
- Spenden sammeln.

Der *Partizipationsindex* ist additiv gebildet – wiederum mit allen Nachteilen, die bereits für die Bildung des Anomieindex in 4.2.1 besprochen wurden. Da die Ausfälle, die einen Umfang von 2 % bis 10,2 % annehmen, insgesamt sechs Variablen betreffen, ist die Nullimputation von wesentlich größerem Umfang: mit einem Ausfall von 15,8 % ist der Anteil der Nullen in der Indexverteilung erheblich vergrößert. Mit Blick auf den Ausfall der anderen Variablen, die dann in das Logitmodell eingefügt werden, zeigt sich ein extremes Bild. Die hypothetischen 15,8 % wären nicht einmal der größte Ausfall für eine einzige Variable, sondern lediglich an zweiter Stelle. Nahezu jeder zweite Befragte hat keine Antwort auf die Einkommensfrage gegeben. In der Literatur wird häufig vermutet, dass die Einkommensvariable wenigstens zum Teil einem NMAR-Ausfallmechanismus folgt (Havasi und Marton 1998). Für die Korrekturmethoden bedeutet dies eine große Herausforderung. Gerade in Deutschland ist die Frage nach dem Einkommen äußerst sensibel und der Ausfall deshalb auch nicht unerwartet, trotzdem äußerst kritisch für die spätere Schätzung. Für das Logitmodell weisen alle berücksichtigten Merkmale folgenden Anteil an fehlenden Werten auf:[73]

[72] Die verwendeten Items entstammen zum Teil dem ALLBUS-Zusatz ISSP, den die Befragten selbst schriftlich ausfüllen; als Ausweichkategorie existiert einheitlich die Kategorie „kann ich nicht sagen“. Für den genauen Wortlaut der Items siehe Anhang 3.

[73] Für den genauen Wortlaut verwendeter Items siehe Anhang 4.

| Item | Ausfall in % | Bezeichnung |
|---|---|---|
| Wahlteilnahme | 8,6 | Wahlteilnahme |
| Demokratiezufriedenheit | 5,4 | Demzufriedenheit |
| Landesteil | 0,0 | Landesteil |
| Geschlecht | 0,0 | Geschlecht |
| Alter | 0,0 | Alter |
| Bildung | 1,3 | Bildunghoch, -niedrig, -mittel |
| Einkommen | 44,5 | Einkommenhoch, -niedrig, mittel |
| politisches Interesse | 0,0 | Pol-Interesse |
| Partizipationsindex | 0,0 | Partizipation |
| n=1.132 | | |

Tabelle 9: Ausfall der Variablen in Beispiel 3

Auch die Frage nach der Wahlbeteiligung bei der letzten Bundestagswahl haben immerhin fast 9 %, auf die Frage nach der Zufriedenheit mit der Demokratie aben 5,6 % nicht beantwortet. Die anderen Variablen weisen nur äußerst geringe oder gar keine Item Nonresponse auf. Gerade auch Variablen, die soziales Verhalten abfragen, leiden unter sozialer Erwünschtheit. So ist es pausibel, dass manche Befragte vorgeben, die Antwort nicht (mehr) zu wissen (Item Nonresponse) oder bewusst etwas Falsches antworten. Das Ausfallmuster wird ebenfalls von der Einkommensvariable definiert. Die Missingraten beziehen sich auf eine Stichprobengröße von 1.332 Befragten (ISSP Fragebogen).

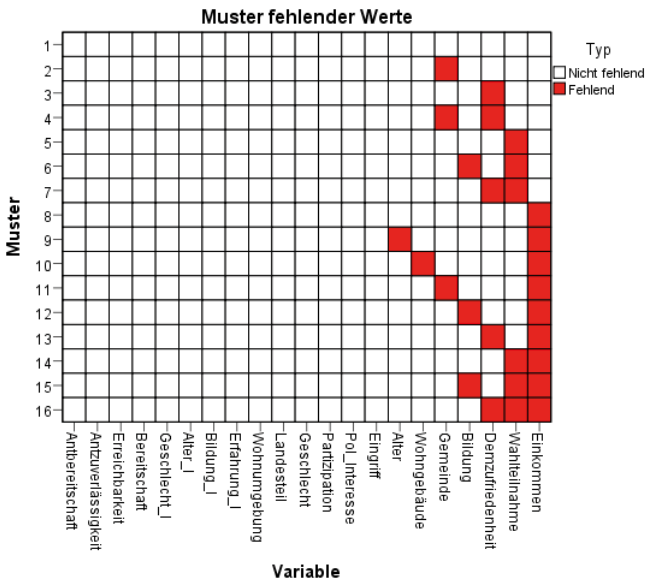



Abbildung 15: Ausfallmuster von Beispiel 3

Ausgewählt werden die Schätzungen individueller Wahlteilnahmewahrscheinlichkeiten in Abhängigkeit von der Demokratiezufriedenheit auf Basis eines Logitmodells, das folgende (großteils dichotomisierte) Variablen enthält. Die Tabelle 10 zeigt daneben die Werte der originalen Berechnung (gewichtet nach Ost-/Westdeutschland) und der ungewichteten replizierten binären logistischen Regression:

| | **gew. Werte (original)** | **replizierte Werte** |
|---|---|---|
| Achsenabschnitt | $-0{,}187^{ns}$ | $-0{,}507^{ns}$ |
| ostdeusch | $-0{,}706^{*}$ | $-0{,}696^{*}$ |
| weiblich | $0{,}219^{ns}$ | $0{,}268^{ns}$ |
| Alter | $0{,}042^{***}$ | $0{,}042^{***}$ |
| mittlere Bildung | $0{,}473^{ns}$ | $0{,}469^{ns}$ |
| hohe Bildung | $0{,}908^{ns}$ | $0{,}943^{ns}$ |
| Einkommen | $0{,}001^{*}$ | $0{,}001^{*}$ |
| Demokratiezufriedenheit | $0{,}197^{*}$ | $0{,}244^{**}$ |
| politisches Interesse | $0{,}819^{**}$ | $0{,}857^{**}$ |
| Partizipationsindex | $0{,}259^{*}$ | $0{,}284^{*}$ |
| | $n = 648$ | $n = 648$ |

Tabelle 10: Replizierte Werte für Beispiel 3 (logistische Regression)

*Ostdeutsch*, *Alter* und *Bildung* sind jeweils dichotom oder wurden dichotomisiert. Die Einkommensvariable ist eine der wenigen stetigen Merkmale im ALLBUS. Demokratiezufriedenheit wird auf einer 11-stufigen Skala gemessen, die Variable *politisches Interesse* besitzt 5 Merkmalswerte, der Partizipationsindex insgesamt 7. Über alle Variablen besitzt das Modell einen multivariaten Ausfall von knapp der Hälfte der Befragten; es bleiben 648 Merkmalsträger zur Analyse übrig. Dies ist die Basis für die Schätzung der Wahlteilnahme für einen ostdeutschen bzw. westdeutschen Mann mittleren Alters, mittlerer Bildung und mittleren Einkommens (als Beispiel für die Ausprägung 0 der Demokratiezufriedenheit):

$$\frac{\exp(b_1 + b_5 \cdot 1 + b_8 \cdot 0)}{(1 + \exp(b_1 + b_5 \cdot 1 + b_8 \cdot 0))}$$

Da diese Schätzung sehr komplizierte Annahmen trifft, werden beim Korrekturmethodenvergleich zunächst das grundlegende Logitmodell geschätzt und anschließend die individuellen Teilnahmewahrscheinlichkeiten. Hierfür wurden zunächst die Werte als Replikation berechnet (Tabelle 11); die Unterschiede zu Werten aus der Publikation ergeben sich wiederum aus der fehlenden Gewichtung, auf die wiederum verzichtet wurde, um den Methodenvergleich nicht zu verkomplizieren:

| Demokratie-zufriedenheit | replizierte Werte (West) | replizierte Werte (Ost) |
|---|---|---|
| 1 | 0,5766 | 0,4045 |
| 2 | 0,6348 | 0,4644 |
| 3 | 0,6894 | 0,5254 |
| 4 | 0,7391 | 0,5856 |
| 5 | 0,7834 | 0,6434 |
| 6 | 0,8220 | 0,6973 |
| 7 | 0,8550 | 0,7462 |
| 8 | 0,8827 | 0,7896 |
| 9 | 0,9057 | 0,8274 |
| 10 | 0,9246 | 0,8595 |
| 11 | 0,9400 | 0,8865 |

Tabelle 11: Replizierte Werte für Beispiel 3 (Wahrscheinlichkeit der Wahlteilnahme)[74]

## 4.3 Verfahren zum Vergleich von Korrekturmethoden

Veröffentlichungen zum Vergleich von Korrekturmethoden beinhalten entweder die bloße Durchführung der Methode anhand eines ausgewählten Datensatzes (z.B. Longford 2000) oder eine Simulationsstudie (z.B. Rässler 2000). Beide Optionen haben Nachteile. Die einfache Durchführung mit einem realen Datensatz berücksichtigt zwar die Herausforderungen realer Datengegebenheiten, sie hat allerdings eine eingeschränkte Aussagekraft über den ausgewählten Datensatz hinaus. Simulationen müssen hingegen immer mit dem Makel des Künstlichen, letztlich auch des häufig weniger Komplexen, leben.[75] Ein weiterer Punkt ist die Selektivität der Modelle und Verteilungen für Simulationen, die nicht unbedingt zur Falsifikation, sondern teilweise unbewusst zur Bestätigung bereits vorhandener Thesen ausgewählt werden. Der Vorteil der Simulation liegt dagegen darin, die „wahren“ Parameterwerte zu kennen, um einen Vergleich geschätzter Parameter ziehen zu können. Die ist bei der Anwendung realer Daten nicht möglich, da zunächst keine „wahren“ Werte bekannt sind.
Für den folgenden Methodenvergleich gilt es einen Weg zu finden, der die Vorteile beider Herangehensweisen miteinander verbindet und einen zuverlässigen Methodenvergleich gangbar macht. Um dem Moment subjektiver Steuerung einer Simulation im Methodenvergleich Rechnung zu tragen, wurde in diesem neuen Verfahren bewusst auf bereits publizierte statistische Analysen zurückge-

[74] Gewichtete Originalwerte siehe Schäfer (2009), S.8.

[75] Es gibt mittlerweile eine Zahl von sehr ausgefeilten und komplexen Simulationsstudien, vgl. Koller-Meinfelder (2010).

griffen. Dieses Verfahren wird sowohl für die Evaluation der Item als auch Unit Nonresponse-Korrekturen angewendet. Entsprechende Unterschiede werden an gegebener Stelle im Kapitel zu Unit Nonresponse diskutiert (Abschnitt 6.1). Das folgende Schema soll das Vorgehen verdeutlichen:

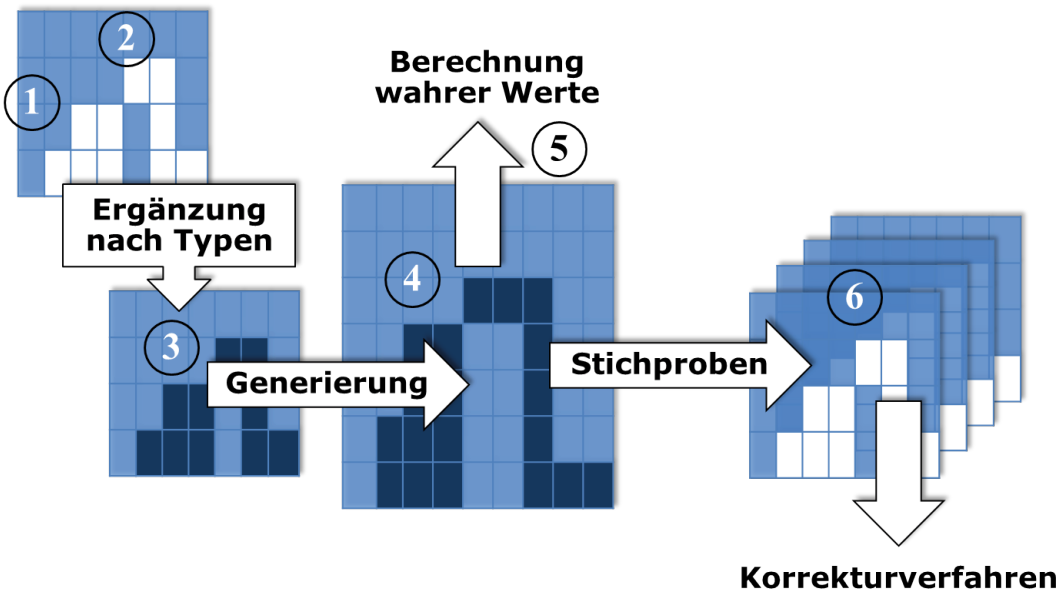



Abbildung 16: Verfahren zum Vergleich von Item Nonresponse-Korrekturmethoden

Der erste Schritt zeigt den Datensatz (konkret: ALLBUS 2002, 2004, 2006) in vereinfachter Form. Zu sehen sind gegebene (blau) und fehlende Daten (weiß). Wie aus der kurzen Beschreibung der Datensätze hervorgeht, hat keiner der Datensätze nur annähernd monotone Ausfallstrukturen. Allerdings existieren Variablen, die keine oder vernachlässigbar wenige (unter 0,5 %) fehlende Werte besitzen (1). Dies variiert jedoch von Datensatz zu Datensatz. Darunter ist in jedem Fall eine Anzahl von Merkmalen, die vorher in der Analyse zu Item Nonresponse (Abschnitt 3.3.3) als mehr oder weniger starke Einflussfaktoren von Item Nonresponse identifiziert wurden. Zur Übersicht sind diese Variablen noch einmal zusammengefasst:[76]

[76] Vorhandene Variablen sind mit „X“, im ALLBUS fehlende Variablen sind mit „-“ gekennzeichnet.

| Variable | ALLBUS 2002 | ALLBUS 2004 | ALLBUS 2006 |
|---|---|---|---|
| Alter | X | X | X |
| Bildung | X | X | X |
| Geschlecht | X | X | X |
| Gesundheitszustand | - | X | X |
| Landesteil | X | X | X |
| Zuverlässigkeit | X | X | X |
| Bereitschaft | X | X | X |
| Teilnahmebereitschaft | X | X | X |
| Erreichbarkeit | X | X | X |
| Wohngebäude | X | X | X |
| Wohngegend | - | - | X |
| Interviewer Bildung | X | X | X |
| Interviewer Alter | X | X | X |
| Interviewer Erfahrung | X | X | X |
| Interviewer Geschlecht | X | X | X |
| Erhebungsmethode | - | - | X |
| Anwesenheit Dritter | X | X | X |
| Eingriff Dritter | X | X | X |
| Gemeindegröße | X | X | X |

Tabelle 12: Übersicht vollständige Variablen

Hinzu kommen die jeweiligen Variablen oder die Variable des geschätzten Parameters oder der geschätzten Parameter, die aus der jeweiligen Publikation ausgewählt und im obigen Abschnitt ebenfalls beschrieben wurden (2).
Dieser Datensatz wird nun bootstrapartig vergrößert, indem Stichproben aus ihm gezogen und immer wieder an den ursprünglichen Datensatz angehängt werden (3). Damit erhält man Ziehung für Ziehung einen immer größeren Datensatz. Um eine reine Replikation zu vermeiden, wird bei jeder Ziehung ein normalverteilter Störterm über ausgewählte Merkmale gelegt. Dies wird so lange wiederholt, bis eine Gesamtheit entsteht, die hinreichend groß ist, um Stichproben gegebener Größe ziehen zu können (Auswahlsatz deutlich unter 0,05) (4). Konsequenterweise behält diese Gesamtheit im Großen und Ganzen die Zusammenhangsstruktur und das Ausfallmuster des jeweiligen ursprünglichen ALLBUS. Als nächstes entsteht das Problem der „wahren“ Werte für die zu schätzenden Parameter, das durch die Zuweisung von Werten gelöst wird. Das bedeutet, dass jeder fehlende Wert zunächst durch eine bestimmte Merkmalsausprägung ersetzt wird. Nach welchen Kriterien dies geschieht, wird in Abschnitt 4.3.1 noch erläutert: praktikabel erscheint eine Füllung der leeren Zellen mit „Extremtypen“, die teilweise in der Literatur dokumentiert sind. Nach der Ersetzung der fehlenden Werte ist es nun möglich, „wahre“ Werte für die künstliche Grundgesamtheit zu berechnen (5). Dies geschieht dann entsprechend den ausgewählten Parametern aus den drei Publikationen; das bedeutet, dass hier die statistischen Analysen noch einmal für die-

se Grundgesamtheit repliziert werden. Nachdem die „wahren“ Werte der Grundgesamtheit jetzt bekannt sind, werden für den simulativen Teil aus dieser Gesamtheit 1.000 Stichproben gezogen (6).[77] Die Stichprobengröße entspricht dabei genau der Größe der Stichproben, wie sie die einzelnen Analysen der drei Aufsätze aufweisen. Für jede Stichprobe wird sodann die jeweilige Analyse zum dritten Mal repliziert, wobei die Stichproben mit den Korrekturverfahren für Item bzw. Unit Nonresponse behandelt werden. Die replizierten Werte werden danach mit den „wahren“ Werten der Grundgesamtheit verglichen, um feststellen zu können, unter welchen Bedingungen welche Korrekturmethode bessere Ergebnisse erzielt. Aus diesem Grund war die Auswahl der Publikationen und der darin enthaltenen statistischen Analysen so wichtig: sie repräsentieren unterschiedliche Typen von Datensituationen und Parameterschätzungen.
Die Vorteile dieses neuen Vorgehens sind zusammengefasst folgende: reale Daten bilden den Ausgangspunkt für den Methodenvergleich, indem ein Datensatz erzeugt wird, der nahezu gleiche Zusammenhangsstrukturen und identische Ausfallmuster besitzt. Durch die Einsetzung von Merkmalsausprägungen lassen sich die „wahren“ Werte dieser Grundgesamtheit berechnen, die dann als Vergleichswerte für die Korrekturmethoden dienen.[78]

### 4.3.1 Konstruktion eines Stresstests

Um für die Grundgesamtheit, wie im vorhergehenden Abschnitt 4.3 beschrieben, „wahre“ Werte berechnen zu können, können die fehlenden Werte, deren Muster aus dem ursprünglichen ALLBUS übernommen wurden, durch Merkmalsausprägungen ersetzt werden. Doch welche Merkmalsausprägungen sollen dies sein? Für diesen Methodenvergleich soll die Antwort darin bestehen, die Methoden mit extremen Werten anstelle der fehlenden Werte zu konfrontieren. Die Methoden durchlaufen damit eine Art „Stresstest“. Der zugrunde liegende Ausfallmechanismus ist für empirische Verhältnisse tendenziell stark NMAR.[79] Die ausgewählten Extremwerte finden teilweise in der Literatur theoretische Begründungen als Manifestation möglicher Typen. Während für die Publikation von Eike Hennig nur zwei dieser Typen angewendet werden, wird für die multivariaten Analysen von Schnabel und Wagner sowie von Armin Schäfer ein Extremtyp mit dem jeweiligen Antitypus und als drittes ein Typ, der etwa in der Mitte angesiedelt ist, konstruiert. In den Übersichten wird noch einmal der jeweilige Ausfall in % ausgeführt sowie alle Merkmalswerte der Variablen. Je mehr Ausfall, desto eher wird ein extremer Wert als Ersatz für die fehlenden Werte den Parameter beeinflussen können. Und je mehr Merkmalswerte existieren, desto größer kann die Entfernung vom Durchschnitt ausfallen. Bei einer dichotomen Variable mit sehr niedrigem Ausfall wird die Beeinflussung durch extreme Werte wesentlich niedriger sein als bei einer Variablen mit sieben Merkmalswerten und einer vergleichsweise hohen Ausfallrate.

[77] Für die Stichprobenziehung werden die vorher aufgefüllten Werte in jeder Stichprobe wieder gelöscht. Eine aufwendige Identifikation ursprünglich fehlender Werte macht dies möglich.

[78] Nachteilig erscheint die Vorabauswahl der Parameter. Dies sollte aber durch die Vielzahl der Parameter kompensiert werden. Dennoch erreicht dieses Verfahren nicht die Flexibilität einer Simulation.

[79] Dennoch darf Multiple Imputation angewendet werden. Für diese Diskussion vgl. Spieß (2008), S.76f.

| Variable | Ausfall in % | Merkmalswerte | Typ 1 | Typ 2 |
|---|---|---|---|---|
| Schlussstrich | 2,2 | 1-2 | 1 | 2 |
| Anomie | 0,0 | 0-3 | - | - |
| Fremd im eig. Land wg. Aus. | 0,7 | 1-7 | 7 | 1 |
| Inflationsbekämpf. | 3,2 | 1-4 | 1 | 4 |
| Meinungsäußerung | 3,4 | 1-4 | 4 | 1 |
| Juden haben zu viel Einfl. | 11,7 | 1-7 | 7 | 1 |
| K. Scham über dt. Untaten an J. | 5,2 | 1-7 | 1 | 7 |
| J. nutzen dt. Vergangenheit aus | 10,2 | 1-7 | 1 | 7 |
| J. an Verfolgung nicht unschul. | 12,4 | 1-7 | 1 | 7 |

Tabelle 13: Bildung von Typen für Beispiel 1

Die ausgewählten Merkmale weisen zum einen sehr erhebliche Unterschiede im Umfang des Datenausfalls auf, zum anderen haben sie eine unterschiedliche Breite der Merkmalswerte. Dabei verbinden gerade die Antisemitismusvariablen eine relativ hohe Ausfallrate mit einer relativ breiten Ordinalskala von 1-7. Konterkariert wird dies durch die geschätzten bedingten Anteilswerte, wie bei der Vorstellung der Publikation beschrieben. Deshalb reichen zwei Typen bei der Evaluation der Korrekturmethoden aus, da jeglicher Wert je nach Definition entweder in die eine oder in die andere Kategorie fällt.
Der Publikation „Who Are the Workers Who Never Joined a Union? Empirical Evidence from Germany“ werden dann drei Typen zugeordnet:

| Variable | Ausfall in % | Merkmalswerte | Typ 1 | Typ 2 | Typ 3 |
|---|---|---|---|---|---|
| niemals Mitglied | 0,4 | 0-1 | 1 | 0 | 1 |
| Uniabschluss | 0,3 | 0-1 | 1 | 0 | 0 |
| niedrige Bildung | 0,8 | 0-1 | 0 | 1 | 1 |
| niedrige Bildung Vater | 9,6 | 0-1 | 0 | 1 | 0 |
| niedrige Bildung Mutter | 4,7 | 0-1 | 0 | 1 | 1 |
| Vater Arbeiter | 13,9 | 0-1 | 0 | 1 | 0 |
| politische Orientierung | 5,0 | 1-11 | 1 | 11 | 5 |
| Öffentlicher Dienst | 0,2 | 0-1 | 1 | 0 | 0 |

Tabelle 14: Bildung von Typen für Beispiel 2

Merkmale mit fehlenden Werten sind im Probitmodell bei Schnabel und Wagner vorwiegend dichotom. Die einzige Variable mit einer relativ großen Skalenspanne ist die *politische Orientierung* (Linksrechtsselbsteinstufung); diese besitzt auch einen nicht unerheblichen Ausfall von fünf

Prozent. Da es in der sozial- oder wirtschaftswissenschaftlichen Literatur keinen Entwurf von bestimmten konkreten Ausfalltypen gibt, sind die Extremtypen zwei Antitypen mit entsprechend extremen *politischen Orientierungen*. Der zusätzliche dritte Typ weist hingegen eine Mischung auf, deren Ausprägungen an den Vorzeichen der Parameter des Originaldatensatzes orientiert sind. Dieser Typ besitzt zudem eine durchschnittliche *politische Orientierung*. Gerade durch die Extremwerte des Merkmals *politische Orientierung* werden die ersten beiden Typen zu wirklichen Extremtypen, während Typ 3 ungefähr in der Mitte zwischen diesen beiden anzusiedeln ist.[80]

| Variable | Ausfall in % | Merkmalswerte | Typ 1 | Typ 2 | Typ 3 |
|---|---|---|---|---|---|
| Wahlteilnahme | 8,6 | 0-1 | 1 | 0 | 1 |
| Demokratiezuf. | 5,4 | 0-10 | 0 | 10 | 5 |
| Alter | 0,1 | stetig | −29,99 | 43,01 | 6,51 |
| niedrige Bildung | 1,3 | 0-1 | 1 | 0 | 0 |
| hohe Bildung | 1,3 | 0-1 | 0 | 1 | 0 |
| Einkommen | 44,5 | stetig | −1011,03 | 6363,97 | 2676,47 |

Tabelle 15: Bildung von Typen für Beispiel 3

In der Publikation von Armin Schäfer wird nicht nur die Demokratiezufriedenheit als 11-stufige Variable verwendet, sondern darüber hinaus die Variable Einkommen. Das monatliche Haushaltseinkommen ist eine stetige Variable. In der Literatur wird vermutet, dass gerade niedrige und hohe Einkommensbezieher keine Angabe zu ihrem Einkommen machen wollen (Becker und Hauser 2003, Pöschl 1993). Die Variable steht hochgradig im Verdacht, einen großen Umfang fehlender Werte aufzuweisen, die im direkten Zusammenhang mit der Ausprägung stehen, also deren Ausfallmechanismus stark zu NMAR tendiert. Zwar besitzt beispielsweise auch die Variable Wahlbeteiligung an der letzten Bundestagswahl einen nicht unerheblichen Anteil an Item Nonresponse von knapp neun Prozent, jedoch übertrifft die Einkommensvariable mit 44,5 % die Ausfälle aller anderen Merkmale bei weitem. Zudem wird das Merkmal durch eine offene Frage erhoben, sodass keine gesicherten Informationen über den höchsten oder niedrigsten Wert bestehen.[81] Aufgrund von Plausibilitätsüberlegungen kann nicht einmal der Fall ausgeschlossen werden, dass jemand ein negatives Einkommen hat, wenn z.B. Rückzahlungen größer sind als das eigentliche Einkommen. Um Merkmalsausprägungen für die Extremtypen zu erhalten, wurden deshalb für Typ 1 die niedrigsten (Typ 1) und höchsten (Typ 2) Merkmalswerte des Datensatzes anstelle der Item Nonresponse eingesetzt. Damit dürfte gewährleistet sein, dass die beiden Typen wirklich extrem sind, obwohl natürlich wesentlich höhere oder niedrigere Einkommen nicht unrealistisch sind (Millionäre etc.). Für Typ 3 wurde das arithmetische Mittel eingesetzt. Dieser dritte Typ ist ein moderater Typus bezogen auf die Merkmalswerte; für den ersten Typus gibt es zudem empirische Evidenz in

[80] Die negativen Werte resultieren aus der wie im Original vorgenommen Zentrierung der Verteilung um den Mittelwert.

[81] Daneben exisitert für diejenigen, die nicht geantwortet haben, die Möglichkeit, sich in einer Liste mit Einkommensklassen einzutragen. Diese Variable wurde aber nicht berücksichtigt.

der Literatur: der junge, ungebildete Nichtwähler mit niedrigem Einkommen, der dies verbergen möchte (De Nève 2009, S.86ff). Für die Korrekturmethoden ist diese Replikation sicherlich die qualitativ und quantitativ anspruchsvollste.

### 4.3.2 Ausgewählte Korrekturverfahren

Der offizielle Überblick über Veröffentlichungen, die den ALLBUS in irgendeiner Form zur Analyse herangezogen haben, verrät die momentane Relevanz von Korrekturverfahren für Item Nonresponse (Tabelle 16). Fast alle Veröffentlichungen schenken dem Problem keine besondere Beachtung und verwenden Complete Cases (Blohm et al. 2010). Die Ausnahme bildet eine Methodenarbeit mit dem ALLBUS 2004, die sich direkt mit Multipler Imputation beschäftigt. Da die Multiple Imputation nach mehreren Jahrzehnten nicht nur theoretisch ausgereift, sondern mittlerweile auch von einer immer größeren praktischen Relevanz ist, wird sie dem Verfahren der Complete Cases gegenübergestellt.[82] Durch die Implementierung verschiedener Multiple Imputation-Verfahren sowohl für stetige als auch für diskrete Variablen und für verschiedene Merkmalstypen, kann auch der nicht speziell geschulte Datennutzer diese Verfahren anwenden.

| | **ALLBUS 2002** | **ALLBUS 2004** | **ALLBUS 2006** |
|---|---|---|---|
| Anzahl | ca. 37 | ca. 24 | ca. 16 |
| MI verwendet | 0 | 1 | 0 |

Tabelle 16: Übersicht über die Verwendung von Multipler Imputation in Veröffentlichungen mit den ALLBUS-Erhebungen

Theoretisch ist die Multiple Imputation der Complete Cases-Methode überlegen. Die MAR-Annahme bei MI ist realistischer als die MCAR-Annahme bei CC. Für den Datennutzer von besonderer Wichtigkeit ist die Effizienz von MI: Die Standardfehler werden bei gegebener Fallzahl kleiner als bei einer Methode, die die Fallzahlen aufgrund von fehlenden Werten reduziert. Gleichzeitig reflektiert Multiple Imputation durch zusätzliche Varianz die durch die Ergänzung auftretende Unsicherheit. Ein Blick auf die Varianzregeln der Multiplen Imputation macht dies anschaulich (Rubin 1987):

$$T = W + \left(1 + \frac{1}{m}B\right),$$

wobei $W$ die durchschnittliche Varianz des geschätzten Parameters $\hat{\theta}_j$ der $j = 1, \ldots, m$ imputierten Datensätze ist, mit

[82] Für die Durchführung von MI existieren mittlerweile mehrere englischsprachige, leicht verständlich Bücher: McKnight et al. (2007) und die jeweiligen Beschreibungen zu den einschlägigen Softwares ICE (Royston 2005), MICE (Van Buuren und Groothuis-Oudshoorn 2009), NORM (Schafer 1997b); vereinzelt auch schon deutschsprachige Literatur: Spieß (2008).

$$W = \frac{1}{m}\sum_{j=1}^{m} \hat{var}\left(\hat{\theta}_{(j)}\right),$$

und $B$ die Varianz von $\hat{\theta}_j$ zu $\hat{\theta}_{MI}$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\theta}^{(j)} - \hat{\theta}_{MI}\right).$$

Für den Methodenvergleich wird zudem die Coverage berechnet, die auf MI-Konfidenzintervallen beruht:

$$\hat{\theta}_{MI} \pm t_{1-\frac{\alpha}{2},\nu}\sqrt{T},$$

wobei $\nu$ die Anzahl der Freiheitsgrade ist und $\nu$

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

Das Konfidenzintervalle für CC lautet dagegen:

$$\hat{\theta} \pm t_{1-\frac{\alpha}{2},\nu}\sqrt{Var}.$$

Die Implementierungen unterscheiden sich noch deutlicher in der Praxistauglichkeit. Prinzipiell scheinen die Pakete in R und ähnlicher Software nutzerfreundlicher und leistungsfähiger zu sein als in allgemeiner Software wie SPSS oder Stata. Für den folgenden Methodenvergleich wird deshalb auch die Software R herangezogen. Da der Methodenvergleich über eine Quasisimulation ein sehr ausgereiftes und effizientes Package benötigt, bietet sich das Package MICE an (Van Buuren und Groothuis-Oudshoorn 2009). Speziell für sozial- und wirtschaftswissenschaftlichen Datensätze, die großenteils diskrete Variable beinhalten, bietet sich zudem das Package BaBoon an (Koller-Meinfelder 2010).
MI wurde in dieser Arbeit mit MICE durchgeführt. Die Imputationsmodelle bestanden stets aus dem interessierenden Parameter (Analysemodell) und zusätzlichen vollständigen Variablen.[83] Neben der entsprechenden Anpassung an die Skalenniveaus der Variablen mit fehlenden Werten, wurden die Grundeinstellungen von $m = 5$ imputierten Datensätzen verwendet.

[83] Damit liegt der Regelfall der Unkongenialität zwischen Analyse- und Imputationsmodell vor; zum Begriff und zur formalen Darstellung vgl. Meng (2002).

### 4.3.3 Ergebnisse des Methodenvergleichs

Um die Fülle von Ergebnissen übersichtlich zu gestalten, wird für jede Parameterschätzung – Anteilswert, Probitmodell, Logitmodell – anfangs immer die Coverage der Multiplen Imputation und der Complete Cases tabellarisch aufgeführt. Die Coverage zeigt den Anteilswert an Schätzwerten, deren Konfidenzintervall den „wahren“ Wert beinhaltet. Dies erfolgt für jeden Typ.
Anschließend werden alle auffälligen Coveragewerte bezüglich der Punktschätzung und der Länge der Konfidenzintervalle untersucht. Auffällig bedeutet, dass entweder eine der Korrekturmethoden wesentlich besser abschneidet als die andere, oder beide Korrekturverfahren schlecht abschneiden. Dies geschieht durch grafische Aufbereitung.
Soweit es erforderlich ist, werden alle Unterschiede zwischen den Korrekturmethoden auf Signifikanz getestet; weitere Analysemöglichkeiten bestehen in der Berechnung des Bias und des MSE.[84]
Die Analysen werden jeweils für alle Parameterschätzungen durchgeführt. Ergebnisse, die keine weiteren, neuen Erkenntnisse liefern, finden sich tabellarisch und grafisch aufbereitet im Anhang.

[84] Definiert als: $BIAS(T_n) = \theta - E(T_n)$ und $MSE = E(T_n - \theta)^2 = Var(T_n) + BIAS(T_n)$.

#### 4.3.3.1 Ergebnis 1: Anteilswerte

Die erste Parameterschätzung wird aus einer bivariaten Verteilung gewonnen. Tabelle 17 enthält die Variablen, die mit der abhängigen Variablen *Schlussstrichziehen* gekreuzt wurden (Details in Abschnitt 4.2.1). In der ersten Zeile befindet sich jeweils ein bedingter Anteilswert – der Anteil der Personen, die für eine Schlussstrichziehung unter die deutsche NS-Vergangenheit sind und gleichzeitig z.B. der Aussage zustimmen, sich wegen Ausländern fremd im eigenen Land zu fühlen. Die zweite Zeile zeigt ebenfalls einen bedingten Anteilswert – den Anteil der Personen, die z.B. der Aussage zustimmen, sich wegen Ausländern fremd im eigenen Land zu fühlen und dabei eine Schlussstrichziehung befürworten. Die dritte und vierte bzw. fünfte und sechste Spalte führt die Coverage für Multiple Imputation (MI) bzw. für Complete Cases (CC) von Typ 1 und Typ 2 auf. In der zweiten Spalte findet sich noch einmal der gemeinsame Ausfall der Variable *Schlussstrichziehen* mit den jeweiligen Meinungsvariablen.

| | | Typ 1 | | Typ2 | |
|---|---|---|---|---|---|
| Schlussstrich mit | Ausfall | CC | MI | CC | MI |
| Fremd | 2,8 | $86{,}6^{ns}$ | $87{,}1^{ns}$ | $87{,}3^{*}$ | $88{,}7^{*}$ |
| | | $73{,}0^{ns}$ | $73{,}5^{ns}$ | $47{,}4^{***}$ | $43{,}7^{***}$ |
| Inflation | 5,0 | $52{,}7^{***}$ | $47{,}1^{***}$ | $69{,}0^{**}$ | $71{,}5^{**}$ |
| | | $76{,}5^{*}$ | $74{,}6^{*}$ | $60{,}9^{***}$ | $55{,}8^{***}$ |
| Meinung | 5,3 | $55{,}6^{*}$ | $59{,}3^{*}$ | $28{,}8^{***}$ | $21{,}6^{***}$ |
| | | $73{,}5^{**}$ | $77{,}2^{**}$ | $82{,}1^{***}$ | $77{,}4^{***}$ |
| Juden Einfluss | 12,4 | $0{,}0^{ns}$ | $0{,}0^{ns}$ | $1{,}1^{**}$ | $1{,}8^{**}$ |
| | | $40{,}9^{*}$ | $38{,}2^{*}$ | $70{,}6^{ns}$ | $69{,}4^{ns}$ |
| Keine Scham | 6,5 | $0{,}0^{ns}$ | $0{,}0^{ns}$ | $73{,}0^{***}$ | $65{,}9^{***}$ |
| | | $52{,}3^{*}$ | $49{,}8^{*}$ | $44{,}8^{**}$ | $42{,}2^{**}$ |
| Juden nutzen aus | 11,0 | $0{,}1^{ns}$ | $0{,}0^{ns}$ | $5{,}4^{***}$ | $9{,}8^{***}$ |
| | | $3{,}7^{*}$ | $4{,}7^{*}$ | $66{,}4^{**}$ | $69{,}2^{***}$ |
| Juden nicht unschuldig | 13,0 | $0{,}0^{ns}$ | $0{,}0^{ns}$ | $11{,}0^{***}$ | $7{,}0^{***}$ |
| | | $19{,}0^{***}$ | $12{,}4^{***}$ | $54{,}0^{***}$ | $47{,}0^{***}$ |
| Anomie | 2,1 | $88{,}7^{ns}$ | $89{,}5^{ns}$ | $88{,}8^{ns}$ | $88{,}8^{ns}$ |
| | | $72{,}0^{ns}$ | $71{,}1^{ns}$ | $49{,}6^{***}$ | $47{,}5^{***}$ |
| im Durchschnitt | | 43,4 | 42,8 | 52,5 | 50,5 |

Tabelle 17: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 1[85]

Mit Blick auf die Coverage als Kriterium fällt das Urteil über beide Methoden sehr schlecht aus. Zunächst zum direkten Vergleich der beiden Methoden nach den beiden Typen: Bei Typ 1 ist CC

[85] Die Signifikanzen beziehen sich auf die Testung der Coveragehöhe zwischen CC und MI.

insgesamt fünfmal signifikant besser als MI. MI zeigt bei drei Parameterschätzungen eine signifikant höhere Coverage. Bei den restlichen acht Parametern ergeben sich keine signifikanten Unterschiede. Noch stärker zu Ungunsten von MI als bei Typ 1 fallen die Ergebnisse bei Typ 2 aus. Hier schneidet MI nach der Coverage fünfmal besser ab, CC dagegen neunmal. Zweimal ergeben sich keine signifikanten Unterschiede in der Höhe der Coverage. Insgesamt ergeben sich für die 32 Parameter, dass in 13 Fällen CC eine höhere Coverage aufweist (44 %), in zehn Fällen gibt es keine signifikanten Unterschiede (ca. 31 %) und in lediglich acht Fällen schneidet MI besser ab (25 %). Diese Ergebnisse können aber mit Blick auf das Niveau der Coverage ziemlich täuschen. Weder bei MI noch bei CC liegt eine Coverage über 90 %. Bei den Coverage, die über 80 % liegen, gibt es keine signifikanten Unterschiede (Ausnahme Typ 2: *Fremd im eigenen Land*; *Inflationsbekämpfung*). Alle anderen Coverages sind indiskutabel niedrig.
Zwei Faktoren bestimmen anscheinend das schlechte Abschneiden beider Methoden bei der Schätzung der bedingten Anteilswerte: die Modellierung der Typen als Extreme in Richtung NMAR in Verbindung mit der Ausfallhöhe. Beide Methoden liegen bei den Antisemitismusvariablen, die die höchsten Ausfälle zeigen, deutlich neben den „wahren“ Werten.
Bevor die eigentliche Parameterschätzung betrachtet wird, um die die Konfidenzintervalle für die Coverage konstruiert werden, soll die Länge der Konfidenzintervalle in der grafischen Darstellung analysiert werden. Die Werte sind die Differenzen der jeweiligen Länge des MI-Konfidenzintervalls und des CC-Konfidenzintervalls.

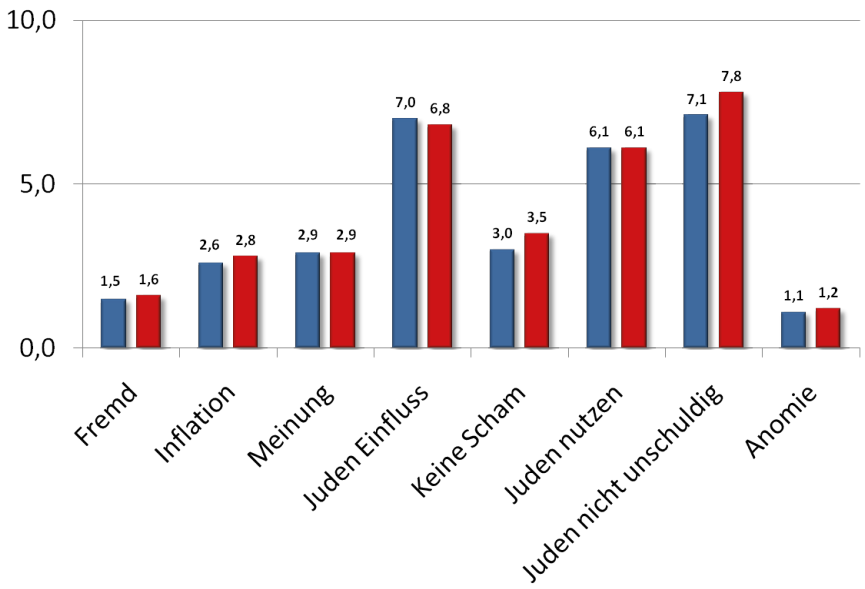



Abbildung 17: Differenz der Konfidenzintervalllängen bei Beispiel 1

Das positive Vorzeichen[86] der Differenzen in Abbildung 17 signalisiert, dass die MI-Konfidenzintervalle stets kleiner sind als die CC-Konfidenzintervalle (blauer Balken: 1. bedingter Anteils-

[86]Da die CC-KI vom jeweiligen MI-KI subtrahiert wurde.

wert, roter Balken: 2. bedingter Anteilswert). Die Stichprobenvarianz der MI-Schätzung ist aufgrund der nicht reduzierten Fallzahl niedriger – nach den Ausführungen aus Abschnitt 4.3.2 ist dies folgerichtig. Die Höhe der Abweichung variiert jedoch – und zwar mit der Ausfallhöhe. Bei den Variablen *Juden nutzen die Vergangenheit aus*, *Juden haben zu viel Einfluss* und *Juden nicht unschuldig an Verfolgung* sind die MI-Konfidenzintervalle teilweise bis zu acht Prozent kürzer als die entsprechenden CC-Konfidenzintervalle. Hier liegt zumindest ein Grund des schlechten Abschneidens von MI. Normalerweise würde – wie in der Varianzformel nachzuvollziehen ist – die MI-Varianz bei größerer Unsicherheit, also bei tendenziell größerem Ausfall, größer werden. Die Folge wäre eine Verringerung des Quotienten der CC- und MI-Intervalle. Diese Unsicherheit kann im Fall der bedingten Anteilswertschätzung dem Anschein nach nicht hinreichend abgebildet werden. Dies erschwert damit aber die Chance auf eine höhere Coverage von Grund auf.
Betrachtet man nun die Parameterschätzwerte, ergibt sich bezüglich der Performanz von CC ein negativeres Bild als das Coverage-Kriterium nahelegt. Der vermeintliche (kleine) Vorteil der Complete Cases Methode ist in der Praxis nicht relevant.

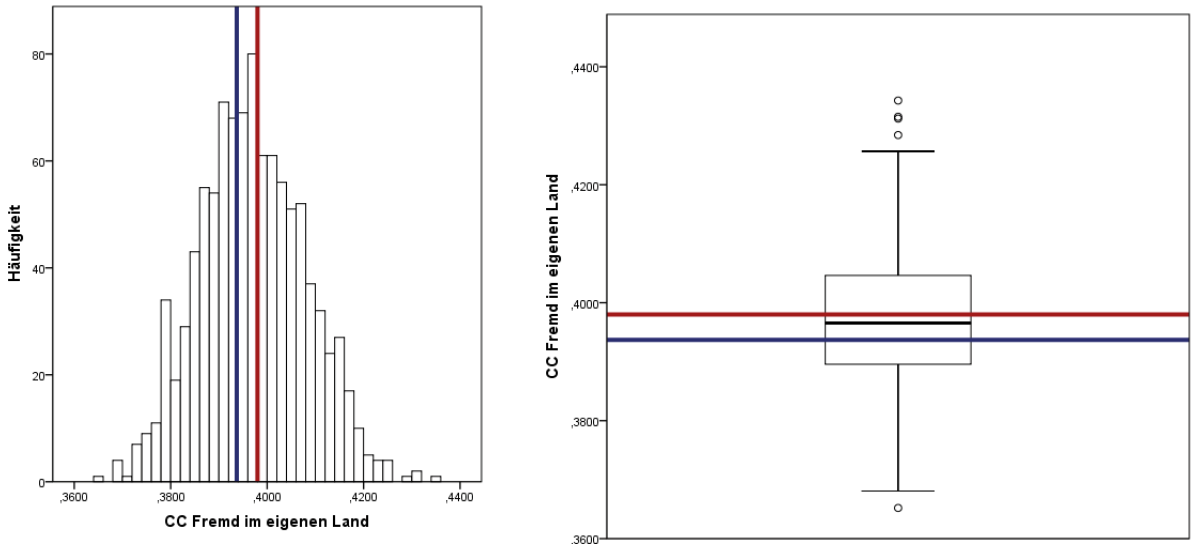



Abbildung 18: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Fremd im eigenen Land*

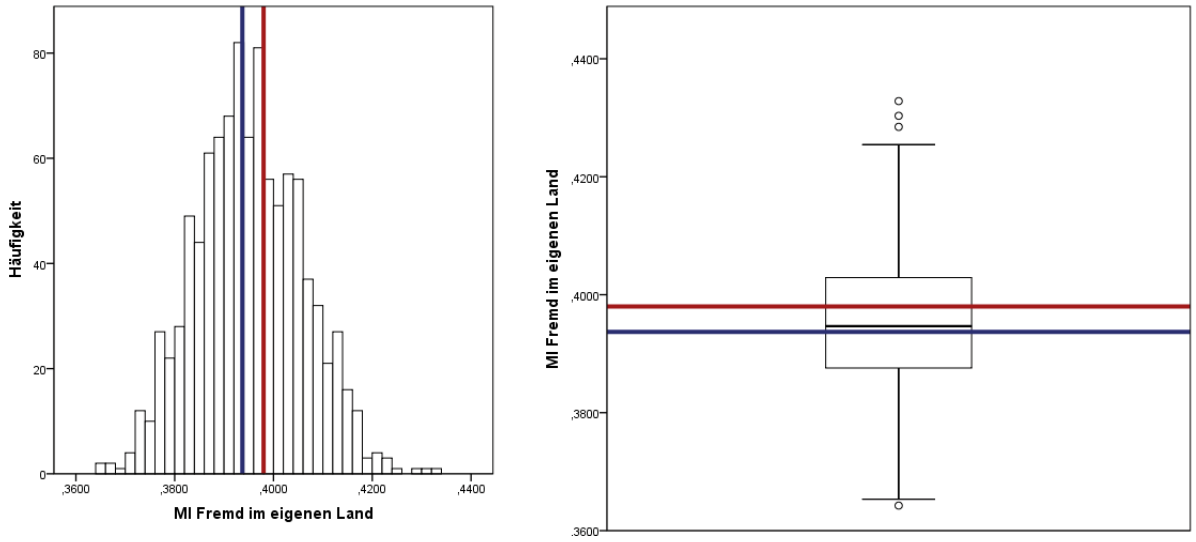



Abbildung 19: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Fremd im eigenen Land*

Bei der grafischen Aufbereitung bildet die Variable *Fremd im eigenen Land* das erste Parameterpaar (1. bedingter Anteilswert).[87] Die rote Linie markiert im Histogramm bzw. im Boxplot den „wahren“ Wert des Typs 1, die blaue Linie den „wahren“ Wert des Typs 2 (Abbildung 18 und 19).[88] Die Coverage bei beiden Typen und für beide Methoden war die höchste für die Schätzung der Anteilswerte.
Aufgrund des geringen Ausfalls liegen die beiden „wahren“ Werte nicht allzu weit voneinander entfernt. In den Histogrammen zeigt sich, wie sich sowohl die MI- als auch die CC-Schätzwerte mittig um die beiden Linien anordnen. Beide liegen durchschnittlich nicht weit von den „wahren“ Werten der beiden Typen entfernt (MI: 0,3951; CC: 0,3970).[89] Die Massen unterscheiden sich in ihrer Breite ebenfalls nur unwesentlich (MI-Standardabweichung: 0,1083; CC-Standardabweichung: 0,1104). Diese Ergebnisse sind ziemlich kongruent mit den entsprechenden Coverages.
Abbildung 20 zeigt eine Parameterschätzung, bei der MI auf einem niedrigen Niveau eine signifikant höhere Coverage bei Typ 1 und auf einem noch niedrigeren Niveau eine signifikant niedrigere Coverage bei Typ 2 aufweist. Es ist der erste bedingte Parameter der Variable *freie Meinungsäußerung*.

[87] Hier nicht aufgeführte grafische Darstellungen finden sich in Anhang 5.

[88] Für die einzelnen Typen wurden an sich jeweils getrennte Durchläufe à 1.000 berechnet, die in der Tabellenübersicht (Tabelle 18 und 19) ausgewertet werden; für die grafische Darstellung reicht einer der Durchläufe, da die Verteilung der Parameterschätzwerte wie erwartet keine großen Abweichungen zeigen.

[89] Wenn nicht explizit erwähnt, ist nicht der BIAS, sondern der Lage- bzw. Streuungsparameter der Schätzwerteverteilung gemeint.

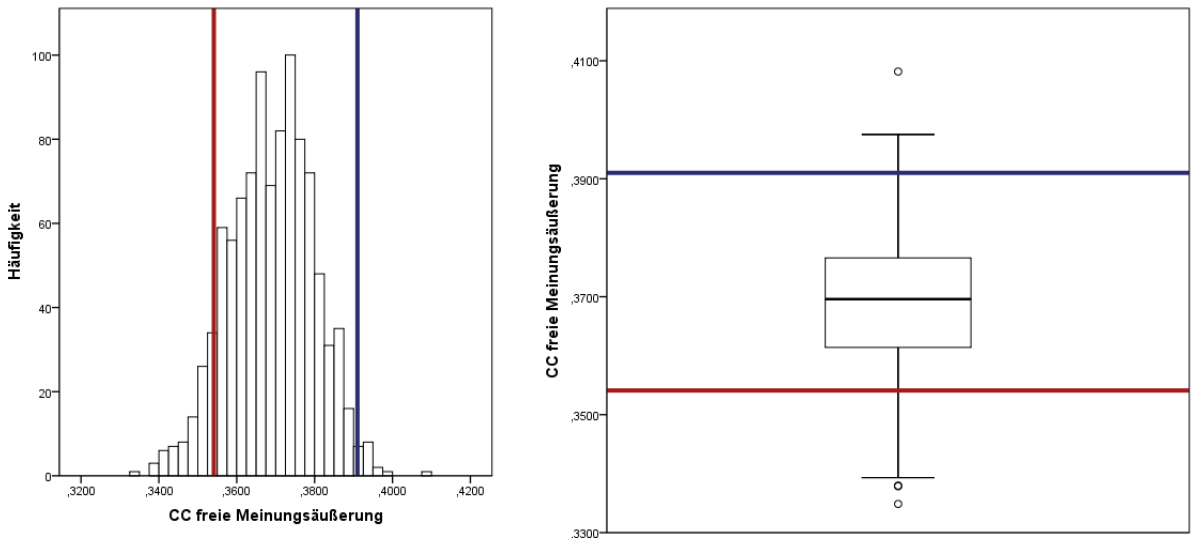



Abbildung 20: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Freie Meinungsäußerung*

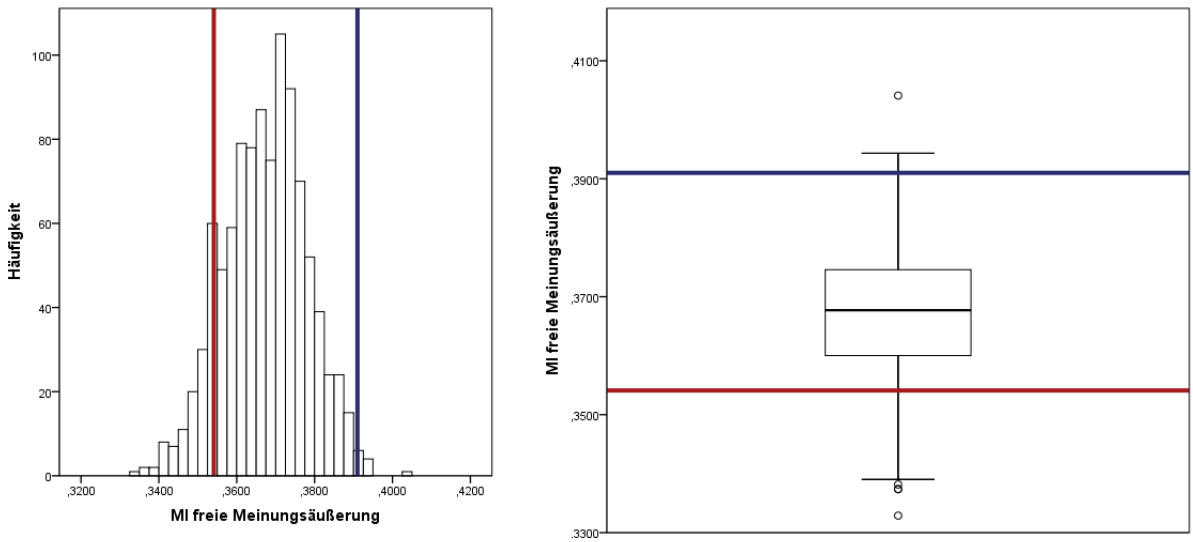



Abbildung 21: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Freie Meinungsäußerung*

Aus der Verteilung im Boxplot und dem Histogramm lassen sich kaum Unterschiede in der Gestalt der Verteilungen herauslesen, tatsächlich besitzt die Verteilung der MI-Schätzwerte ein Mittel von 0,3673 (Standardabweichung: 0,0108), CC dagegen 0,3690 (Standardabweichung: 0,0107). Beide Verfahren verschätzen sich für beide Typen. Bei Typ 1 liegt eine Überschätzung vor, bei Typ 2 eine noch gravierendere Unterschätzung der jeweils „wahren“ Werte. Selbst wenn die Coverage für Typ 1 und Typ 2 signifikant zu Gunsten der einen bzw. der anderen Korrekturmethode ausfällt, dürften diese Unterschiede in der Praxis keine Relevanz haben.
Diese Verzerrung fällt im folgenden Beispiel noch wesentlich gravierender aus. Für die Variable „Juden haben zu viel Einfluss“ (1. bedingter Anteilswert) beträgt die Coverage von MI und von CC null für den Typ 1 und nahe null für Typ 2. Die Grafiken 22 und 23 verdeutlichen, wie sehr dabei der „wahre“ Wert unterschätzt (Typ 1) oder überschätzt (Typ 2) wird. Die Gestalt der Schätzwertverteilung von MI und CC unterscheiden sich dabei wiederum sehr wenig (MI: 0,3866 mit Standardabweichung 0,0118; CC: 0,39 mit Standardabweichung 0,0118).

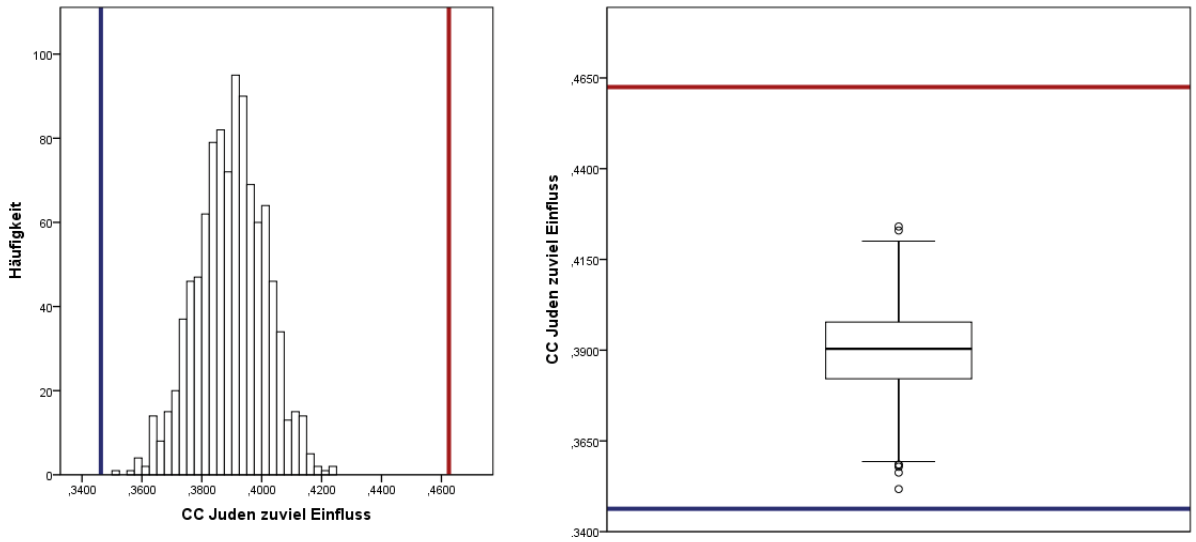



Abbildung 22: CC Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Juden zuviel Einfluss*

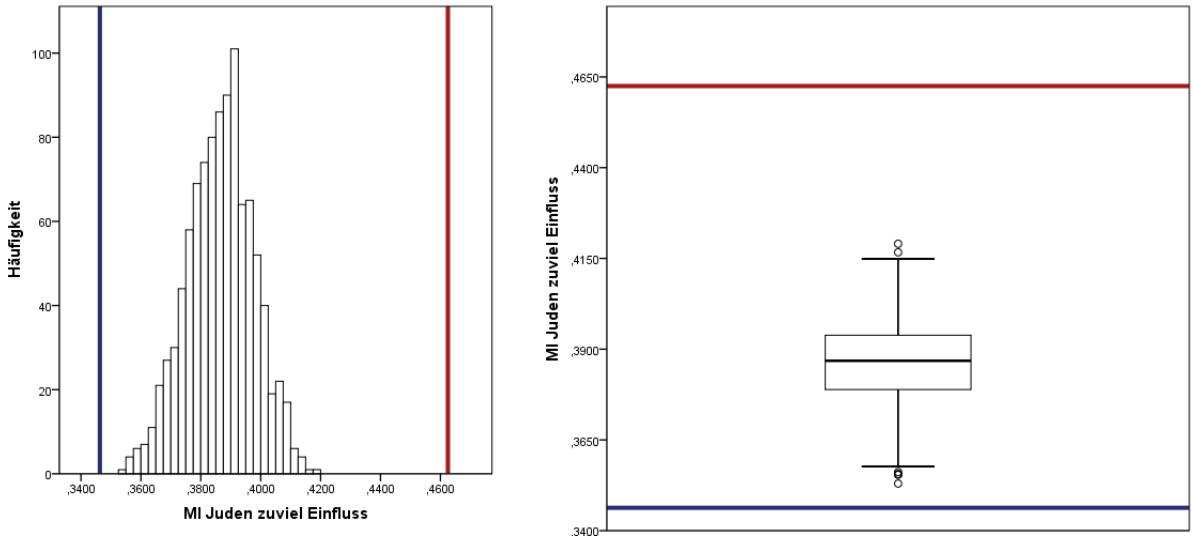



Abbildung 23: MI Histogramm und Boxplot des bedingten Anteilswerts zur Variablen *Juden zuviel Einfluss*

Um einen abschließenden Überblick zu erhalten, werden in den Tabellen 18 und 19 die mittleren Abweichungen von den „wahren“ Werten und die Variation für die mittlere Abweichung der Schätzwerte (Bias) für beide Methoden bei Typ 1 und Typ 2 abgetragen:

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| Fremd | 0,0010 | 0,01104 | 0,0028 | 0,01038 |
| | 0,0447 | 0,01266 | 0,0034 | 0,01257 |
| Inflation | 0,0159 | 0,01800 | 0,0171 | 0,01796 |
| | −0,0011 | 0,01346 | −0,0024 | 0,01336 |
| Meinung | −0,0149 | 0,01083 | −0,0132 | 0,01070 |
| | 0,0076 | 0,01416 | 0,0027 | 0,01403 |
| Juden Einfluss | 0,0726 | 0,01145 | 0,0759 | 0,01115 |
| | −0,0182 | 0,01423 | −0,0184 | 0,01349 |
| Keine Scham | 0,0516 | 0,00820 | 0,0502 | 0,00814 |
| | −0,0035 | 0,02052 | −0,0044 | 0,02003 |
| Juden nutzen | 0,0514 | 0,01149 | 0,0565 | 0,01128 |
| | −0,0382 | 0,01229 | −0,0351 | 0,01205 |
| Juden nicht unschuldig | 0,1012 | 0,01042 | 0,1003 | 0,01017 |
| | −0,0304 | 0,01737 | −0,0335 | 0,01642 |
| Anomie | −0,0021 | 0,01041 | −0,0014 | 0,01031 |
| | −0,0014 | 0,01357 | 0,0043 | 0,01355 |

Tabelle 18: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| Fremd | −0,0036 | 0,01049 | −0,0019 | 0,01038 |
| | −0,0150 | 0,01315 | −0,0164 | 0,01306 |
| Inflation | −0,0123 | 0,01046 | −0,0110 | 0,01047 |
| | −0,0119 | 0,01354 | −0,0133 | 0,01340 |
| Meinung | 0,0228 | 0,01010 | 0,0244 | 0,01009 |
| | −0,0006 | 0,01322 | −0,0056 | 0,01297 |
| Juden Einfluss | −0,0440 | 0,01154 | −0,0406 | 0,01163 |
| | −0,0051 | 0,01373 | −0,0054 | 0,01312 |
| Keine Scham | −0,0081 | 0,00797 | −0,0095 | 0,00791 |
| | −0,0149 | 0,02037 | −0,0157 | 0,01967 |
| Juden nutzen | −0,0369 | 0,01169 | −0,0317 | 0,01140 |
| | −0,0085 | 0,01247 | −0,0055 | 0,01216 |
| Juden nicht unschuldig | −0,0282 | 0,00990 | −0,0291 | 0,00970 |
| | −0,0107 | 0,01716 | −0,01308 | 0,01609 |
| Anomie | −0,0003 | 0,01051 | 0,0005 | 0,01048 |
| | −0,0153 | 0,01337 | −0,0159 | 0,01329 |

Tabelle 19: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI

Die Visualisierung der Schätzergebnisse in den Grafiken und die mittlere Abweichung der geschätzten Werte vom „wahren“ Wert zeigen, dass auch signifikant höhere Coverages praktisch keine relevante Bedeutung haben müssen. Für die Schätzung der bedingten Anteilswerte kann keines der beiden Verfahren wirklich empfohlen werden. Obwohl der Anteilswert als „einfacher“ Parameter angesehen wird (bedingt univariat), haben MI und CC massive Probleme bei der Schätzung. Die Erklärung hierfür ist relativ einfach: Die MCAR-Annahme für CC ist unrealistisch und für die Konstruktion des Stresstests, dessen Konzept NMAR ist, ungeeignet. Die Tendenz zu NMAR stellt auch eine Herausforderung für MI dar, die hier allem Anschein nach nicht bewältigt wird.
Relativierend ist die Tatsache, dass die Typen, gegen die beide Verfahren getestet wurden, für Anteilswerte extrem sind. Auch angesichts der geschätzten Parameter müsste der Datennutzer selbst entscheiden, welchen Spielraum des Irrtums er subjektiv seinen Ergebnissen erlauben will.

#### 4.3.3.2 Ergebnis 2: Parameter des Probitmodells

Anders als bei der ersten Parameterschätzung, die sich mit dem bedingten Anteilswert im uni- bzw. bivariaten Raum bewegt hat, wird nun ein multivariates Modell mit 14 Parametern und Achsenabschnitt (Details in Abschnitt 4.2.1) für CC und MI bei nunmehr drei Typen verglichen. Zur Erinnerung: zwei der Typen (Typ 1 und Typ 2) sind extreme Typen, der Typ 3 liegt gemäßigt dazwischen (Details in Abschnitt 4.3.1). Tabelle 20 verzeichnet wiederum den Ausfall der einzelnen Variablen und den Gesamtausfall (zweite Spalte) sowie die Coverage für die drei Typen nach CC und MI.

| | | Typ 1 | | Typ 2 | | Typ 3 | |
|---|---|---|---|---|---|---|---|
| | Ausfall | CC | MI | CC | MI | CC | MI |
| niem. Gew.-Mitgl. | 0,4 | $91{,}1^{*}$ | $93{,}4^{*}$ | $93{,}8^{ns}$ | $94{,}2^{ns}$ | $93{,}8^{ns}$ | $94{,}4^{ns}$ |
| Geschlecht | 0,0 | $92{,}3^{ns}$ | $94{,}1^{ns}$ | $91{,}5^{ns}$ | $91{,}8^{ns}$ | $89{,}4^{***}$ | $94{,}8^{***}$ |
| Alter | 0,0 | $92{,}4^{ns}$ | $92{,}2^{ns}$ | $93{,}3^{ns}$ | $93{,}4^{ns}$ | $93{,}0^{ns}$ | $93{,}7^{ns}$ |
| $\text{Alter}^2$ | 0,0 | $94{,}2^{*}$ | $92{,}0^{*}$ | $92{,}4^{ns}$ | $93{,}8^{ns}$ | $91{,}8^{**}$ | $94{,}1^{***}$ |
| jem. arbeitslos | 0,0 | $86{,}1^{***}$ | $93{,}9^{***}$ | $86{,}0^{***}$ | $95{,}4^{***}$ | $88{,}1^{***}$ | $94{,}3^{***}$ |
| Bild.: niedrig | 0,8 | $93{,}1^{ns}$ | $95{,}7^{ns}$ | $92{,}1^{**}$ | $95{,}1^{**}$ | $90{,}1^{***}$ | $95{,}5^{***}$ |
| Uni.-Abschluss | 0,3 | $92{,}2^{***}$ | $96{,}2^{***}$ | $95{,}4^{ns}$ | $96{,}0^{ns}$ | $94{,}5^{ns}$ | $95{,}5^{ns}$ |
| Bild. Vt.: niedrig | 9,6 | $78{,}0^{ns}$ | $76{,}9^{ns}$ | $93{,}2^{ns}$ | $93{,}2^{ns}$ | $92{,}5^{ns}$ | $93{,}6^{ns}$ |
| Bild. M.: niedrig | 4,7 | $81{,}9^{***}$ | $90{,}5^{***}$ | $95{,}0^{ns}$ | $94{,}5^{ns}$ | $96{,}3^{ns}$ | $95{,}2^{ns}$ |
| Vt. Arbeiter | 13,9 | $87{,}1^{*}$ | $85{,}1^{*}$ | $92{,}7^{ns}$ | $92{,}5^{ns}$ | $91{,}7^{ns}$ | $91{,}3^{ns}$ |
| Arbeiter | 0,0 | $92{,}2^{***}$ | $95{,}7^{***}$ | $90{,}6^{***}$ | $94{,}3^{***}$ | $93{,}1^{ns}$ | $94{,}5^{ns}$ |
| Ostdeutsch | 0,0 | $86{,}2^{***}$ | $94{,}6^{***}$ | $74{,}2^{***}$ | $95{,}7^{***}$ | $76{,}0^{***}$ | $94{,}7^{***}$ |
| pol. Orientierung | 5,0 | $71{,}5^{*}$ | $69{,}0^{*}$ | $88{,}4^{ns}$ | $87{,}6^{ns}$ | $96{,}4^{ns}$ | $96{,}2^{ns}$ |
| Vollzeit | 0,0 | $96{,}0^{ns}$ | $96{,}0^{ns}$ | $95{,}1^{ns}$ | $95{,}0^{ns}$ | $95{,}7^{ns}$ | $96{,}3^{ns}$ |
| Ö. Dienst | 0,2 | $93{,}6^{ns}$ | $93{,}0^{ns}$ | $90{,}4^{***}$ | $95{,}3^{***}$ | $89{,}7^{***}$ | $94{,}4^{***}$ |
| | 23,6 | 88,5 | 90,6 | 90,9 | 93,9 | 91,5 | 94,6 |

Tabelle 20: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 2

Schaut man sich die Coverages in Tabelle 20 getrennt nach Typen an, schneidet bei Typ 1 MI noch am schlechtesten ab. MI hat nur doppelt so oft wie CC eine signifikant höhere Coverage – nämlich sechsmal. Bei Typ 2 schneidet MI fünfmal signifikant besser ab, CC kein einziges Mal; ebenso wenig bei Typ 3. Hier hat MI wiederum sechsmal eine signifikant höhere Coverage. Damit zählt man für MI bei 17 von 45 Parametern (31 %) über die drei Typen eine signifikant bessere Coverages, für CC lediglich 3 von 45 (7 %); diese 3 Parameter weisen zudem eine schwache Signifikanz auf. Bei 56 % der Schätzungen gab es keinen signifikanten Unterschied. Betrachtet man nun das

Coverageniveau, wird das Bild deutlicher: Während bei Typ 1 von 15 Parametern 12 Parameter über 90 % der Konfidenzintervalle den „wahren“ Wert überdecken, schafft dies CC bei nur sechs Parametern. Sechsmal schafft MI sogar ca. 95 % zu schätzen – was den theoretisch bestmöglichen Wert entspricht;[90] CC gelingt dies nur zweimal. Typ 2 gewährt MI 14 von 15 Parametern über 90 % richtig zu schätzen, bei CC sind es 12. Geht man auf das Niveau von ca. 95 %, erreicht dies MI bei 9 Parametern, CC bei nur zweien. Waren die beiden ersten Typen von extremer Natur, so ist Typ 3 gemäßigt. Dies wirkt sich vor allem für MI aus: MI weist für alle Parameter eine Coverage von 90 % oder besser auf, wohingegen die CC-Coverage bei nur 11 von 15 Parametern über 90 % liegt. Sogar dreizehnmal schätzt MI zu etwa 95 % richtig, CC lediglich viermal. Diese oberflächliche Analyse macht bereits die Überlegenheit von MI bei der Schätzung des multivariaten Probitmodells deutlich – deutlicher als die durchschnittlich 2,1 bis 3,1 %-Punkte bessere Coverage. Es gibt aber durchaus Unterschiede bei der Schätzung einzelner Parameter.
Bei einigen Variablen schätzen MI und CC auf hohem Niveau sehr gut: *Vollzeit* über alle Typen, etwas schlechter *Alter* und mit Ausnahme des ersten Typs für den Achsenabschnitt (hier ist MI schwach signifikant besser). Daneben gibt es Variablen, die alle bis zu einem gewissen Grad Ausfälle zeigen und teilweise relativ niedrige Coverage aufweisen. Vor allem bei Typ 1 und teilweise bei Typ 2 ist dies bei den Variablen *politische Orientierung*, *Vater niedrige Bildung*, *Vater Arbeiter* und *Mutter niedrige Bildung* der Fall. Die Variablen *Alter*$^2$, *Bildung niedrig* sowie *Universitätsabschluss* zeigen unspektakuläre Ergebnisse über alle Typen bei beiden Verfahren, wobei insgesamt MI die bessere Performanz besitzt. Eindeutige Ergebnisse zeigen die Variablen *Geschlecht*, *Ostdeutsch* und *Öffentlicher Dienst*. Bei der ersten und letzten Variablen zeigt CC ein unsystematisches Verhalten. Der gemäßigte Typ 3 wird von CC sogar am schlechtesten geschätzt. Sehr schlecht fallen für CC die Ergebnisse bei *Ostdeutsch* aus. Ebenfalls zeigt sich hier das unsystematische Verhalten, das bei CC für dieses Modell öfter zu sehen ist.

[90] Die Coverage wurde mit einem Signifikanzniveau von 95 % geschätzt.

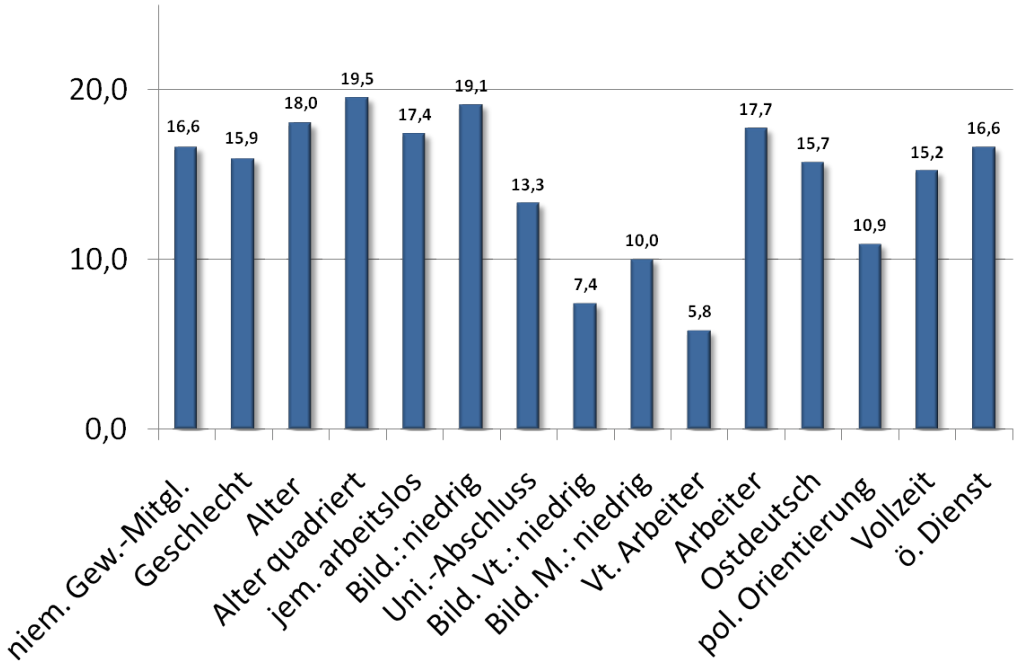



Abbildung 24: Differenz der Konfidenzintervalllängen bei Beispiel 2

Setzt man die Ergebnisse noch in Bezug zur Länge der jeweiligen Konfidenzintervalle, ergibt sich mustergültig, was nach der Theorie sein muss (Abbildung 24): Die Länge der CC-KI ist bei den Variablen mit sehr geringem Ausfall oder mit keinem Ausfall zwischen 15 % und 20 % größer als die Länge der MI-KI. Mit steigendem Ausfall wird der Quotient geringer, da MI die Unsicherheit berücksichtigt. Folglich besitzt die Variable *Vater Arbeiter* mit dem höchsten Ausfall (13,9 %) auch den geringsten Unterschied zwischen den MI- und CC-Konfidenzintervallen (lediglich 5,8 %).
Wie für die Anteilswerte sollen auch für die Parameter des Probitmodells einzelne auffällige Ergebnisse mit den geschätzten Parameterwerten visualisiert werden, um besser verstehen zu können, warum die Ergebnisse so ausfallen.[91] Begonnen werden soll mit der Variablen $Alter^2$, die mit Blick auf die Coverage MI leichte Probleme zu bereiten scheint.

[91] Hier nicht aufgeführte grafische Darstellungen finden sich in Anhang 6.

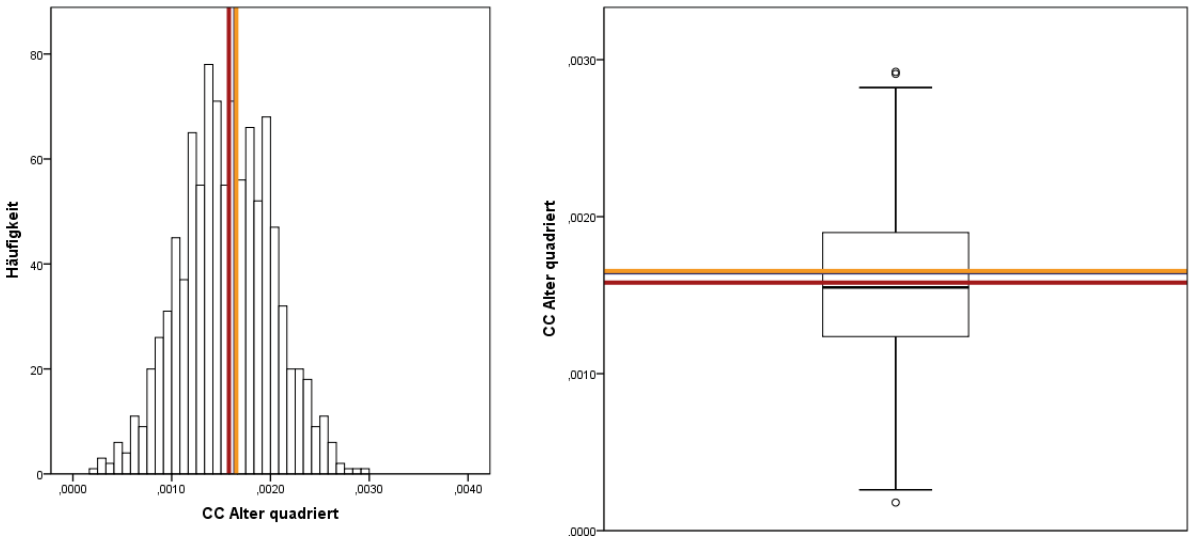



Abbildung 25: CC Histogramm und Boxplot des Parameters zur Variablen *Alter*$^2$

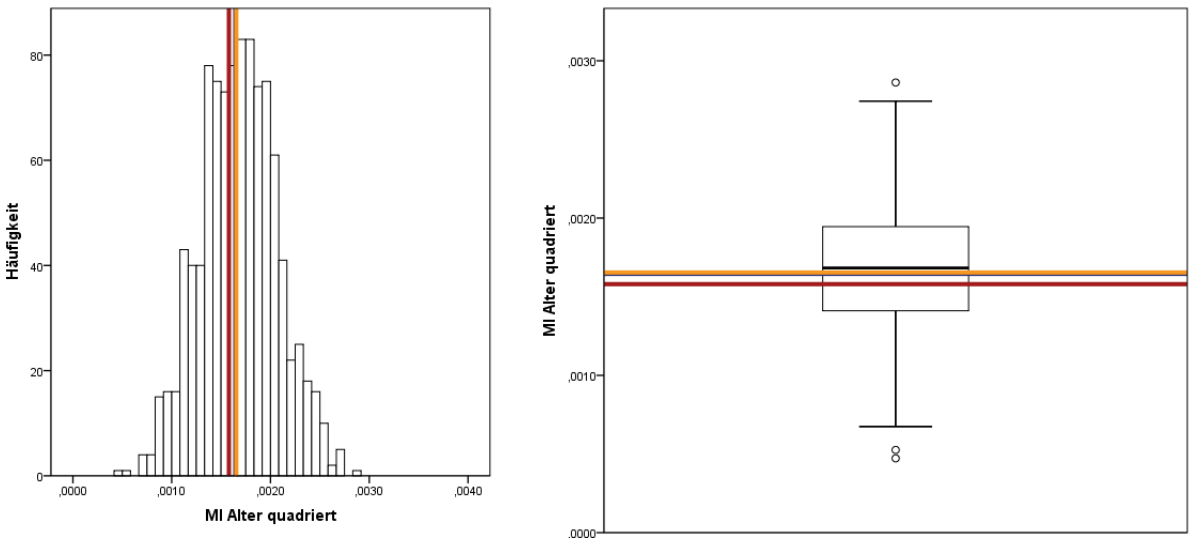



Abbildung 26: MI Histogramm und Boxplot des Parameters zur Variablen *Alter*$^2$

*Alter* und *Alter*$^2$ sind die einzigen stetigen Variablen des Modells. In der Visualisierung sind zwei Dinge erwähnenswert (Abbildung 25 und 26): Zunächst lässt sich konstatieren, dass beide Methoden im Schnitt kaum Abweichungen von den „wahren“ Werten aufweisen (MI: 0,0016799 und CC: 0,0015580). Auffällig ist jedoch, dass die Gestalt der MI-Werteverteilung schlanker wirkt und sich dies in der Standardabweichung durchaus niederschlägt (MI-Standardabweichung: 0,00039 und CC-Standardabweichung: 0,00046). Die Zahlen signalisieren, dass die vermeintliche leichte Schwäche von MI z.B. bei diesem Parameter zu vernachlässigen ist.
Als nächstes betrachten wir die Schätzwerte von MI und CC für die Variable *Arbeitslos*.

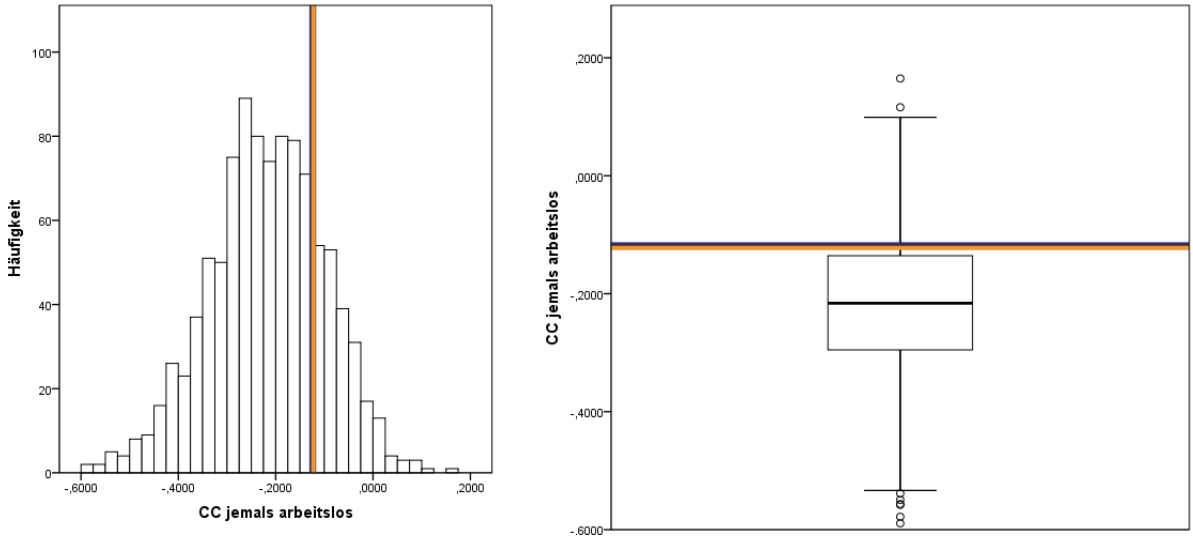



Abbildung 27: CC Histogramm und Boxplot des Parameters zur Variablen *Arbeitslos*

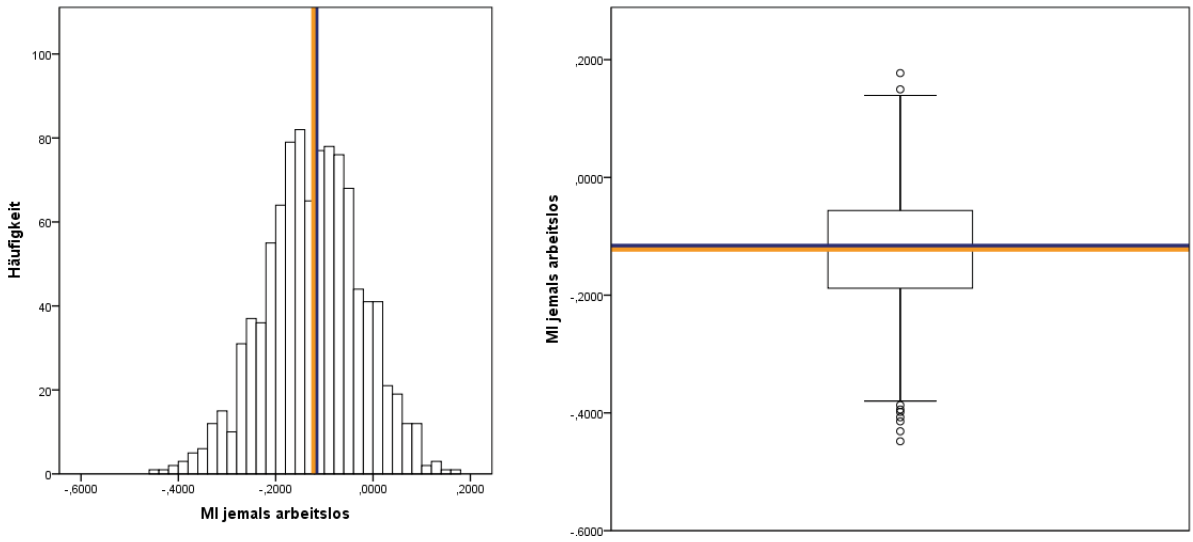



Abbildung 28: MI Histogramm und Boxplot des Parameters zur Variablen *Arbeitslos*

Im Histogramm und im Boxplot (Abbildung 27 bzw. 28) sieht man, wie gut MI die „wahren“ Parameter trifft (durchschnittlich wird der Wert -0,1239 geschätzt), während die geschätzten CC-Parameterwerte einen zu starken negativen Zusammenhang anzeigen würden (-0,2185). Verstärkt wird diese Verzerrung durch die Gestalt: die Standardabweichung der MI-Schätzwerte beträgt 0,0996, die der CC-Schätzwerte dagegen 0,1191. Die unabhängige Variable besitzt an sich keine fehlenden Werte. Wie kommt es dann, dass CC derart falsch schätzt? Dies liegt an der Fallreduktion, die bei der Variablen *jemals arbeitslos* anscheinend zu einer starken Verzerrung führt, wohingegen MI stabil über alle Typen die Information aller Fälle ausnutzt.
Die nächste unabhängige Variable des Modells – *Vater niedrige Bildung* – weist einen Ausfall von knapp 10 % auf.

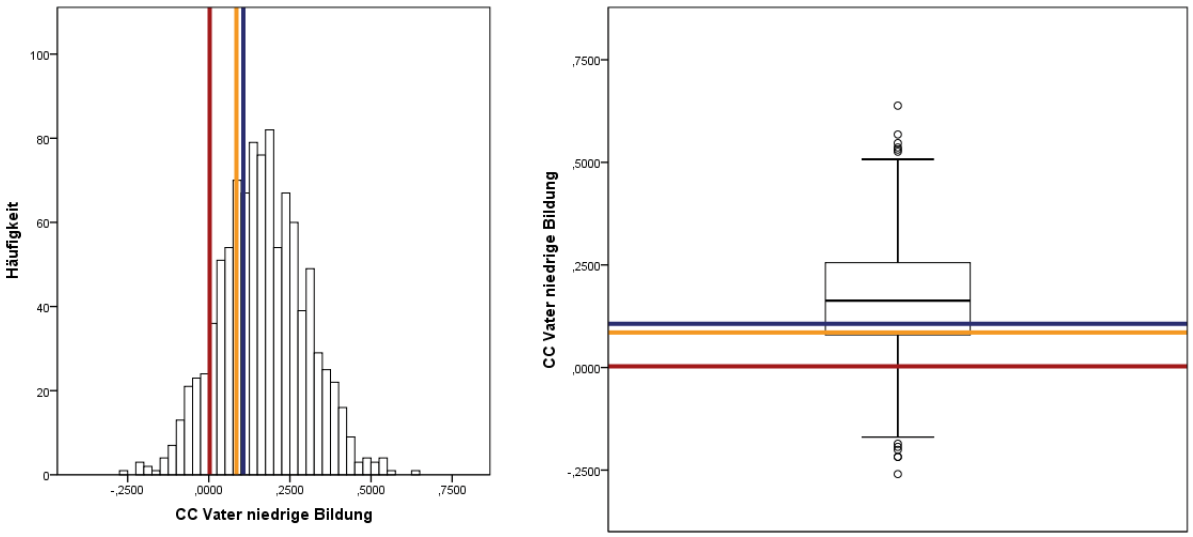



Abbildung 29: CC Histogramm und Boxplot des Parameters zur Variablen *Vater niedrige Bildung*

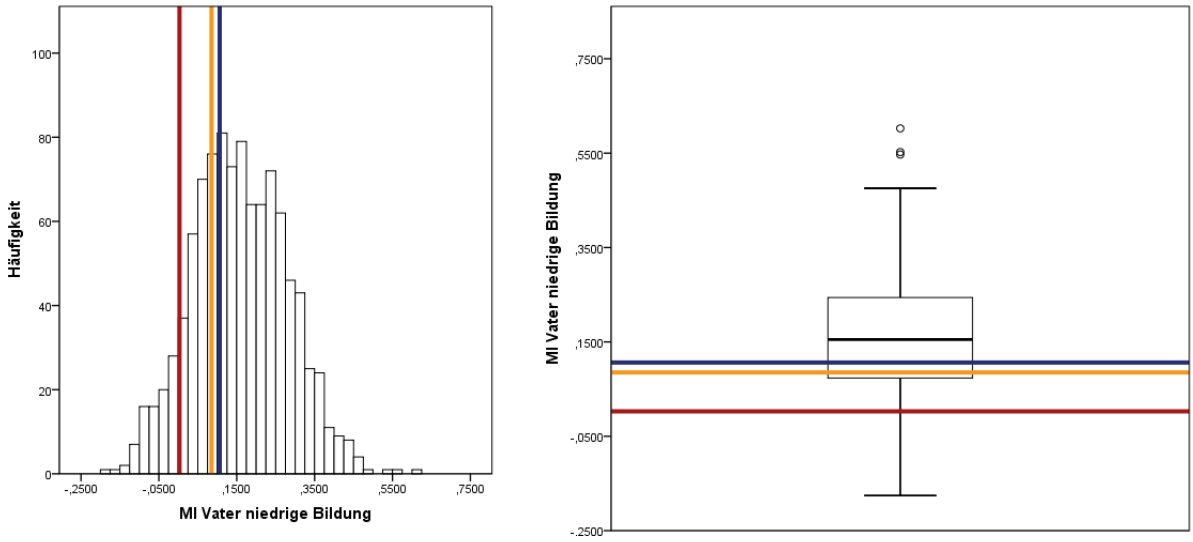



Abbildung 30: MI Histogramm und Boxplot des Parameters zur Variablen *Vater niedrige Bildung*

Gerade bei Typ 1 wird mit dieser Variable deutlich, wie sehr ein Extremtyp die Coverage senken kann. Zur Erinnerung: Typ 1 unterstellt für alle fehlenden Werte die Ausprägung 0 (keine niedrige Bildung). Realistischer ist sicherlich Typ 2, bei dem der Befragte aufgrund von Scham die niedrige Bildung des Vaters nicht angeben will. Beide Verfahren kommen mit Typ 2 und 3 jedoch wesentlich besser zu recht. Im Durchschnitt überschätzt MI den Einfluss der Variablen auch weniger als CC (0,1590 zu 0,1674) bei gewohnt niedrigerer Standardabweichung von MI (0,1210 zu 0,1325). Der Unterschied in der Standardabweichung ist konsequenterweise aufgrund des nicht geringen Ausfalls von 10 % geringer als bei den bisher analysierten Variablen. Die nächste Variablen *Vater Arbeiter* hat mit fast 14 % Ausfall die meisten fehlenden Werte aller Explanantia.

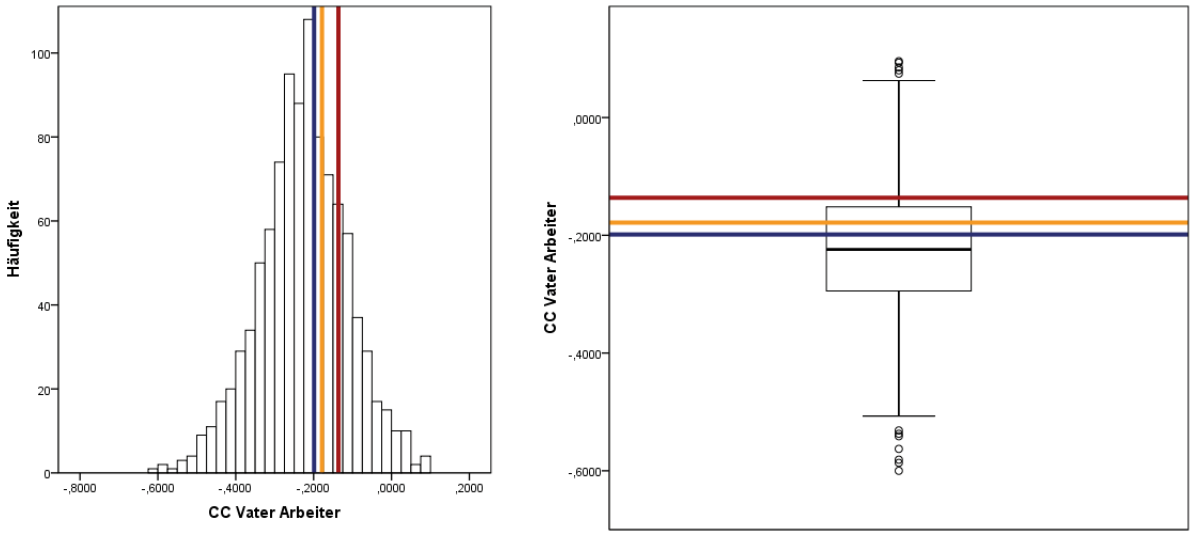



Abbildung 31: CC Histogramm und Boxplot des Parameters zur Variablen *Vater Arbeiter*

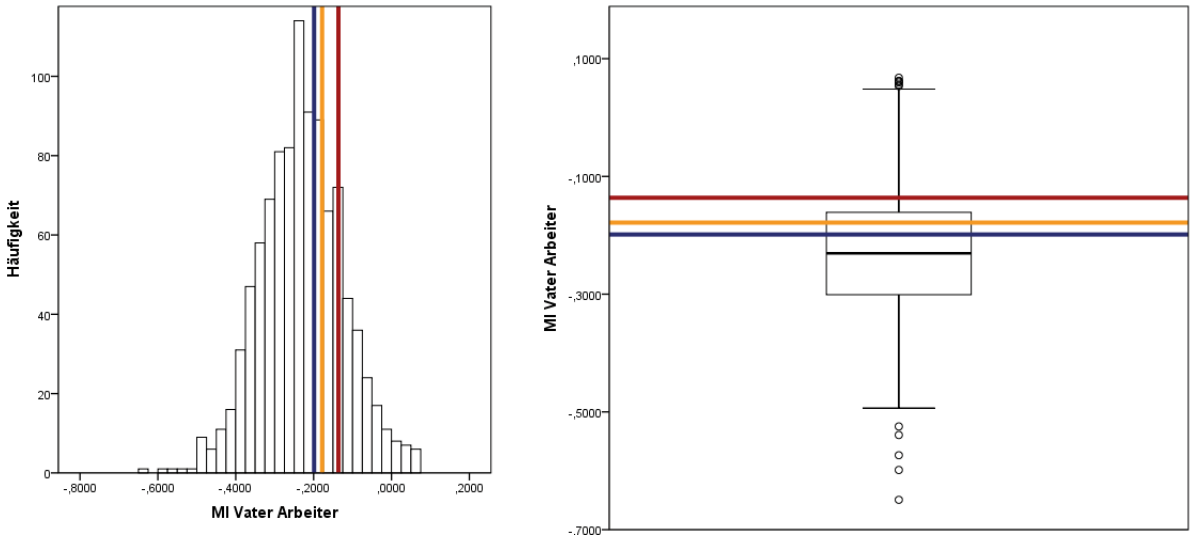



Abbildung 32: MI Histogramm und Boxplot des Parameters zur Variablen *Vater Arbeiter*

Die grafische Darstellung der von MI und CC geschätzten Parameterwerte zeigen keine großen Unterschiede (Abbildung 31 und 32). Beide Methoden führen bei dieser Variable zu einer Überschätzung ihres negativen Einflusses auf die abhängige Variable (MI: -0,2298 und CC: -0,2264). Je eher Typ 1 der Realität entsprechen sollte, desto größer wäre auch die Überschätzung. Die nun betrachtete Variable *Ostdeutsch* zeigt in der Coverage für MI sehr gute und stabile Werte, CC schätzt unsystematisch und ziemlich schlecht. Grafisch umgesetzt wird dies noch deutlicher:

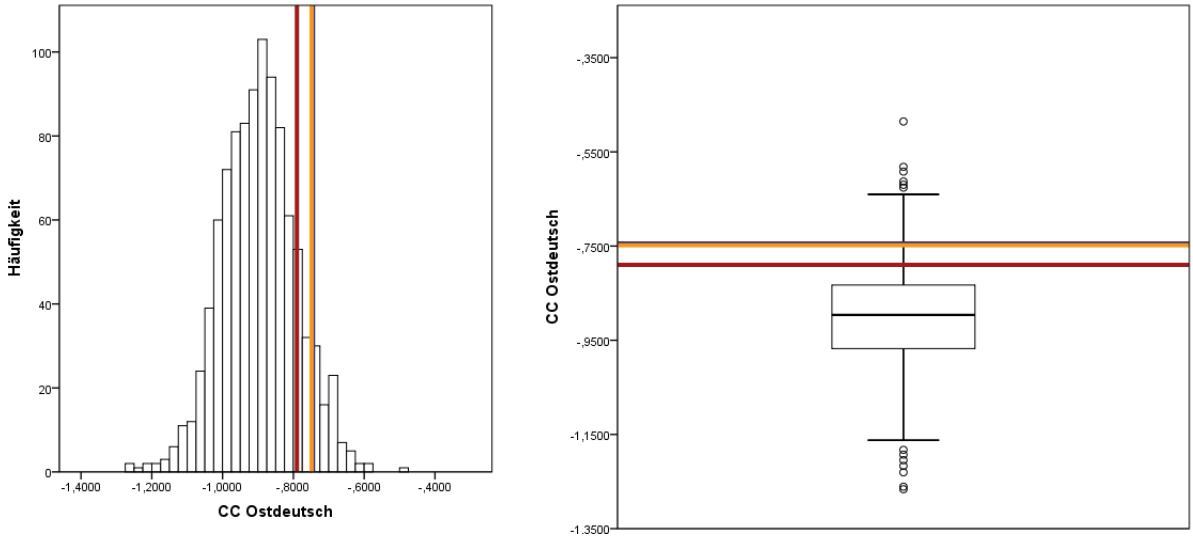



Abbildung 33: CC Histogramm und Boxplot des Parameters zur Variablen *Ostdeutsch*

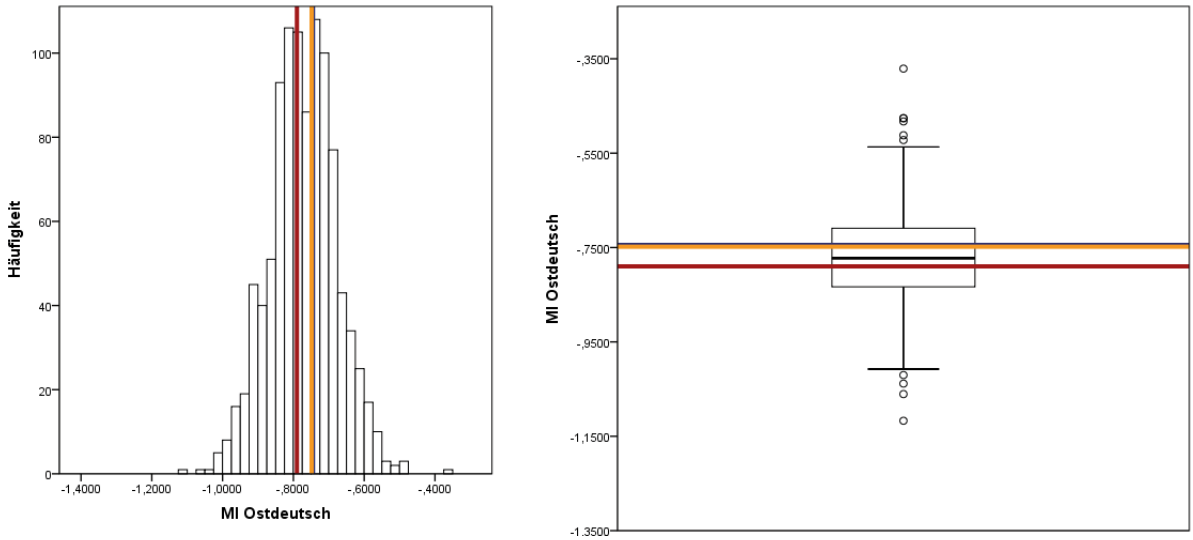



Abbildung 34: MI Histogramm und Boxplot des Parameters zur Variablen *Ostdeutsch*

Was für die Variable *jemals arbeitslos* konstatiert wurde, trifft bei der Variablen *Ostdeutsch* noch mehr zu. Die Fallreduktion bei CC führt zu einer starken Verzerrung bei Schätzung der Parameter für das Probitmodell. Wiederum zeigen die geschätzten Werte eine Überschätzung der „wahren" Werte (CC: -0,8975), während MI stabil über alle Typen schätzt. Die Gestalt der Schätzwerteverteilung für CC – dies lässt sich gut im Histogramm erkennen – hat einen verhältnismäßig breiten linken Rand, der die Wahrscheinlichkeit einer starken Überschätzung steigen lässt (CC-Standardabweichung: 0,1056; MI-Standardabweichung: 0,0947).
Die letzte visuell aufbereitete Variable ist *politische Orientierung* (Abbildung 35 und 36). Anders als die vorhergehenden unabhängigen Variablen, die dichotom oder dichotomisiert in das Modell eingebracht wurden, besitzt diese insgesamt 11 Ausprägungen (nur die stetigen Variable *Alter* und $Alter^2$ besitzen mehr).

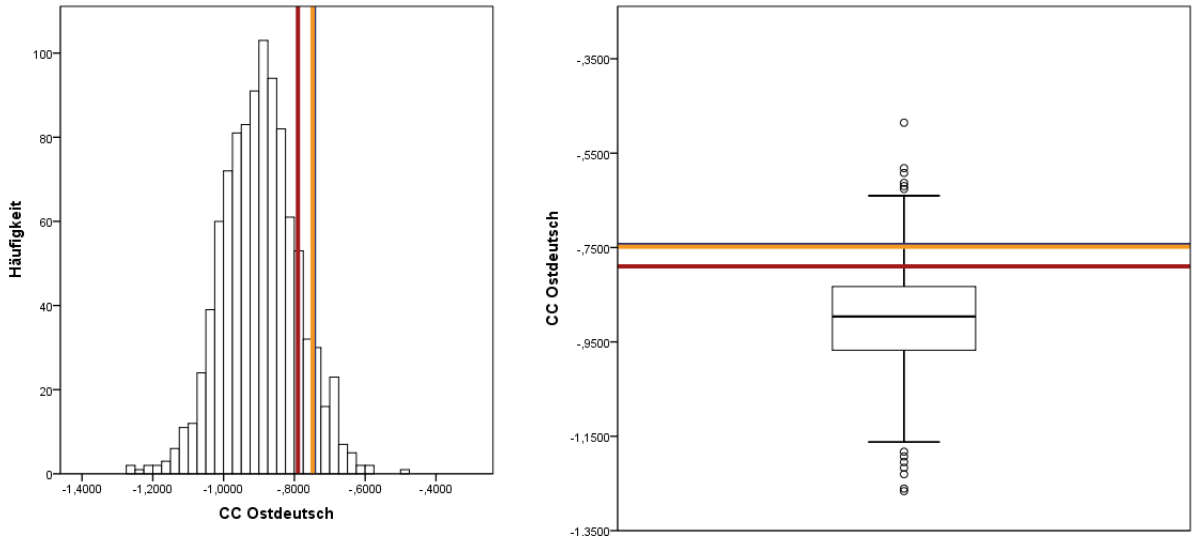



Abbildung 35: CC Histogramm und Boxplot des Parameters zur Variablen *Politische Orientierung*

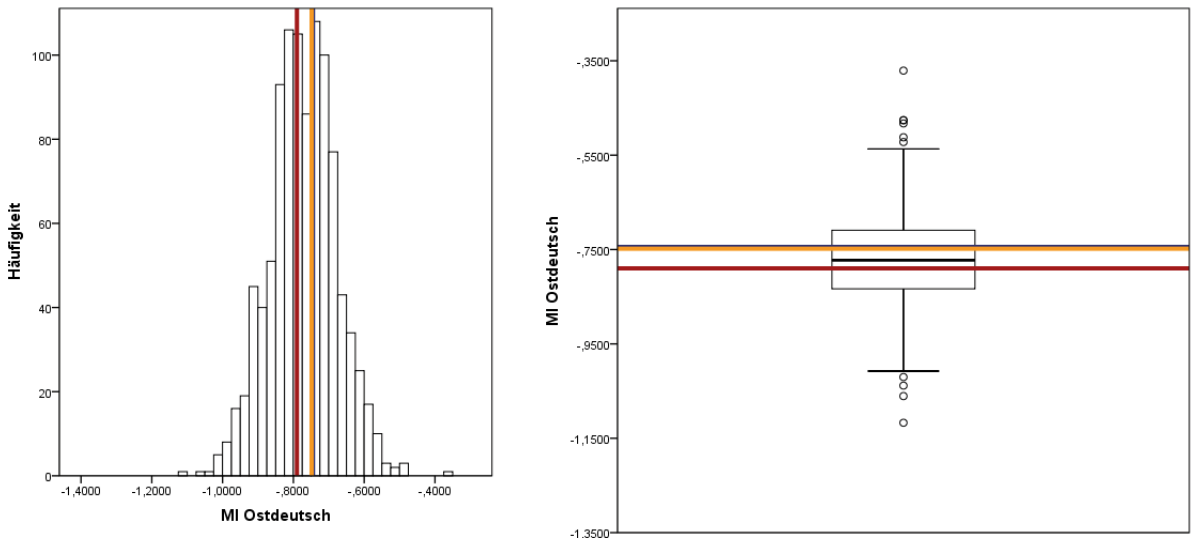



Abbildung 36: MI Histogramm und Boxplot des Parameters zur Variablen *Politische Orientierung*

Auch bei dieser Variablen unterscheiden sich CC und MI in der Gestalt der Schätzwerteverteilung. Für Typ 1 und 2 würden beide Methoden den Einfluss überschätzen, bei Typ 3 treffen beide Verfahren den „wahren“ Wert im Mittel sehr gut (CC: 0,1143; MI: 0,1128), wobei die MI-Schätzwerte auch hier weniger streuen (CC: 0,0279; MI: 0,0243).
Bevor für dieses Modell das Fazit gezogen wird, soll nach der Visualisierung einzelner Variablen, die Abweichung und deren Standardabweichung für alle drei Typen berechnet werden (Tabelle 21, 22 und 23).

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| niem. Gew.-Mitgl. | 0,1994 | 0,81140 | 0,0608 | 0,70321 |
| Geschlecht | −0,0617 | 0,11440 | 0,0268 | 0,09307 |
| Alter | 0,0054 | 0,03872 | 0,0103 | 0,03305 |
| $\text{Alter}^2$ | 0,000022 | 0,00046 | −0,0001 | 0,00039 |
| jem. arbeitslos | 0,0973 | 0,11907 | 0,0021 | 0,09960 |
| Bild.: niedrig | −0,0535 | 0,13692 | 0,0161 | 0,11534 |
| Uni.-Abschluss | −0,0664 | 0,12558 | −0,0307 | 0,11004 |
| Bild. Vt.: niedrig | −0,1645 | 0,13246 | −0,1561 | 0,12099 |
| Bild. M.: niedrig | −0,1429 | 0,13764 | −0,0828 | 0,11999 |
| Vt. Arbeiter | 0,0901 | 0,11366 | 0,0934 | 0,10627 |
| Arbeiter | 0,0645 | 0,12188 | 0,0013 | 0,10403 |
| Ostdeutsch | 0,1076 | 0,10560 | −0,0187 | 0,09470 |
| pol. Orientierung | −0,0383 | 0,02787 | −0,0368 | 0,02431 |
| Vollzeit | −0,0450 | 0,14591 | −0,0201 | 0,12937 |
| Ö. Dienst | −0,0372 | 0,11302 | 0,0263 | 0,09773 |

Tabelle 21: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| niem. Gew.-Mitgl. | 0,0330 | 0,78578 | −0,0774 | 0,65732 |
| Geschlecht | −0,0529 | 0,11410 | 0,0408 | 0,09785 |
| Alter | 0,0003 | 0,03861 | 0,0043 | 0,03195 |
| $\text{Alter}^2$ | 0,000082 | 0,00047 | 0,000030 | 0,00038 |
| jem. arbeitslos | 0,1074 | 0,24441 | 0,0085 | 0,09897 |
| Bild.: niedrig | −0,0741 | 0,13158 | −0,0051 | 0,11033 |
| Uni.-Abschluss | −0,0640 | 0,13541 | −0,0518 | 0,12371 |
| Bild. Vt.: niedrig | −0,0619 | 0,13541 | −0,0518 | 0,12371 |
| Bild. M.: niedrig | −0,0209 | 0,13553 | 0,0470 | 0,11949 |
| Vt. Arbeiter | 0,0288 | 0,11701 | 0,0319 | 0,11217 |
| Arbeiter | 0,0738 | 0,12821 | 0,0037 | 0,10772 |
| Ostdeutsch | 0,1463 | 0,10846 | 0,0163 | 0,09402 |
| pol. Orientierung | −0,0219 | 0,02753 | −0,0214 | 0,02416 |
| Vollzeit | −0,0362 | 0,15239 | −0,0117 | 0,13759 |
| Ö. Dienst | −0,0652 | 0,11433 | −0,0010 | 0,09838 |

Tabelle 22: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| niem. Gew.-Mitgl. | 0,0146 | 0,78116 | −0,1018 | 0,66780 |
| Geschlecht | −0,0790 | 0,11172 | 0,0111 | 0,09675 |
| Alter | −0,0016 | 0,03818 | 0,0027 | 0,03227 |
| Alter$^2$ | 0,000085 | 0,00046 | 0,000030 | 0,00039 |
| jem. arbeitslos | 0,0885 | 0,12126 | −0,0074 | 0,10121 |
| Bild.: niedrig | −0,0830 | 0,13603 | −0,0169 | 0,10945 |
| Uni.-Abschluss | −0,0420 | 0,12638 | −0,0087 | 0,11208 |
| Bild. Vt.: niedrig | −0,0745 | 0,13153 | −0,0612 | 0,12329 |
| Bild. M.: niedrig | −0,0168 | 0,12831 | 0,0498 | 0,11672 |
| Vt. Arbeiter | 0,0526 | 0,11293 | 0,0560 | 0,10697 |
| Arbeiter | 0,0475 | 0,12662 | −0,0199 | 0,10697 |
| Ostdeutsch | 0,1446 | 0,11156 | 0,0185 | 0,09662 |
| pol. Orientierung | 0,0010 | 0,02608 | 0,0018 | 0,02334 |
| Vollzeit | −0,2060 | 0,14942 | 0,0001 | 0,13010 |
| Ö. Dienst | −0,0570 | 0,11366 | 0,0082 | 0,09397 |

Tabelle 23: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3 und Standardabweichung der Abweichung bei CC und MI

Der Gesamteindruck der Coverageergebnisse bestätigt bzw. verstärkt sich bei der grafischen Darstellung und Berechnung der Abweichung der Schätzwerte von den „wahren“ Werten. Es gibt einige Variablen, bei denen der Bias bei MI größer ist als bei CC. Dies kommt bei Typ 1 drei-, bei Typ 2 und 3 fünfmal vor. Für diese Fälle soll noch der MSE berechnet werden:

| Variable (Typ) | MSE CC | MSE MI |
|---|---|---|
| Alter (Typ 1) | 0,00151 | 0,00120 |
| Alter$^2$ (Typ 1) | 0,000000213 | 0,000000161 |
| Vt. Arbeiter (Typ 1) | 0,0211 | 0,0197 |
| niem. Gew.-Mitgl. (Typ 2) | 0,6190 | 0,4380 |
| Alter (Typ 2) | 0,00149 | 0,00104 |
| Bild. M.: niedrig (Typ 2) | 0,01844 | 0,00162 |
| Vt. Arbeiter (Typ 2) | 0,01483 | 0,01402 |
| niem. Gew.-Mitgl. (Typ 3) | 0,6102 | 0,4564 |
| Alter (Typ 3) | 0,00146 | 0,00104 |
| Bild. M.: niedrig (Typ 3) | 0,0152 | 0,0065 |
| Vt. Arbeiter (Typ 3) | 0,01674 | 0,01610 |
| pol. Orientierung (Typ 3) | 0,000681 | 0,000548 |

Tabelle 24: MSE für ausgewählte Parameter

Von den 12 nochmals untersuchten Variablen besitzt MI stets den kleineren MSE (Tabelle 24). Der Vorteil von MI ist dabei die Effizienz der Schätzung, die die Verzerrungen mehr als aufhebt.
MI als Imputationsverfahren ist im multivariaten Raum CC als Fallreduktionsverfahren überlegen. Dies zeigt sich bei näherer Betrachtung über alle drei Typen hinweg, selbst bei denjenigen, die stark NMAR geprägt sind (Typ 1 und 2). Etwas weniger sicher schätzt MI bei größerem Ausfall und bei größer werdender Ausprägungszahl. Dagegen lässt sich CC als unzuverlässig charakterisieren. Durch die Fallreduktion werden auch Parameter, die nicht von Ausfall betroffen sind, teilweise stark verzerrt. Analysiert man noch die Verteilung der Schätzwerte, fallen bei der Verteilung der CC-Schätzwerte durchgehend die breiteren Ränder auf. Nach Auswertung der Ergebnisse muss MI deshalb für multivariate Modelle dieser Art empfohlen werden, selbst wenn NMAR vorliegt.

#### 4.3.3.2 Ergebnis 3: Parameter des Logitmodells und individuelle Eintrittswahrscheinlichkeiten

In der dritten Veröffentlichung sind beide Korrekturverfahren mit zwei Herausforderungen konfrontiert: Zum einen gibt es wesentlich weniger dichotome oder dichotomisierte Variablen. Das Basismodell besteht unter anderem aus zwei stetigen Variablen (*Einkommen* und *Alter*) und aus mehreren Merkmalen, die in der sozialwissenschaftlichen Analysepraxis als „quasimetrisch" bezeichnet werden: *Politisches Interesse*, *Demokratiezufriedenheit* und der *Partizipationsindex*. Wie bereits besprochen fallen zum anderen die Ausfälle der einzelnen unabhängigen Variablen sehr unterschiedlich aus. Für das Logitmodell, das die Grundlage für die später berechneten Eintrittswahrscheinlichkeiten bei zunehmender Demokratiezufriedenheit darstellt, bedeutet dies, dass fast die Hälfte der Merkmalsträger im multivariaten Raum verloren gehen.
Zunächst werden die Ergebnisse für das Logitmodell untersucht (Tabelle 25). Wiederum beginnend mit der Coverage werden dann ausgewählte Ergebnisse auch visualisiert analysiert. Aufgrund der schwierigen Varianzbestimmung für die individuellen Eintrittswahrscheinlichkeiten wird bei deren Analyse auf die Coverage verzichtet und die Punktschätzungen näher betrachtet.

| | | Typ 1 | | Typ 2 | | Typ 3 | |
|---|---|---|---|---|---|---|---|
| | Ausfall | CC | MI | CC | MI | CC | MI |
| Wählen | 8,6 | $96{,}2^{***}$ | $81{,}5^{***}$ | $90{,}5^{***}$ | $53{,}5^{***}$ | $96{,}1^{ns}$ | $95{,}7^{ns}$ |
| Ost | 0,0 | $38{,}3^{***}$ | $78{,}1^{***}$ | $95{,}5^{***}$ | $46{,}7^{***}$ | $91{,}6^{***}$ | $72{,}3^{***}$ |
| Weiblich | 0,0 | $92{,}2^{**}$ | $95{,}0^{**}$ | $91{,}9^{***}$ | $95{,}5^{***}$ | $88{,}4^{***}$ | $94{,}4^{***}$ |
| Alter | 0,1 | $93{,}5^{***}$ | $65{,}0^{***}$ | $55{,}0^{***}$ | $84{,}2^{***}$ | $50{,}0^{***}$ | $81{,}0^{***}$ |
| Bild. mittel | 1,3 | $86{,}3^{**}$ | $84{,}1^{**}$ | $93{,}0^{ns}$ | $93{,}6^{ns}$ | $92{,}8^{ns}$ | $92{,}8^{ns}$ |
| Bild. niedrig | 1,3 | $91{,}6^{*}$ | $94{,}1^{*}$ | $91{,}2^{ns}$ | $92{,}5^{ns}$ | $92{,}7^{ns}$ | $94{,}2^{ns}$ |
| Einkommen | 44,5 | $78{,}2^{***}$ | $93{,}3^{***}$ | $32{,}1^{***}$ | $56{,}0^{***}$ | $31{,}1^{***}$ | $55{,}4^{***}$ |
| Demokratiezuf. | 5,4 | $52{,}2^{***}$ | $67{,}8^{***}$ | $35{,}3^{*}$ | $38{,}4^{*}$ | $90{,}5^{**}$ | $94{,}3^{**}$ |
| Pol. Interesse | 0,0 | $62{,}5^{***}$ | $43{,}0^{***}$ | $94{,}0^{ns}$ | $96{,}1^{ns}$ | $95{,}1^{*}$ | $96{,}1^{*}$ |
| Partizipation | 0,0 | $91{,}9^{***}$ | $87{,}7^{***}$ | $95{,}5^{*}$ | $97{,}2^{*}$ | $95{,}6^{ns}$ | $95{,}8^{ns}$ |
| | 48,2 | 78,3 | 77,1 | 77,4 | 75,4 | 82,4 | 87,2 |

Tabelle 25: Vergleich der Korrekturmethoden anhand der Coverage bei Beispiel 3

Im Durchschnitt weisen beide Korrekturverfahren teilweise schlechte Werte über die drei Typen auf. Wie bei den Analysen zuvor bilden Typ 1 und Typ 2 eher die Extremtypen, Typ 3 einen gemäßigten Typ.
Insgesamt gibt es bei sieben der 30 Coverages keine signifikanten Unterschiede (ca. 23 %); bei acht Coverages schneidet CC signifikant besser ab als MI (27 %) und in 15 Fällen MI besser (50 %). Im Vergleich zu den Schätzungen im letzten Kapitel fallen zwei Dinge auf: die Anzahl der

Nichtsignifikanzen ist deutlich geringer (nur ca. 17 %) und MI schneidet auf den ersten Blick besser ab als CC, für die jeweiligen Parameter sind die Unterschiede aber wesentlich größer. Global gibt es Unterschiede zwischen den Korrekturmethoden bezogen auf die Typen: fünf zu zwei zu eins ist CC und fünf zu fünf zu fünf ist MI jeweils signifikant besser. Wie eingangs angedeutet zeigen beide Verfahren Schwächen. Das Coverage-Niveau für den Typ 1 liegt in fünf von zehn Fällen unter 90 % bei CC, bei MI sind es sieben von zehn – vor allem für die Variable *politisches Interesse* und *Alter* sind die Ergebnisse recht ungünstig. Die bestmögliche Coverage von etwa 95 % erreicht CC bei Typ 1 nur in einem, MI in zwei Fällen. Bei Typ 2 ist das Verhältnis drei (CC) zu fünf (MI) Coverages, die unter einem Niveau von 90 % bleiben. Bei Typ 3 fällt das Verhältnis 3 (CC) zu 3 (MI) aus. Typ 2 weist für CC dreimal, für MI ebenfalls dreimal die Spitzencoverage auf. Typ 3 zeigt dieselben Werte bei der Häufigkeit der ca. 95 %-Coverage: drei (CC) zu drei (MI).
Betrachtet man nun die einzelnen Variablen, so ergibt sich kaum ein Muster wie bei der Analyse des Probitmodells aus Abschnitt 4.3.3.2. Die beiden dichotomisierten Variablen zum Bildungsniveau fallen unspektakulär für Typ 2 und 3 aus; gleiches gilt für die Variable *Weiblich* und teilweise für *Partizipation*. Erstaunlich erscheint jedoch das schlechte Abschneiden von MI bei der Schätzung der Achsenabschnittswerte (*Wahlteilnahme*) mit Ausnahme des gemäßigten Typs 3. Hier ist eine weitere Analyse geboten. Nicht weniger überraschend kommen die Ergebnisse für die Variablen *Ostdeutschland* und *Alter*, die eigentlich kaum oder keine Ausfälle aufweisen. MI schätzt hier über alle Typen hinweg sehr unbefriedigend, wenn man nur nach der Coverage ginge; CC verhält sich unsystematisch. Aufgrund des massiven Ausfalls bei der Variable *Einkommen* ist das Abschneiden der Korrekurmethoden von besonderem Interesse. MI liefert zwar stets bessere Ergebnisse, aber auf unterschiedlichem Niveau. Zunächst verblüffend ist der Umstand, dass bei Typ 1 und 2 gänzlich verschiedene Coverage erreicht werden, der gemäßigte Typ 3 allerdings fast dieselben Ergebnisse wie Typ 2 zeigt. Dies gilt auch für CC, jedoch auf noch einmal niedrigerem Niveau. Einem sprunghaften Auf und Ab gleichen die Coverages der Variable *Demokratiezufriedenheit*, wobei MI auf verschiedenen Niveaus über die drei Typen höhere Coverages aufweist. Anders als im letzten Abschnitt variiert bei einer Variable wie *politisches Interesse*, die selbst keinen Ausfall zeigt, die Coverage beträchtlich. Zudem weist MI nicht automatisch die besseren Ergebnisse auf.
An dieser Stelle sei daran erinnert, dass nach der Berechnung von Wagner (2007) eine spezielle Eintrittswahrscheinlichkeit für die Wahlteilnahme prognostiziert werden soll, die sich auf den Achsenabschnitt, die mittlere Bildungsvariable und die *Demokratiezufriedenheit* stützt, während alle andere Koeffizienten gleich null gesetzt werden.
Bereits in der Analyse zum Probitmodell von Schnabel und Wagner war ein Misstrauen gegenüber der Coverage als einzigem Kriterium für die Leistungsfähigkeit von Korrekturverfahren angebracht. Deshalb zunächst wieder der Blick auf die Länge der Konfidenzintervalle (Abbildung 37).

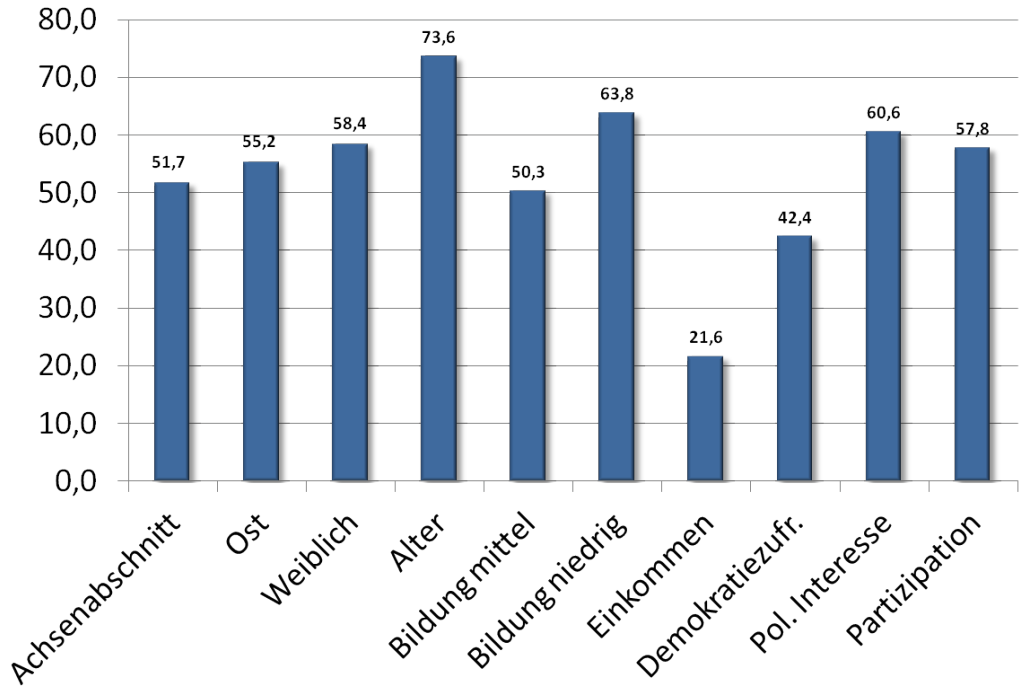



Abbildung 37: Differenz der Konfidenzintervalllängen bei Beispiel 3

Die hier für alle Variablen abgetragenen Differenzen zwischen der Länge der CC-Konfidenzintervalle und der MI-Konfidenzintervalle offenbaren weitaus größere Unterschiede als beim Probitmodell. Mit wenigen Ausnahmen sind alle CC-Intervalle durchschnittlich um 50 % bis 75 % länger als die entsprechenden MI-Intervalle. Die Ausnahmen bilden zwei Variablen mit Ausfall verschiedenen Umfangs: *Demokratiezufriedenheit* (5,4 %) und *Einkommen* (44,5 %). Aus den Ausführungen zur Theorie der MI-Stichprobenvarianz ist es damit konsequent, dass die Variable *Einkommen* die geringste Abweichung zwischen den CC- und MI-Konfidenzintervallen zeigt: Die Länge der Intervalle hat direkten Einfluss auf die Coverage-Ergebnisse.
Im Folgenden werden ausgewählte Ergebnisse noch einmal visualisiert und genauer analysiert. Wie auch schon in den Abschnitten 4.3.3.1 und 4.3.3.2 sind die Werte für die Punktschätzung der Koeffizienten abgetragen und zwar einmal in einem Histogramm sowie in einem Boxplot für MI und CC; darüber hinaus werden für jeden Koeffizienten dieselben Achsenbemessungen beibehalten. Dabei sind wiederum die „wahren“ Werte der drei Typen eingezeichnet: Typ 1 (rot), Typ 2 (blau) und Typ 3 (orange). Die Stresstestkonzeption hat es bei diesem Beispiel mit sich gebracht, dass in manchen Fällen die „wahren“ Werte dicht beieinander liegen. Dies hat sich bereits in der Coverage-Übersicht angedeutet.
Die erste Visualisierung befasst sich mit dem Achsenabschnitt, hinter dem die abhängige Variable *Wahlteilnahme bei der letzten Bundestagswahl* steht.[92]

[92]Hier nicht aufgeführte grafische Darstellungen finden sich in Anhang 7.

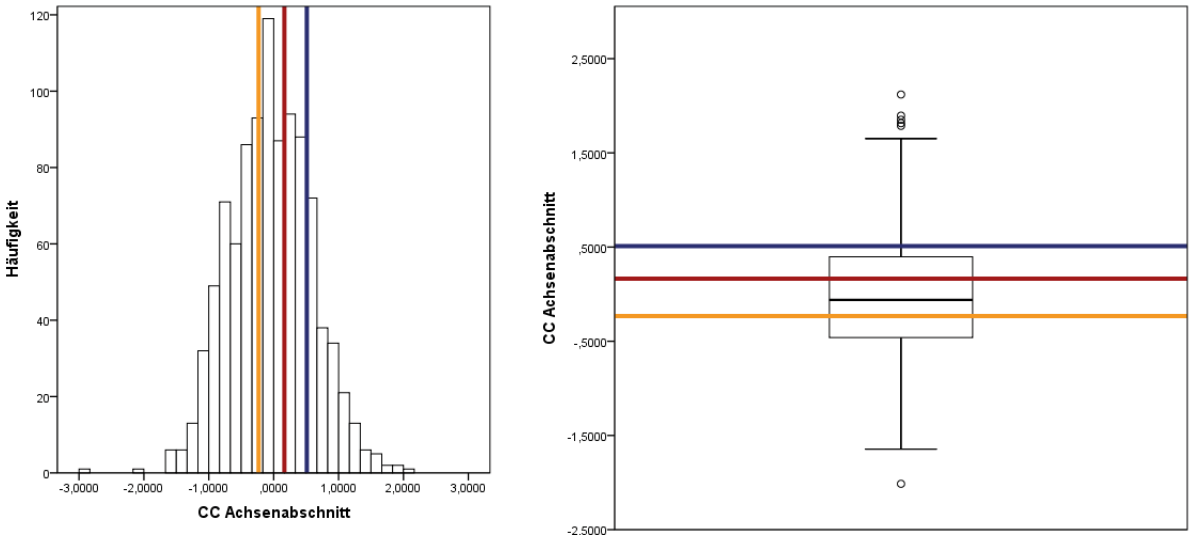



Abbildung 38: CC Histogramm und Boxplot des Achsenabschnitts

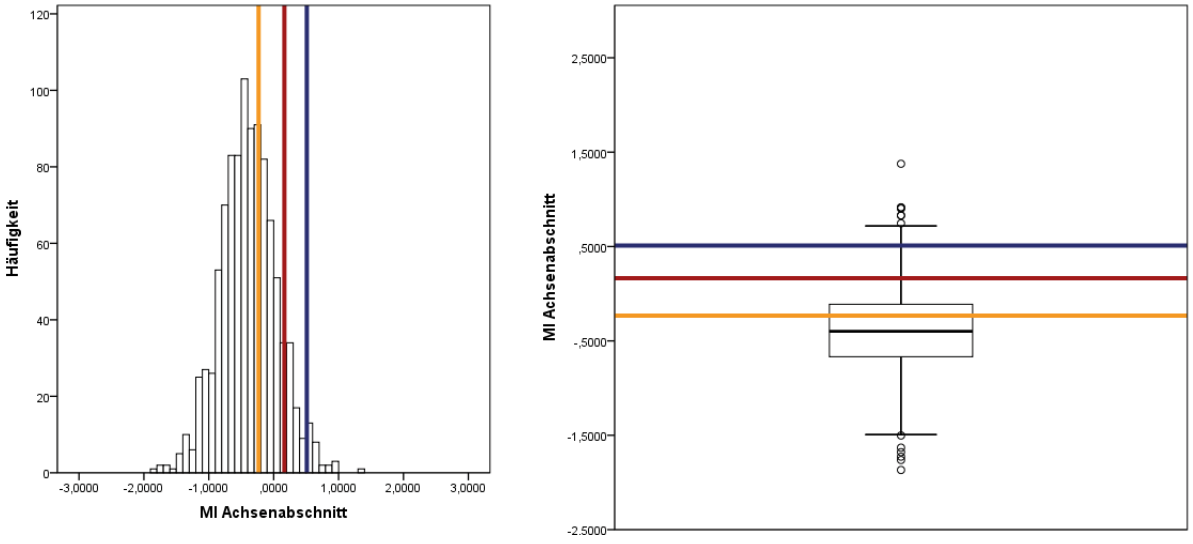



Abbildung 39: MI Histogramm und Boxplot des Achsenabschnitts

Bei der Schätzung des Achsenabschnittparameters zeigen sich bereits erhebliche Unterschiede in der Verteilung von MI- und CC-Schätzwerten (Abbildung 38 und 39). Durchschnittlich wird bei MI der Achsenabschnitt etwas näher an null geschätzt als bei CC (0,0387 bei MI und -0,0421 bei CC). Auffällig ist jedoch die breite Gestalt der CC-Verteilung im Histogramm, die sich in der Standardabweichung widerspiegelt: 0,6318 beträgt die Standardabweichung bei CC, wohingegen die MI-Werte nur mit 0,4358 streuen. Setzt man die Verteilungen in Bezug zu den „wahren“ Werten, ergibt sich ein Bild, das nicht eindeutig ist. Der Betrachter mag zunächst zum Schluss kommen, MI schätze effektiver aber nicht erwartungstreu, während die Masse der CC-Verteilung aufgrund ihrer Breite die „wahren“ Werte mehr oder weniger überdecken kann. Letztlich führt nur die Messung des Bias zur klaren Beantwortung der Frage nach dem für das Beispiel überlegenen Korrekturmethode: Typ 1 CC = 0,2061; MI = 0,5022, Typ 2 CC = 0,5579; MI = 0,8741, Typ 2 CC = -0,1946; MI = 0,1417. Die Zahlen lassen sich für die endgültige Interpretation der Ergebnisse des Achsenabschnitts folgendermaßen bewerten: MI liefert effizientere Schätzwerte, die aber im Falle des Achsenabschnitts verzerrt sind (zumindest für die Extremtypen 1 und 2). Ähnliche Ergebnisse, wenn auch in abgeschwächter Form, ließen sich auch für das Probitmodell in Abschnitt 4.3.3.2 beobachten.
Der nächste Koeffizient befasst sich mit dem Einfluss der Variable *Alter* auf die Wahlteilnahme. Die Coverage-Ergebnisse zeigten hier sehr unregelmäßige Ergebnisse.

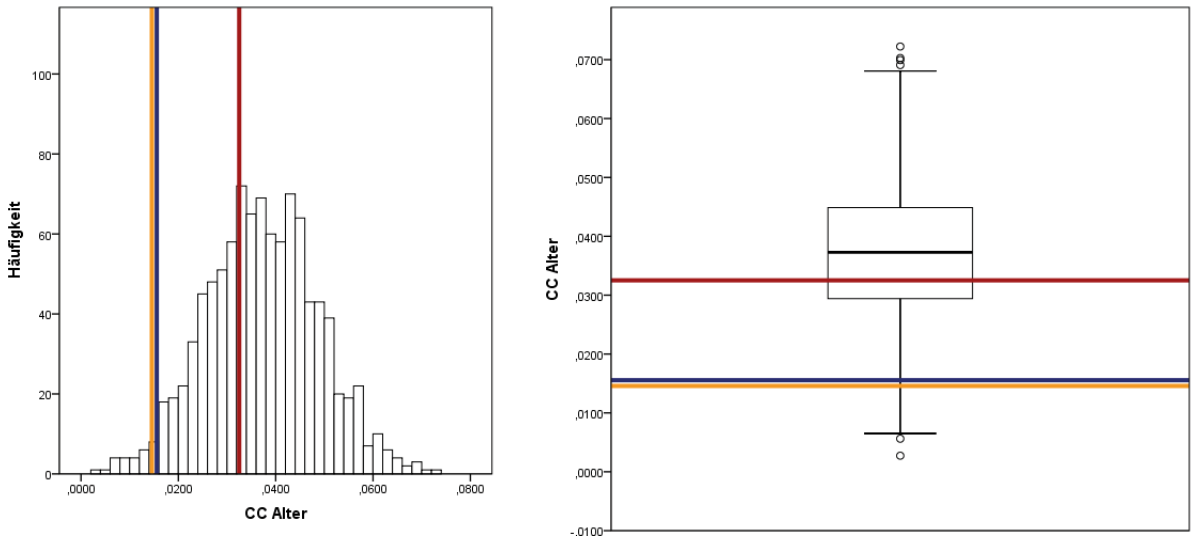



Abbildung 40: CC Histogramm und Boxplot des Parameters zur Variablen *Alter*

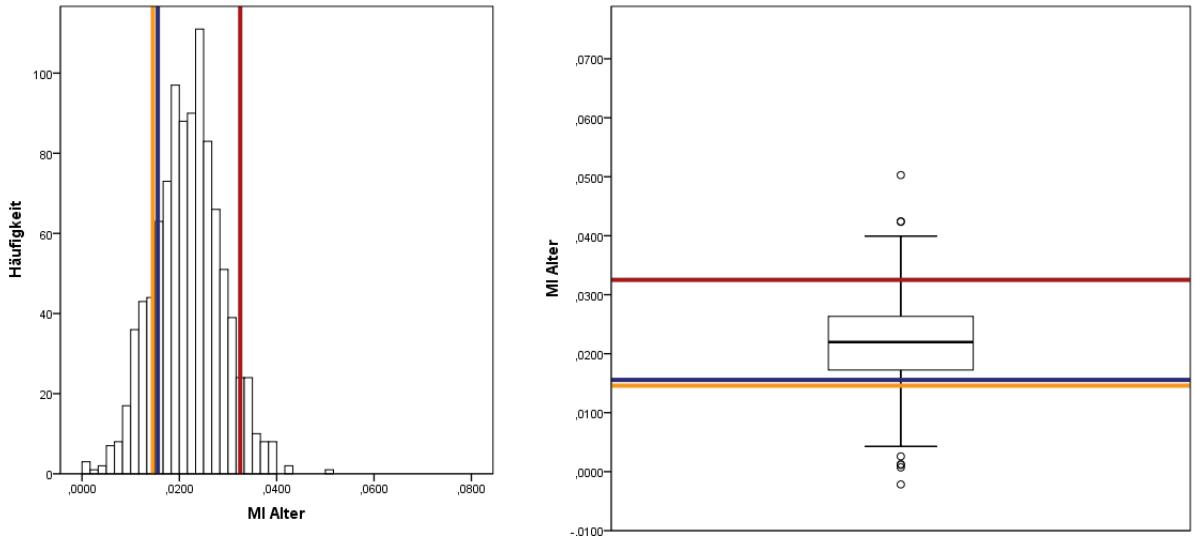



Abbildung 41: MI Histogramm und Boxplot des Parameters zur Variablen *Alter*

Da die Achsen exakt dieselbe Skalierung aufweisen, wird bei der Variablen *Alter* besonders deutlich, wie stark die Verteilungen der geschätzten Werte unterschiedlich bei den beiden Verfahren streuen (Abbildung 40 und 41). Während CC eine Standardabweichung von 0,0115 misst, liegt diese bei MI um das 1,7-fache niedriger. Die eingezeichneten „wahren“ Werte führen vor dem Hintergrund dieser Verteilung zur Annahme, dass die Coverage ein unzureichendes Bild gibt. Die MI-Schätzwerte liegen optisch etwa in der Mitte der „wahren“ Werte. Der Bias beträgt hier: Typ 1 CC = 0,0373; MI = 0,0107, Typ 2 CC = -0,0211; MI = -0,0064, Typ 3 CC = -0,0211; MI = -0,0073. Diese Zahlen bergen eine Überraschung, weil die Coverage für MI bei Typ 2 und 3 tatsächlich wesentlich besser sind, jedoch durchschnittlich die MI-Schätzwerte bei Typ 3 um einiges weniger vom „wahren“ Wert abweichen als die CC-Schätzwerte, die eigentlich eine höhere Coverages besitzen. MI schätzt aber für diesen Koeffizienten unverzerrter und präziser als CC.
Die nächsten visualisierten Schätzwerte zeigen die Ergebnisse für den Einfluss des Einkommens. Die Variable *Einkommen* ist stetig und leidet mit fast 45 % unter massivem Datenausfall.

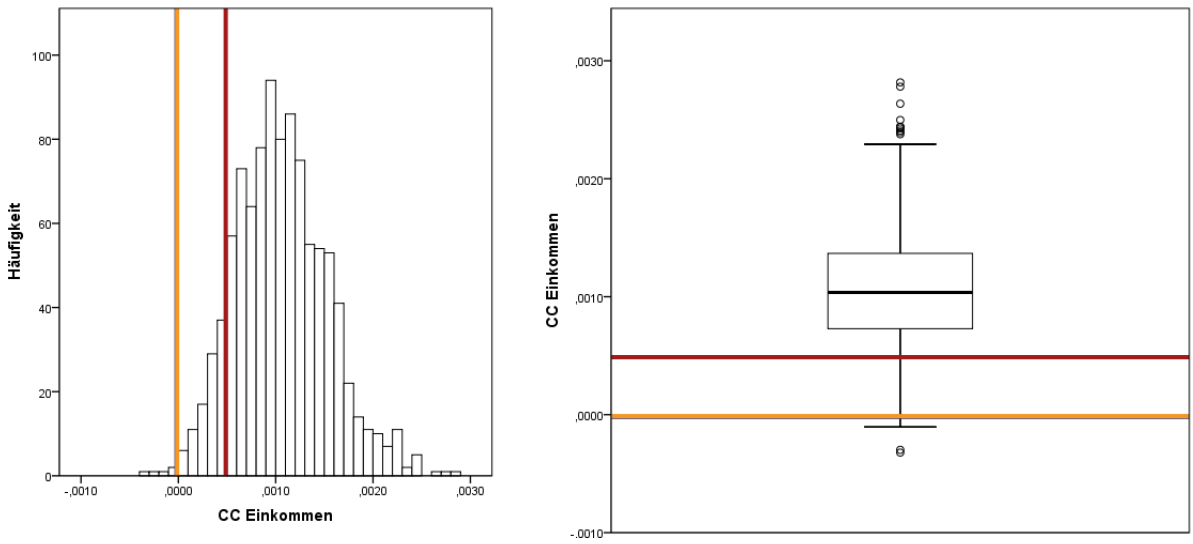



Abbildung 42: CC Histogramm und Boxplot des Parameters zur Variablen *Einkommen*

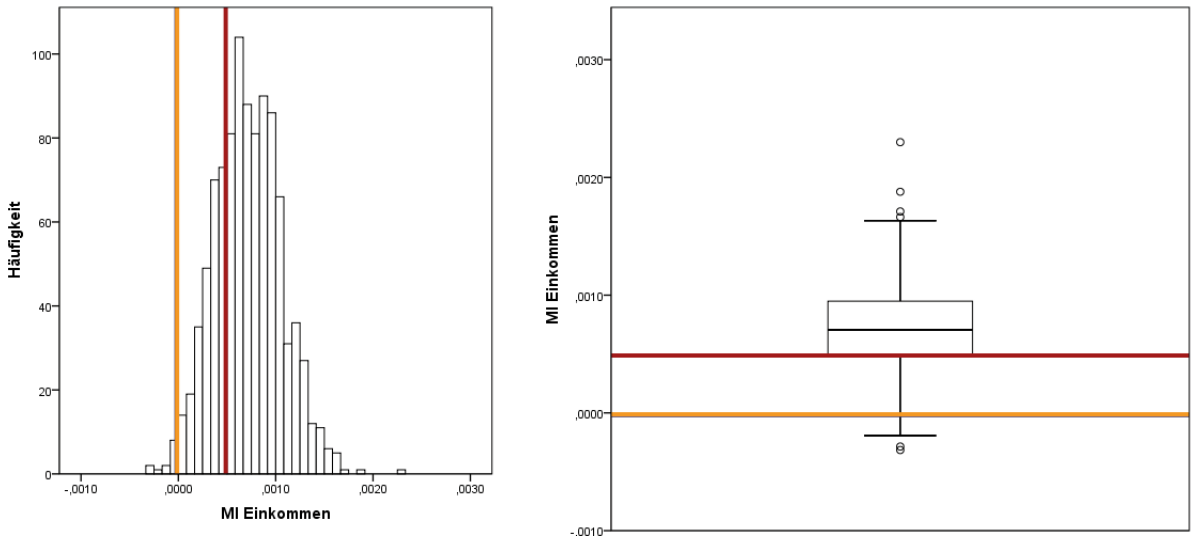



Abbildung 43: MI Histogramm und Boxplot des Parameters zur Variablen *Einkommen*

Die Histogramme und Boxplots zeigen wieder eine deutlich breitere Streuung der Schätzwerte von CC als die Streuung der MI-Schätzwerte (Standardabweichung: CC = 0,00048; MI = 0,0740) (Abbildung 42 und 43). Beide Verfahren überschätzen jedoch über alle Typen hinweg die Höhe des Koeffizienten und damit den Einfluss der Variablen auf die Wahlteilnahme. Mit Blick auf die durchschnittlichen Abweichungen von den „wahren“ Werten schneidet MI aber insgesamt viel besser ab, als es die Coverage-Ergebnisse nahe legen.

Zuletzt werden noch die Ergebnisse für die Variable *Demokratiezufriedenheit* visualisiert. Zwar zeigt MI über alle Typen stets die höhere Coverage – jedoch bei Typ 1, mehr noch bei Typ 2, auf einem niedrigen Niveau.

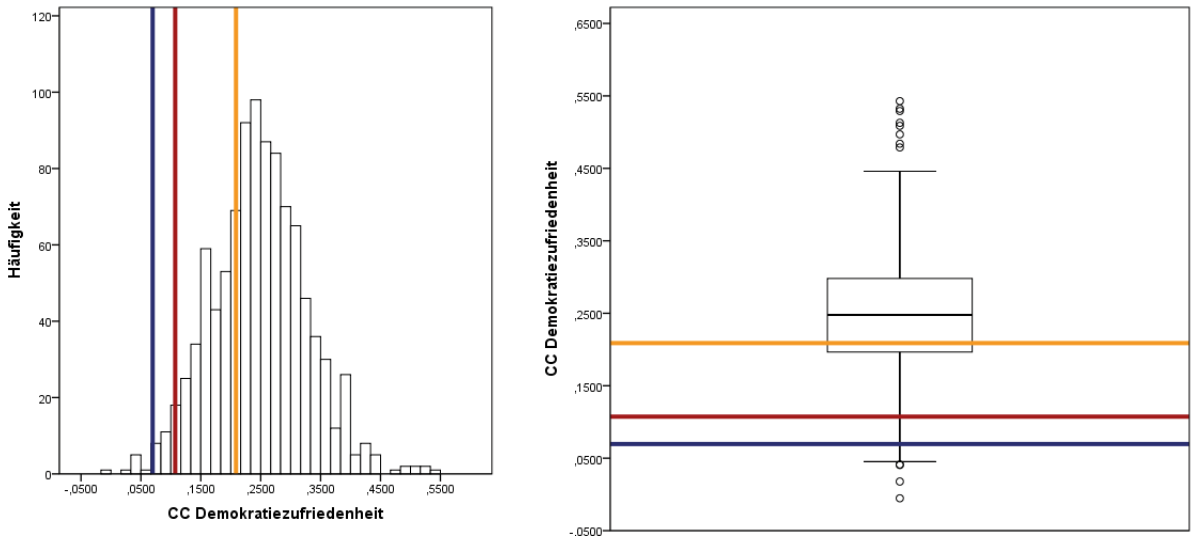



Abbildung 44: CC Histogramm und Boxplot des Parameters zur Variablen *Demokratiezufriedenheit*

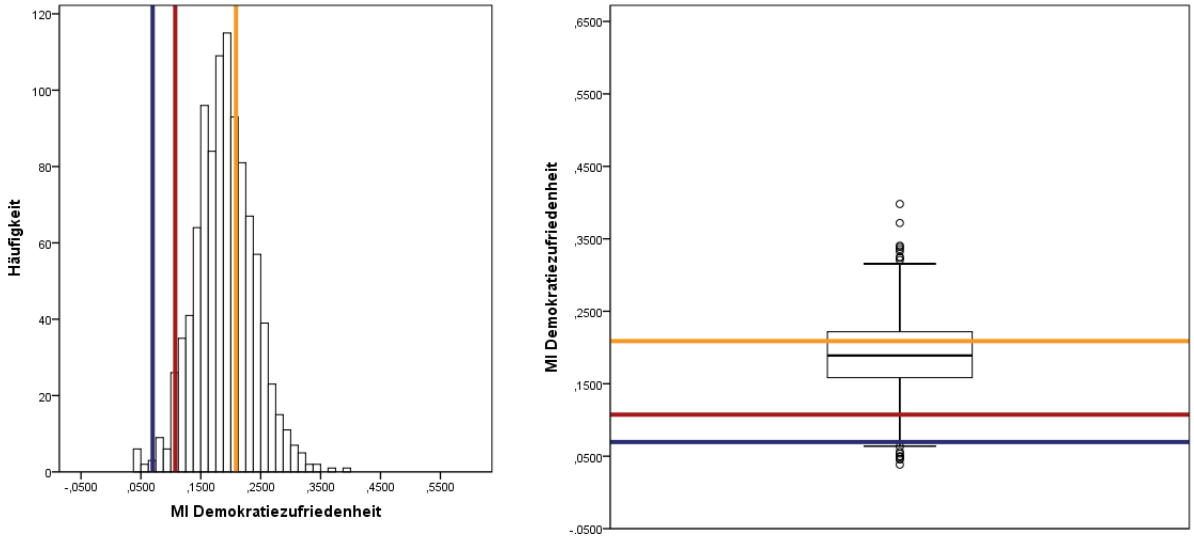



Abbildung 45: MI Histogramm und Boxplot des Parameters zur Variablen *Demokratiezufriedenheit*

Sowohl in den Histogrammen als auch im Boxplot ist gut zu erkennen, wie sich die beiden Verteilungen schon in der Streuung unterscheiden (CC = 0,07892; MI = 0,04915) (Abbildung 44 und 45). Während CC die „wahren“ Werte aller drei Typen recht stark überschätzt, überschätzt zwar auch MI den Einfluss der *Demokratiezufriedenheit* auf die abhängige Variable bei Typ 1 und 2. Für den gemäßigten Typ 3 liefert MI jedoch durchaus gute Schätzwerte. Bezieht man die durchschnittliche Abweichung noch in die Analyse mit ein, dann zeigt sich, wie stark die Coverage die Schwäche von CC verdeckt. Durchschnittlich weicht für Typ 1 CC um fast 70 % mehr vom „wahren“ Wert ab als MI, für Typ 2 um immerhin etwa 45 % und für Typ 3 sogar um 145 % (Tabelle 26, 27 und 28). Aufgrund der verhältnismäßig großen Intervalle kommt viel schlechter zur Geltung, um wie viel größer die Verzerrung und – wie bei den Variablen zuvor auch – die Ungenauigkeit von CC ist.

Dieses Phänomen lässt sich bei allen drei Typen sehr häufig finden: MI schätzt insgesamt wesentlich genauer und mit weniger Abweichung von den „wahren“ Werten. Im Folgenden ist die Übersicht über alle Typen für den Bias aufgeführt:

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| Wählen | 0,1942 | 0,63974 | 0,5283 | 0,42400 |
| Ost | 0,7964 | 0,36629 | 0,2834 | 0,21879 |
| Weiblich | 0,1599 | 0,32684 | −0,0698 | 0,19377 |
| Alter | −0,0048 | 0,01149 | 0,0107 | 0,00694 |
| Bild. mittel | 0,2435 | 0,39428 | 0,2286 | 0,25292 |
| Bild. niedrig | −0,4216 | 0,73542 | −0,0704 | 0,40138 |
| Einkommen | −0,0006 | 0,00048 | −0,0002 | 0,00034 |
| Demokratiezuf. | −0,1407 | 0,07892 | −0,0830 | 0,04915 |
| Pol. Interesse | −0,4448 | 0,26611 | −0,3467 | 0,16212 |
| Partizipation | 0,0691 | 0,11759 | 0,0577 | 0,07792 |

Tabelle 26: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI (logistische Regression)

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| Wählen | 0,5487 | 0,63475 | 0,8850 | 0,43623 |
| Ost | −0,0264 | 0,35581 | −0,4656 | 0,22110 |
| Weiblich | 0,1983 | 0,32566 | −0,0332 | 0,20855 |
| Alter | −0,0211 | 0,01211 | −0,0064 | 0,00702 |
| Bild. mittel | −0,0682 | 0,40542 | −0,0993 | 0,25456 |
| Bild. niedrig | −0,2541 | 0,77157 | 0,0784 | 0,41259 |
| Einkommen | 0,0011 | 0,00046 | −0,0007 | 0,00034 |
| Demokratiezuf. | −0,1766 | 0,08382 | −0,1215 | 0,05379 |
| Pol. Interesse | −0,0804 | 0,25935 | 0,0105 | 0,15623 |
| Partizipation | 0,0005 | 0,11526 | −0,0088 | 0,07245 |

Tabelle 27: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI (logistische Regression)

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| Wählen | −0,1946 | 0,63475 | 0,1417 | 0,43623 |
| Ost | 0,1782 | 0,37680 | −0,3206 | 0,22276 |
| Weiblich | 0,2539 | 0,32065 | 0,0332 | 0,20933 |
| Alter | −0,0211 | 0,01211 | −0,0064 | 0,00702 |
| Bild. mittel | −0,0949 | 0,40070 | −0,1149 | 0,24621 |
| Bild. niedrig | −0,3113 | 0,77252 | 0,0415 | 0,40382 |
| Einkommen | −0,0011 | 0,00046 | −0,0007 | 0,00034 |
| Demokratiezuf. | −0,0427 | 0,08215 | 0,01663 | 0,05186 |
| Pol. Interesse | −0,1126 | 0,25604 | −0,0106 | 0,15487 |
| Partizipation | −0,0137 | 0,12422 | −0,0226 | 0,07715 |

Tabelle 28: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3 und Standardabweichung der Abweichung bei CC und MI (logistische Regression)

Es gibt vom vorher Konstatierten insgesamt acht Ausnahmen, die sich vor allem auf Typ 2 konzentrieren. Für diese Variablen soll außerdem noch der mittlere quadratische Fehler ermittelt werden:

| Variable (Typ) | MSE CC | MSE MI |
|---|---|---|
| Wählen (Typ 1) | 0,4467 | 0,4671 |
| Alter (Typ 1) | 0,00017 | 0,00016 |
| Wählen (Typ 2) | 0,7251 | 0,9773 |
| Ost (Typ 2) | 0,1267 | 0,2658 |
| Bild. mittel (Typ 2) | 0,1687 | 0,0749 |
| Partizipation (Typ 2) | 0,0130 | 0,0051 |
| Ost (Typ 3) | 0,1737 | 0,1528 |
| Partizipation (Typ3) | 0,0152 | 0,0065 |

Tabelle 29: MSE für ausgewählte Parameter

Bis auf den Achsenabschnitt und die Variable *Ostdeutschland* in Typ 2 besitzt MI entweder einen sehr ähnlich hohen, meist aber einen wesentlich niedrigeren MSE (Tabelle 29.
Das hier ausführlich analysierte Logitmodell stellt eigentlich nur die Basis für die folgende Schätzung individueller Teilnahmewahrscheinlichkeiten dar. Wie in Abschnitt 4.2.1 beschrieben, werden bis auf drei alle Koeffizienten null gesetzt. Der Koeffizient für *Demokratiezufriedenheit* wird mit 0 bis 10 variierend multipliziert. Aufgrund dieses Vorgehens lässt sich nicht ohne Weiteres eine Varianz schätzen und somit auch keine Coverage errechnen.[93] Allerdings können die „wahren“ Werte (vgl. Abschnitt 4.3) mit den Schätzwerten verglichen werden, und diese sehen der Vollständigkeit halber für die drei Typen wie folgt aus:

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| 0 | 0,1085 | 0,15016 | 0,1800 | 0,10561 |
| 1 | 0,0763 | 0,13710 | 0,1573 | 0,09865 |
| 2 | 0,0455 | 0,12292 | 0,1340 | 0,09106 |
| 3 | 0,0108 | 0,10872 | 0,1109 | 0,08336 |
| 4 | −0,0064 | 0,09542 | 0,0888 | 0,07602 |
| 5 | −0,0261 | 0,08357 | 0,0683 | 0,06943 |
| 6 | −0,0414 | 0,07335 | 0,0498 | 0,06375 |
| 7 | −0,0525 | 0,06467 | 0,0338 | 0,05896 |
| 8 | −0,0600 | 0,05734 | 0,0201 | 0,05493 |
| 9 | −0,0644 | 0,05112 | 0,0089 | 0,5144 |
| 10 | −0,0665 | 0,04581 | −0,0002 | 0,04835 |

Tabelle 30: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 1 und Standardabweichung der Abweichung bei CC und MI (Eintrittswahrscheinlichkeit)

[93] Der Aufwand für eine approximative Varianzschätzung würde den Rahmen sprengen.

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| 0 | 0,1087 | 0,14687 | 0,1824 | 0,10854 |
| 1 | 0,0688 | 0,13318 | 0,1516 | 0,10051 |
| 2 | 0,0310 | 0,11876 | 0,1208 | 0,09203 |
| 3 | −0,0031 | 0,10480 | 0,0908 | 0,08366 |
| 4 | −0,0326 | 0,09223 | 0,0626 | 0,07591 |
| 5 | −0,0571 | 0,08151 | 0,0366 | 0,06914 |
| 6 | −0,0764 | 0,07263 | 0,0135 | 0,06349 |
| 7 | −0,0910 | 0,06527 | −0,0065 | 0,05887 |
| 8 | −0,1014 | 0,05910 | −0,0234 | 0,05508 |
| 9 | −0,1083 | 0,05381 | −0,0373 | 0,05189 |
| 10 | −0,1124 | 0,04919 | −0,0485 | 0,04906 |

Tabelle 31: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 2 und Standardabweichung der Abweichung bei CC und MI (Eintrittswahrscheinlichkeit)

| | CC | | MI | |
|---|---|---|---|---|
| Variable | ØAbweichung | Stdv. | ØAbweichung | Stdv. |
| 0 | 0,1110 | 0,14828 | 0,1811 | 0,10534 |
| 1 | 0,0698 | 0,13422 | 0,1499 | 0,09776 |
| 2 | 0,0309 | 0,11937 | 0,1187 | 0,8972 |
| 3 | −0,0042 | 0,10496 | 0,0886 | 0,08174 |
| 4 | −0,0345 | 0,09193 | 0,0601 | 0,07433 |
| 5 | −0,0594 | 0,08072 | 0,0347 | 0,06781 |
| 6 | −0,0791 | 0,07130 | 0,0111 | 0,06232 |
| 7 | −0,0940 | 0,06344 | −0,0089 | 0,05780 |
| 8 | −0,1045 | 0,05684 | −0,0258 | 0,05407 |
| 9 | −0,1114 | 0,05125 | −0,0396 | 0,05092 |
| 10 | −0,1154 | 0,04646 | −0,0506 | 0,04815 |

Tabelle 32: Durchschnittliche Abweichung vom „wahren“ Wert bei Typ 3 und Standardabweichung der Abweichung bei CC und MI (Eintrittswahrscheinlichkeit)

Die Verläufe der durchschnittlichen Abweichungen sind bei allen drei Typen ähnlich: Bei Typ 1 zeigt MI die geringere durchschnittliche Abweichung ab einer *Demokratiezufriedenheit* von 7, bei Typ 2 und 3 ab fünf und höher. Insgesamt besitzt die durchschnittliche Abweichung von MI zu Beginn der Skala, wenn die *Demokratiezufriedenheit* noch mit keinem oder geringerem Gewicht in die Berechnung der Teilnahmewahrscheinlichkeit einfließt, eine größere Streuung. Bei beiden Verfahren wird sie dann aber sukzessive kleiner. Nachdem die Berechnung auf den drei Koeffizienten Achsenabschnitt, *mittlere Bildung* und *Demokratiezufriedenheit* basiert, schlägt natürlich die Verzerrung unter der MI beim ersten Koeffizienten durch. Im Gegensatz zur Bildungsvariable wiesen die Coverages bei der Variablen *Demokratiezufriedenheit* keine befriedigenden Ergebnisse auf; mit Blick auf den MSE wird aber deutlich, dass bei diesen beiden Variablen MI jeweils leicht oder wesentlich bessere Werte liefert. Die Bewertung der Ergebnisse der Teilnahmewahrscheinlichkeiten bezogen auf das Problem fehlender Werte ist schwierig, da die Berechnung sehr willkürlich wirkt. Die Probleme, die dabei auftreten, haben genuin nichts mit fehlenden Werten zu tun. Deshalb soll zum Abschluss der dritten Analyse auch noch einmal festgehalten werden, dass Schätzungen nicht nur unter Nonresponse leiden, sondern auch durch die Auswahl ungeeigneter Analysemethoden eingeschränkte Aussagekraft haben können. Dennoch wurden so weit wie möglich auch diese Parameter berechnet, um die Analyse der Veröffentlichung von Schäfer (2007) entsprechend der Fragestellung in Hinblick auf die Nonresponsekorrektur zu komplettieren.

#### 4.3.3.4 Zusammenfassung

Die Ergebnisse des Methodenvergleichs zerfallen ziemlich klar in zwei Teile. Bei der Schätzung multivariater Modelle schneidet MI deutlich besser ab als CC. Dies war mit Blick auf die Coverage nicht gleich sichtbar. Zog man den MSE als Kriterium heran, schnitt MI in der Regel besser ab. Bei der Schätzung des Anteilswerts lässt sich keine eindeutige Empfehlung abgeben. Bei univariaten Schätzungen zeigte MI keine dominant besseren Ergebnisse. Bedenkt man den zusätzlichen Aufwand für eine Imputation, muss für den jeweiligen Fall entschieden werden, ob es sich lohnt, MI einzusetzen. Faktoren, auf denen eine praktikable und sinnvolle Lösung basieren sollte, sind der Ausfallumfang und das – nach plausiblen Überlegungen – Vorliegen der Bedingungen zur Durchführung einer validen Imputation (Ignorierbarkeit und MAR).[94] MI hat im Vergleich zu CC noch einen weiteren Vorteil: ursprünglich geht man für die Imputation von einem Datenerzeuger, der die Imputation durchführt, und einem Datennutzer, der mit diesen Daten Parameter schätzt, aus. Der Datenerzeuger imputiert auf Grundlage eines Imputationsmodells die fehlenden Werte und erzeugt dabei auch *m* vollständige Datensätze, die der Datennutzer als Datengrundlage für seine Schätzung verwendet (Analysemodell). In gezeigten Beispielen handelt es sich – wie aus der Beschreibung hervorgeht (Abschnitt 4.2.1, Abschnitt 4.2.2 und Abschnitt 4.2.3) – bei den Imputationsmodellen um die entsprechenden Analysemodelle zuzüglich einer Anzahl von vollständigen Variablen. Zwar gibt es neben den formalen Voraussetzungen, wie sie Rubin nennt (Rubin 1985, 1996), keine Garantie für eine gute Imputation, jedoch sollte so viel Information wie möglich in Form zusätzlicher Variablen für die Imputation herangezogen werden: „Dabei sind vor allem solche Variablen in die Imputationsmodelle aufzunehmen, die in den inhaltich interessierenden Analysen eine Rolle spielen, aber auch solche, die für die Stichprobenziehung oder die Schätzung von Responsewahrscheinlichkeiten relevant sind [...]"(Spieß 2008, S. 64).[95] Ein sehr großer Datensatz wie der ALLBUS liefert hierzu denkbar günstige Voraussetzungen. Dies gilt gerade für das Beispiel 3 mit der problematischen Einkommensvariablen und dem damit verbundenen hohen Ausfall. Dass – wie aus Abschnitt 4.3.1 hervorgeht – für die hier durchgeführten Quasisimulationen, entgegen der allgemeinen Forderung, sparsame Imputationsmodelle verwendet wurden, ist der Handhabbarkeit und dem Rechenaufwand geschuldet. Die Sparsamkeit des Imputationsmodells hat zur Folge, dass für die hier beschriebenen Ergebnisse konstatiert werden kann, dass im Einzelfall MI ein weitaus höheres Potential verspricht.
Als klare Empfehlung lässt sich deshalb für die Korrektur von Item Nonresponse generell die Anwendung von MI aussprechen. Nur im Falle sehr geringen Datenausfalls bei der Schätzung multivariater Parameter sollte der Aufwand im Verhältnis zum plausibel erwartbaren Nutzen bedacht werden.

[94] Spieß empfiehlt nach Rubin MI bis maximal 30 % Ausfall, auf keinen Fall aber 50 % oder mehr, vgl. Spieß (2008), S.64.

[95] Dennoch sollte ein hohes Maß an Sorgfalt bei der Auswahl walten; das Ideal liegt dabei dann eher auf der Personalunion von *Imputer* und *Analyst*.

## 4.4 Zwischenfazit Item Nonresponse

Zu Beginn des Kapitels wurde Item Nonresponse als ein spezielles Ergebnis des S-P-R-Modells betrachtet (Abschnitt 3.2.2), das durch kognitive Zustände, aber auch Erwartungshaltungen und andere Faktoren begünstigt oder erschwert (Abschnitt 3.2.1) wird. Die Vielzahl von theoretischen Aspekten konnte durch den Survey Lifecycle systematisiert werden. Die theoretischen Überlegungen führten zu einigen Annahmen über die Plausibilität möglicher Ausfallmechanismen (Abschnitt 3.2.4).
Die anschließende Item Nonresponse-Analyse am Beispiel des ALLBUS 2006 bestätigte eine Vielzahl von Hypothesen, die in der älteren und neueren Literatur aufgestellt wurden, erneut. Bei der Schätzung des Erklärungsmodells für Item Nonresponse hat sich dabei die modifizierte Poissonregression als Lösung gezeigt, die sowohl nullinflationierte Daten als auch Overdispersion am besten modellieren konnte. Dabei konnten eine Reihe von wichtigen, den Datenausfall erklärenden Variablen identifiziert werden.
Im letzten Teil des Abschnitts über Item Nonresponse wurde die CC-Methode mit MI anhand dreier Beispiele, deren Autoren mit ALLBUS-Daten gearbeitet haben, verglichen. Die Ergebnisse sprechen zum Teil stark für die Verwendung der MI, und das, obwohl die Arbeit mit zu NMAR tendierenden Extremtypen alles andere als MI-freundliche Bedingungen bedeutete. Die Ergebnisse sprechen in diesem Zusammenhang auch gegen die starke Trennung von Datenerzeuger und Datennutzer sondern für den Datennutzer, der aufgrund weitreichender Kenntnis über den Datenentstehungsprozess zu sinnvollen Entscheidungen über die Anwendung von komplexeren Korrekturmethoden kommt. Die Hürde zur Anwendung von MI sinkt dabei zusehens: Es soll abschließend noch einmal darauf hingewiesen werden, dass die für die drei Beispiele verwendeten MI-Routinen lediglich Standardroutinen waren, die mittlerweile für nahezu jede Statistiksoftware implementiert sind.

# 5 Unit Nonresponse: Theorie und Determinanten

Nachdem in Kapitel 3 und 4 intensiv die Ursachen, Auswirkungen und Korrektur von Item Nonresponse diskutiert wurden, befassen sich Kapitel 5 und 6 der vorliegenden Arbeit mit dem Problem Unit Nonresponse. Im Survey Lifecycle manifestiert sich Unit Nonresponse vor Item Nonresponse. Tritt Unit Nonresponse auf, schließt das im Weiteren Item Nonresponse ein. Aus der Item Nonresponse-Perspektive bedeutet Unit Nonresponse gleichzeitig 100 % Item Nonresponse, nachdem aus verschiedenen Gründen die Befragung erst gar nicht zustande kommt. Da über Unit Nonresponse nur sehr begrenzt Informationen für Analysen vorhanden sind und zudem in den letzten Jahrzehnten diese Ausfallart tendenziell zunimmt (Schnell 1997; de Heer 1999), stellt Unit Nonresponse eine große Herausforderung für alle Disziplinen dar, die mit quantitativen Methoden arbeiten.
Kapitel 5 und 6 sind im Wesentlichen ähnlich aufgebaut wie Kapitel 3 und 4. Abweichungen wie z.B. bei der theoretischen Übersicht des Unit Nonresponse-Problems ergeben sich aus dem Gegenstand. Anders als bei Item Nonresponse wird beispielsweise bei Unit Nonresponse stärker nach den Ausfallgründen getrennt: Ist die Person nicht befragungsfähig, nicht erreichbar oder verweigert sie die Teilnahme? Auch mit dieser Differenzierung muss noch diskutiert werden, ob eine feinere Gliederung sinnvoller sein könnte. Nach der kurzen Vorstellung wichtiger theoretischer Konzepte zur Erklärung von Unit Nonresponse werden für den ALLBUS empirische Ergebnisse präsentiert. Wie schon in den Item Nonresponse-Kapiteln folgt auf die Theorie und die empirischen Ergebnisse ein Vergleich von Korrekturverfahren.

## 5.1 Kontext: Unit Nonresponse im Survey Lifecycle

Auch für die Einordnung des Unit Nonresponse-Problems wird der Survey Lifecycle herangezogen (Abbildung 46), wenngleich die Systematisierung in *nichtbefragungsfähig*, *nichterreichbar* und *verweigert* in der Literatur dominiert und im weiteren ebenso berücksichtigt wird.

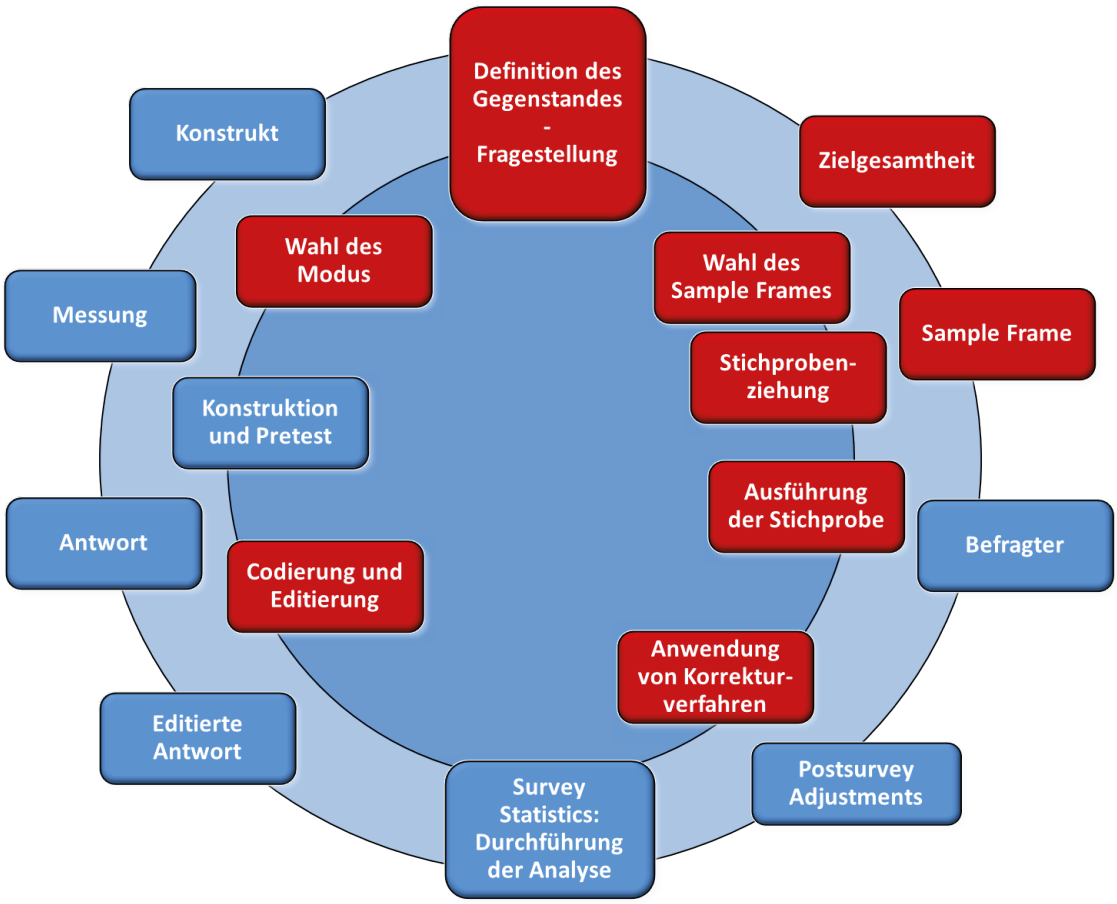



Abbildung 46: Survey Lifecycle und Unit Nonresponse

Nach dem Survey Lifecylce finden sich die ersten Determinanten für Unit Nonresponse an der Spitze des Kreises mit der **Definition des Gegenstandes** und der **Fragestellung**. Hierbei muss zwischen mit nur einem sensiblen Thema befassten Erhebungen (z.B. Erhebungen zu Sexual- oder Suchtverhalten, Erhebungen zu medizinischen Themen) und Mehrthemenbefragungen (z.B. eben ALLBUS, SOEP) unterschieden werden. Bei ersteren beeinflusst häufig die sensible Fragestellung das Auftreten von Unit Nonresponse (Lensveldt-Mulders 2008, S.464f). Auf der linken Seite, die die Messfunktion der Erhebung zeigt, beeinflusst die **Wahl des Erhebungsmodus** nach den empirischen Ergebnissen der Literatur die Wahrscheinlichkeit von Unit Nonresponse (Groves et al. 2004, S.153ff; De Leeuw 2008, S.127ff). Auf der anderen Seite, die die Repräsentationsfunktion der Erhebung abbildet, bilden die Definition der **Zielgesamtheit**, die **Wahl des Sample Frames** und die **Stichprobenziehung** Faktoren des Unit Nonresponse-Auftretens (Lynn 2008). Manifest wird Unit Nonresponse dann bei der Ausführung der Stichprobe vor dem Interview, wenn versucht wird, mit der Zielperson Kontakt für das Interview herzustellen. Je nach Definition ist auch nach dem Interview eine **Editierung** als Unit Nonresponse möglich, wenn sich beispielsweise herausstellt, dass die Zielperson nicht zur Zielgesamtheit gehört (Groves et al. 2004, S.76f). Dabei werden in der Regel diese Elemente nicht als eigentliche Unit Nonresponse gezählt. Schließlich wird häufig versucht, Unit Nonresponse durch **Postsurvey Adjustments** nach der Erhebung zu

korrigieren. Neben den Designgewichten finden sich in vielen Datensätzen auch Gewichte, die Unit Nonresponse auszugleichen suchen(Rösch 1994, S.9f; Biemer und Christ 2008, S.327). Auf der anderen Seite muss beachtet werden, dass – anders als bei Item Nonresponse – der Datennutzer über Art und Umfang der Unit Nonresponse im Normalfall nur durch die Dokumentation der Erhebung informiert wird. Unit Nonresponse werden nahezu immer bei der **Editierung** des Datensatzes vollständig aus dem Datensatz getilgt, ehe der Datennutzer den Datensatz verwendet. Eine bisher starke Fixierung auf Item Nonresponse und deren Korrektur ist eine der Folgen, ohne dabei den Informationswert vieler Erhebungsvariablen, die auch für Unit Nonresponse vorliegen, in die Korrekturüberlegungen mit einzubeziehen.[96]

Um den theoretischen Rahmen Survey Lifecycle angemessen verwenden zu können, muss zunächst über die Differenzierung von Unit Nonresponse diskutiert werden. Die übliche Unterscheidungsweise in *neutrale* und *systematische* Ausfälle wird ergänzt durch die Unterscheidung nach dem Ausfallgrund, der sich in folgende Reihenfolge bringen lässt: Nichterreichbare zur Befragung ausgewählte Personen (*Nichterreichbare*), ausgewählte zu Befragende, die nicht teilnehmen konnten (*Nichtbefragbare*) und schließlich Personen, die nach der Kontaktierung die Teilnahme an der Erhebung verweigern (*Verweigerer*) (Groves et al. 2004, S.169f). Das Adjektiv *neutral* meint in diesem Zusammenhang zufällige Ausfälle. So kann die Nichterreichbarkeit einer Person damit zusammenhängen, dass die Adresse falsch oder die Person verzogen ist, was nach Lesart der meisten Erhebungsinstitute einen *neutralen* Ausfall darstellen würde; Nichterreichbarkeit kann allerdings auch auf ungewöhnliche Arbeitszeiten der Zielperson hindeuten – ein Zusammenhang mit einzelnen Themen und Fragestellungen der Erhebung wäre dann nicht mehr ausgeschlossen. Dann könnte der Ausfall als systematischer bezeichnet werden. Da derlei Unterscheidungen der Komplexität der Ausfälle nicht gerecht werden, schlägt Schnell folgende Einteilung der Unit Nonresponse-Ursachen, die sich ebenfalls chronologisch ordnen lassen (Schnell 1997, S.18), vor:

[96] Dabei wird zum einen das Thema Paradaten und Metadaten tangiert, zum anderen gehört eine ausführliche Unit Nonresponse-Dokumentation zu den Schlüsselindikatoren der Erhebungsqualität, vgl. Mohler et al. (2008), S.406.

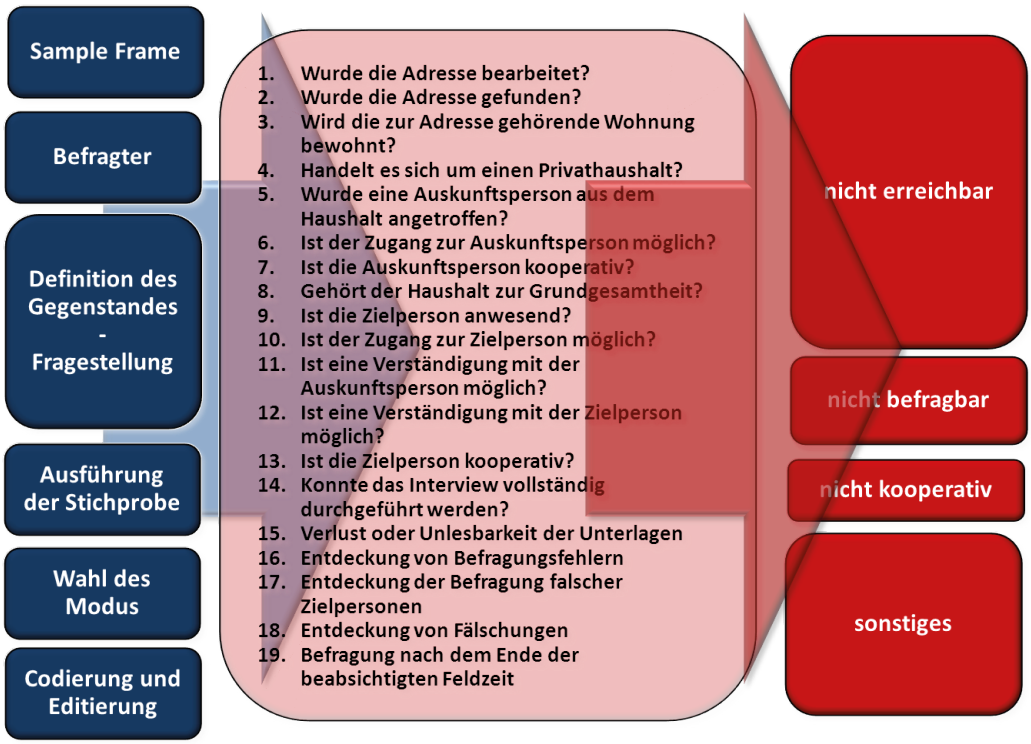



Abbildung 47: Übersicht über Gründe von Unit Nonresponse nach Schnell (1997) mit Zuordnung zu Elementen des Survey Lifecycle

Die 19 Fragen, hinter denen 19 verschiedene Ausfallgründe stehen, korrespondieren im hohen Maß mit der üblichen Differenzierung in der Literatur (mittlere Spalte). Um eine eindeutige Systematisierung von Unit Nonresponse im Kontext der gesamten Erhebung zu leisten, werden für die Ausfallgründe einzeln die Faktoren aus dem Survey Lifecycle analysiert (angedeutet bereits in der letzten Spalte), auch wenn der Umfang der Literatur zu Ausfallgründen und Elementen des Survey Lifecycle sehr unterschiedlich ist und keine strenge Chronologie eingehalten wird. Wenn eine Person erst nach der Feldzeit als eigentlich nicht zum Frame gehörig identifiziert wird, wurde der Fehler z.B. schon bei der Framebearbeitung begangen, aber erst bei der Editierung der Daten bemerkt und korrigiert.

### 5.1.1 Nichterreichbarkeit im weiteren Sinne (Undercoverage)

Die zunächst betrachtete Gruppe mit fehlenden Werten wird teilweise als eigenständige Gruppe definiert oder findet keine Beachtung. Es geht zunächst um *Undercoverage*, sodann um *Overcoverage* (Lessler und Kalsbeek 1992). An dieser Stelle soll eine Visualisierung diese Begriffe veranschaulichen (Abbildung 48):

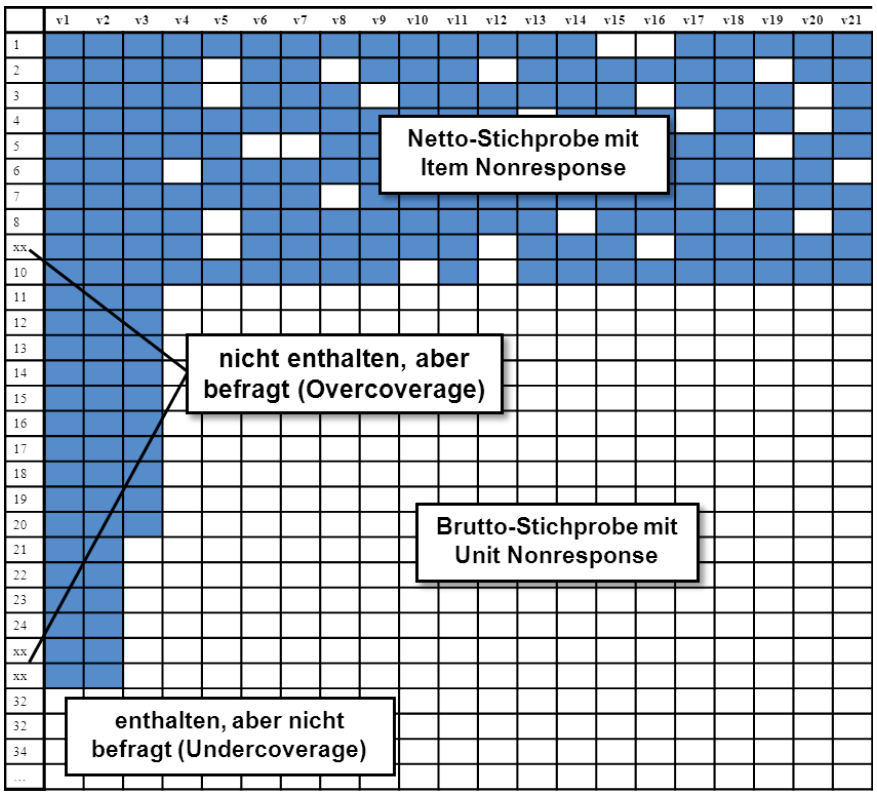



Abbildung 48: Einordnung von Ausfallgründen

Gemäß dem Survey Lifecycle werden zu Beginn einer Erhebung neben der Fragestellung auch die Objekte, über die Informationen erhoben werden, definiert. Zwar lässt sich relativ einfach in der Theorie die **Zielgesamtheit** bestimmen, jedoch ist die Auswahl des **Sample Frames** schwierig, da in der Praxis oftmals Abstriche zwischen der Ziel- und der Auswahlgesamtheit gemacht werden müssen. Obwohl von empirischer Seite dieser Aspekt häufig vernachlässigt wird, wird in diesem Erhebungsabschnitt bereits Unit Nonresponse generiert, ohne dass sich dies vermeiden ließe.[97] Zwei Möglichkeiten kommen für eine Abweichung in Frage: zum einen ist es denkbar, dass der Sample Frame Personen enthält, die nicht zur Zielgesamtheit gehören (Overcoverage), zum anderen können Elemente aus irgendeinem Grund nicht im Sample Frame enthalten sein,

[97] Als ziemlich kostspieliger und aufwendiger Ausweg wird bisweilen die Benutzung mehrerer Frames vorgeschlagen, vgl. Groves et al. (2004), S.86ff.

obwohl sie laut Zielgesamtheit darin enthalten sein sollten (Undercoverage) (Madow et al. 1983, S.16). Beide Fehler stellen Formen von Unit Nonresponse dar, wenngleich sie bisweilen nicht explizit als solche aufgeführt werden. Der Fall von Undercoverage kann allerdings sehr häufig nicht identifiziert werden, es besteht keine Korrekturmöglichkeit (Lohr 2008, S.100f). Da in der Realität kein Sample Frame perfekt ist, kann aber davon ausgegangen werden, dass auf diesem Weg Unit Nonresponse entsteht, wenn auch „ungezählte"; Madow weist beispielsweise darauf hin, dass der weit verbreitete Frame *Adressenliste für Anwesen* zum Zeitpunkt der Befragung häufig bereits nicht mehr aktuell ist (Madow et al. 1983, S.17). Dies gilt nicht nur für den **Erhebungsmodus** des persönlichen Interviews, sondern auch für telefonische Erhebungen, soweit sich der Frame zu sehr an das Telefonbuch hält. Im Falle der Overcoverage kann nach der Erhebung der fälschlich **Befragte** noch aus der Erhebung ausgeschlossen werden, wenn eine sorgfältige Prüfung bei der **Editierung** vonstattengeht und der Interviewer den Umstand der irrtümlichen Befragung selbst nicht bemerkt hat. Zwar wird diese Art von Unit Nonresponse als *neutraler* Ausfall gewertet, der andere Fall jedoch nicht. So ist es vorstellbar, dass bei einer Erhebung zu Armut Register der Einwohnermeldeämter gezogen werden, jedoch eine Gruppe, die besonders von Armut geprägt ist, nämlich Obdachlose, kaum in den Sample Frame gelangen.[98] Schnell (1997) führt daneben die Entscheidung des **Interviewers** an, der z.B. fälschlicherweise Ausländer der Kategorie „Haushalt gehört nicht zur Grundgesamtheit" (gilt als *neutral*) oder der Kategorie „Zielperson spricht nicht deutsch" (unter Umständen *systematischer* Ausfall) zuteilen könnte.[99] Eine interviewerabhängige Mischform stellt bei persönlichen Interviews der Ausfall eines Sample Point dar. Zwar sind die Elemente framegerecht in der Stichprobe, sie werden jedoch nicht befragt, da der Interviewer beispielsweise längerfristig schwer erkrankt (Schnell 1997, S.23f).[100]
Daneben gibt es allerdings kaum empirische Erkenntnisse über die soziodemografischen Merkmale von Personen, die, obwohl intendiert, nicht im Sample Frame enthalten sind. Undercoverage bildet somit in verschärfter Form eine methodische Herausforderung.

[98] Zwar muss in der Bundesrepublik Deutschland jede Person mit einer Adresse gemeldet sein, doch bedeutet dies nicht, dass die Person dort wohnen muss.

[99] Ähnliche Grauzonen existieren häufiger in der Standard-ADM-Stichprobe, vgl. Schnell (1997), S.25.

[100] Beim ALLBUS werden beispielsweise in diesem Fall Ersatzadressen bearbeitet; mehr zu diesem Thema in Abschnitt 5.2.2.

### 5.1.2 Nichterreichbarkeit im engeren Sinne

Wie beim Thema Coverage angedeutet, findet auch die *Nichterreichbarkeit* weniger Aufmerksamkeit in der Fachliteratur als das später noch zu diskutierende Thema Verweigerung (Schnell 1997, S.217). Ein nicht erreichbarer **Befragter** gehört nach Schnell (1997) zu einer der folgenden Gruppen: Personen, die sich aufgrund langandauernder Abwesenheit oder aufgrund ihres Lebensstils nicht in der Wohnung aufhalten. Desweiteren gibt es die Gruppe von Personen, die sich in der Wohnung aufhalten, aber nicht auf Kontaktversuche reagieren. Die letzte Gruppe, die Schnell benennt, sind Verweigerer, die vom Interviewer als nicht erreichbar deklariert werden, was natürlich eine Fälschung von Seiten des Interviewers darstellt (Schnell 1997, S.218). Von besonderer Bedeutung sind quantitativ und qualitativ die ersten beiden Gruppen, da sie nach der Literatur zu den nichtzufälligen Ausfällen gehören. Im Survey Lifecycle konzentriert sich die Diskussion um Nichterreichbare vor allem auf den Faktor **Erhebungsmodus** und **-design** auf der Messseite und auf die Realisation des Sample Frames durch den Befragten (auf der Repräsentationsseite).
Die empirischen Befunde konzentrieren sich auf wenige Variablen: Haushaltsgröße, Haushaltszusammensetzung und Erwerbsstatus (Engel et al. 2004, S.53) und passen damit ins Bild der oben aufgeführten Gruppenbildung. Wesentlich ergiebiger gestalteten sich allerdings die Untersuchung von Effekten des **Interviewerverhaltens** und des Designs (De Leeuw 2008, S.123, S.127ff). Für die Designeigenschaften einer Studie lässt sich je nach dem Erhebungsmodus konstatieren, dass Follow-ups, Vorankündigungen der Befragung, gewisse Personalisierungen und Erinnerungstechniken sowie Incentives die Erreichbarkeit erhöhen können (Traugott et al. 1987; Singer et al. 2000; Singer 2002; Harkness et al. 1998; Cantor et al. 1998; White et al. 1998).
Da die Erreichbarkeit (als Kontaktwahrscheinlichkeit) als Funktion von Kontaktzeiten von **Interviewer** und **Befragtem** betrachtet werden kann (Groves und Couper 1998, S.79ff), wird diese durch einige Designfaktoren erhöht oder vermindert, wenn Zeiten, in denen wenigstens eine Kontaktperson angetroffen wird, Zeiten, in denen der Interviewer den Kontaktversuch startet, und die Möglichkeit physischer Hindernisse (z.B. Anrufbeantworter, hohe Mauern) berücksichtigt werden (Groves et al. 2004, S.170; Groves und Couper 1998, S.105ff; Heberlein und Baumgartner 1978). Bei telefonischen Erhebungen sind dies die Anzahl der Anrufe sowie das Anrufmuster, bei persönlichen Interviews entsprechend Kontaktanzahl und Besuchsmuster, generell die Länge der Feldzeit, aber auch die Belastung der Interviewer (Engel et al. 2004, S.53f; Botman und Thornberry 1992).
Um die Vielzahl aller als relevant geltender Faktoren in eine Übersicht zu bringen, haben Groves und Couper „A Simple but Powerful Model of Contactability“ grafisch dargestellt (Groves und Couper 1998, S.81):

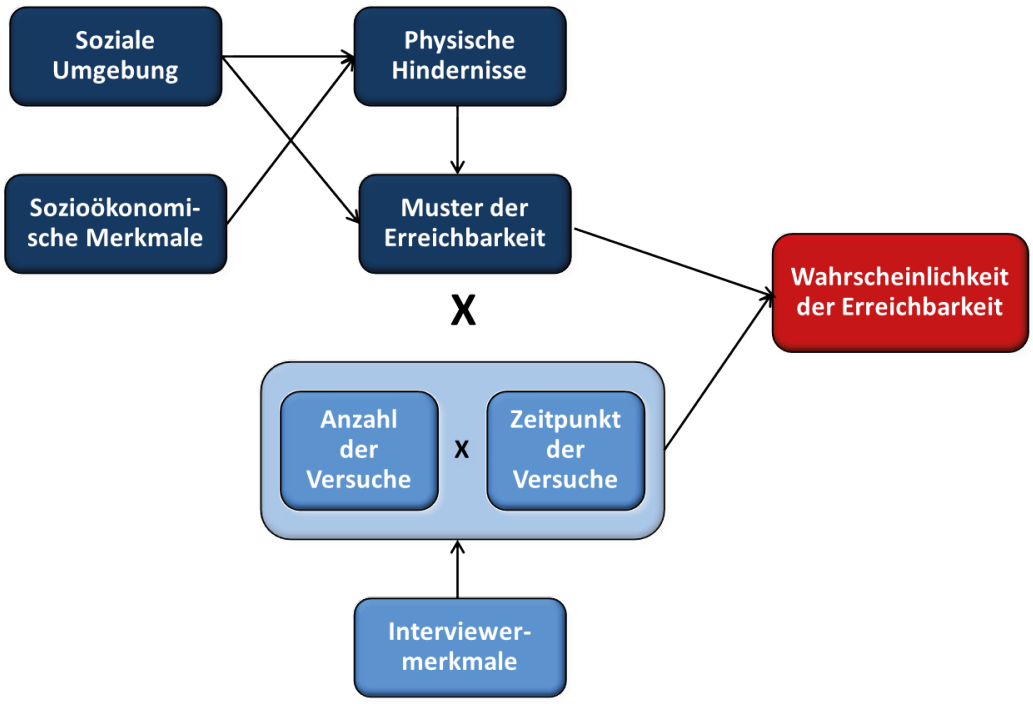



Abbildung 49: Schema Kontaktwahrscheinlichkeit nach Groves und Couper (1998)

Um noch einmal auf die Rolle des **Interviewers** bzw. dessen Kontaktverhalten zurückzukommen: „Die Erreichbarkeit der Zielperson hängt offenbar auch vom Kontaktverhalten der Interviewer ab“(Schnell 1997, S.224). Ähnlich wie im Kapitel über Item Nonresponse kann die Kontaktaufnahme des Interviewers mit dem potentiellen Befragten bereits als verkürzte Form der Interaktion aufgefasst werden und empirische Befunde legen nahe, dass manche Interviewer eine höhere Kontaktrate aufweisen als andere (Lievesley 1986). Ob es nun zu einer Verweigerung kommt oder nicht, wird später noch behandelt, wichtig ist zunächst die Frage nach der erfolgreich hergestellten Interaktion. Die einschlägige Literatur zu Kontaktstrategien nennt relativ einhellig den Umstand, dass Kontaktversuche am späten Nachmittag oder Abend und am Wochenende wesentlich erfolgversprechender sind (Groves und Couper 1998, S.95ff). Gerade diese Zeiten bedeuten jedoch auch einen höheren Aufwand für den Interviewer und damit für das Erhebungsinstitut. Dies würde unter anderem den deutlichen „Institutseffekt“ erklären, den Schnell in seiner großen Studie als Schlüsselindikator für diese Art von Unit Nonresponse beschreibt, da sich die Institute gerade in dieser Hinsicht unterscheiden (Schnell 1997, S.219ff).

### 5.1.3 Nichtbefragbarkeit

*Nichtbefragbarkeit* ist ein Urteil, das der **Interviewer** über die Person, die eigentlich befragt werden soll, fällt. Als wichtigste Gründe der Nichtbefragbarkeit oder Befragungsunfähigkeit gelten: Defizite, die Sprache zu verstehen oder lesen zu können, physische oder psychische Krankheiten sowie intellektuelle Defizite (Groves et al. 2002, S.178). Die Gruppe der nichtbefragbaren Personen ist in der Regel gering, jedoch kann sich ihr Umfang verzerrend in Hinblick auf ein Erkenntnisinteresse auswirken, wenn es im direkten Zusammenhang mit dem Grund der Nichtbefragbarkeit steht. Die Defintion von *Nichtbefragbarkeit* von Personen leidet vermutlich unter der Ungenauigkeit und der allzu subjektiven Einordnung der Interviewer. Zudem gilt bespielsweise „Krankheit" der Zielperson als gängige Ausrede von Kontaktpersonen. Dies bedeutet darüber hinaus, dass die Anzahl der im Haushalt wohnenden Personen ausschlaggebend sein kann (Schnell 1997, S.108). Das Beispiel „Krankheit" verdeutlicht auch die Problematik der Abgrenzung, da sich eine kranke Person auch im Krankenhaus befinden kann und damit *nichterreichbar* im engeren Sinn ist. Es gibt Erhebungen, die Nichtbefragbarkeit generell als *neutralen* Ausfall bewerten, ohne zu berücksichtigen, dass viele Gründe für Nichtbefragbarkeit mit anderen Variablen korrelieren: Alter, Einkommen, Nationalität. Ausfallneutralität ist in diesem Zusammenhang deshalb sehr fragwürdig.

### 5.1.4 Verweigerung

Von den bisher behandelten Ausfällen, die eine ganze Einheit betreffen, unterscheidet sich die letzte Gruppe, die Befragungsverweigerer, in einem wichtigen Punkt: anders als Nichtbefragbare und Nichterreichbare und anders als Personen, die nicht im Frame enthalten sind, entscheiden sich diese Personen bewusst zur Nichtteilnahme an einer Erhebung.[101] Damit sind Verweigerer zunächst den **Befragten** ähnlich, die in einer Befragung keine **Antwort** geben (Item Nonresponse). Der Stimulus bzw. die Stimuli unterscheiden sich aber quantitativ grundlegend. Während innerhalb einer Befragung der Befragte vielen Fragen (Stimuli) ausgesetzt wird, stellt die Anbahnung des **Interviews**, der Kontakt, lediglich ein Stimulus dar, auf den der potentielle Befragte mit lediglich zwei Optionen konfrontiert wird: entweder nimmt er an der Erhebung teil oder nicht. Diese Analogie verantwortet wohl auch die teilweise ähnlichen Theoriezugänge bei Item Nonresponse und dem Unit Nonresponse-Spezialfall Verweigerung.
Zwar gibt es Bestrebungen, für Unit Nonresponse eine allgemeine Theorie aufzustellen, die bisherige Literatur divergiert aber grob im grundlegenden Ansatz zwischen zweckrationalen und wertrationalen Erklärungsmodellen (Engel et al. 2004, S.57). Die jeweiligen Ansätze beanspruchen jedoch für sich, jeweils andere Theorien zu beinhalten bzw. diese aus dem eigenen Ansatz ableiten zu können. Damit gestaltet sich eine theoretische Übersicht nicht ganz einfach. So sind Rational-Choice-Ansätze in der Methodenforschung seit geraumer Zeit im häufigen Gebrauch (Green et al.

[101] Lynn stellt einen sehr detaillierten Ereignisbaum für Unit Nonresponse auf. Verweigerung zählt qualitativ dabei zu einem ganz speziellen Fall von Unit Nonresponse, der quantitativ aber den weitaus größten Teil der Ausfälle ausmacht, vgl. Lynn (2008), S.39.

1999); so auch bei der Frage nach einem Modell zur Teilnahme oder Nichtteilnahme an Erhebungen: „Danach wären Teilnahme oder Nichtteilnahme an einer Umfrage als Entscheidungsalternativen aufzufassen, deren Wahl im wesentlichen einer Kosten-Nutzen-Kalkulation folgt.“(Engel et al. 2004, S.55). Die erste grundlegende Annahme der Rational-Choice-Theorie, die Nutzenmaximierung, lässt sich konkret herunter brechen (Green et al. 1999, S.25). Engel et al. (2004) nennen insgesamt folgende Kosten und Nutzen einer Teilnahme:

Kosten:

- Zeit für das Interview
- Verlust an anderen Handlungsoptionen zur selben Zeit (Opportunitätskosten)
- kognitive Last
- negative Konsequenzen offen gelegter Informationen und Antworten
- Offenbarung von Informationen und Verlust der Privatheit
- Missbrauch mit Informationen

Nutzen:

+ Vermeidung lästiger anderer Aufgaben
+ Befriedigung durch Teilnahme an etwas Nützlichem
+ Befriedigung der eigenen Neugierde
+ Befriedigung des Interesses an einem Thema
+ Lust an der Interaktion mit Interviewer
+ Bestätigung der Wichtigkeit der eigenen Meinung
+ Beitrag zur Erfüllung einer Bürgerpflicht

Diese potentiellen Kosten- und Nutzenfaktoren finden immer wieder als „empirische Einzelhypothesen“(Schnell 1997, S.165) mehr oder weniger Evidenz.[102] Verschiedene Untersuchungen haben sich zunächst vorgenommen, die Erhebungen als Belastung, sei es als Störung der „privacy“, sei es als Gefährdung anderer Ziele oder als Belastung durch die Befragung als solche, zu untersuchen – ohne wirklich durchschlagenden empirischen Befund. Zudem lässt sich plausiblerweise davon

[102]Eine sehr ausführliche Zusammenfassung findet sich bei Stoop et al. (2010), S.25.

ausgehen, dass zweitens die Präferenzen in eine Reihenfolge gebracht werden können. Es dürfte die absolute Ausnahme sein, dass eine Person Teilnahme und Nichtteilnahme an einer Erhebung genau gleich präferiert (Konsistenzbedingung). Dagegen spielt bei zwei Handlungsalternativen die Transitivität keine Rolle. Drittens ist für die Teilnahmeentscheidung die Erwartung an das Ereignis Interview ausschlaggebend. Natürlich weiß der Befragte nicht restlos, was genau bei einer Befragung auf ihn zukommt. Genuin verbunden mit der Rational-Choice-Theorie (RC-Theorie) ist viertens die Annahme, „dass es sich bei den relevanten Maximierern um Individuen handelt"(Green et al. 1999, S.26), wie es bei einer Personenerhebung der Fall ist. Einige empirische Ergebnisse legen allerdings nahe, dass die Entscheidungsfindung für oder gegen eine Teilnahme auch von anderen Faktoren abhängt als der individuellen Abwägung von Kosten und Nutzen (Groves und Couper 1998, S.122). Doch schon aufgrund der begrenzten kognitiven Ressourcen von Individuen bedarf es einer Modifikation des RC-Ansatzes, um ihn wirklichkeitsnäher zu modellieren. Modifikation im Sinne einer Konkretisierung wird im folgenden Abschnitt 5.1.4.1 vorgestellt. Der RC-Ansatz ist aber auch in die Leverage-Salience-Theorie (Abschnitt 5.1.4.2) und die Theorie wertrationalen Handelns eingebettet (Abschnitt 5.1.4.3).

#### 5.1.4.1 Konkretisierung von RC in habitualisierten Verhaltenstendenzen und skripttheoretischer Spezifizierung

Wo liegen die Schwächen des RC-Ansatzes? Da Rational-Choice-Ansätze zwar unter vollständigen Informationen individuelle Prognosen über die Teilnahme denkbar erscheinen lassen, ist, wenn diese Informationen nicht gegeben sind, eine individuelle Verhaltensprognose unmöglich. Als weiteren Punkt kann man anführen, dass die Vorstellung einer ständigen Abwägung von derlei Entscheidungen im Alltag die kognitiven Fähigkeiten überreizen würde.[103] Deshalb ergänzt Schnell den RC-Ansatz, um eine realistische Modellierung von Teilnahme an Erhebungen zu erhalten (Schnell 1997, S.161ff). Habitualisierte Verhaltenstendenzen stellen eine Entscheidungsentlastung dar. Durch gewisse Reize werden Verhaltensweisen ausgelöst, im konkreten Fall eben die Verweigerung oder die Teilnahme. In der Literatur werden sechs Prinzipien genannt, die als Auslöser eines habitualisierten Verhaltens gelten (Cialdini und Sagarin 2005; Cialdini 1994; Cialdini 1989): So spielen die Prinzipien Reziprozität, Konsistenz, sekundärer Vergleich, Knappheit und Autorität die entscheidenden Rollen bei der Systematisierung von Entscheidungsprozessen. Ähnliche Faktoren werden auch in umfassenderen Entwürfen zu Teilnahmetheorien verwendet (Couper und Groves 1998). Kritisch lässt sich hinterfragen, ob tatsächlich ein Automatismus sinnvoll die Realität einer Interaktion wie des Befragungskontakts abbildet.
Deshalb werden statt habitualisierten Verhaltenstendenzen auch Skripte als Spezifikation der Rational-Choice-Theorie genannt: „Im Unterschied zu Habits als reinen Verhaltensprogrammen sind Skripte Wissensstrukturen" (Schnell 1997, S.163). Was es heißt, nach einem Skript zu handeln, fasst Abelson in folgenden Punkten zusammen: erstens die Existenz einer stabilen Repräsentation eines Skripts, zweitens die Auslösung des Skripts durch eine Situation und drittens die Ver-

[103] Dies sind auch klassische Kritikpunkte gegen RC-Ansätze in anderen sozialwissenschaftlichen Kontexten, vgl. Green et al. (1999).

knüpfung eines solchen Skripts mit einer Verhaltensregel (Abelson 1981; Bredenkamp und Vaterrodt 1992). Schnell merkt an, dass auch die skripttheoretische Ergänzung zur Rational-Choice-Theorie keine wirklichen Verhaltensprognosen zulässt, wenn sie nicht noch durch weitere Parameter spezifiziert wird (Schnell 1997, S.164).
Zwei weitere nennenswerte Ansätze versuchen das Dilemma zwischen Überspezifikation, Generalisierung und fehlenden Informationen zu lösen, indem beim ersten Ansatz Parameter für eine rationale Teilnahmeentscheidung identifiziert werden und beim zweiten Ansatz die rationale Komponente zu einer wertrationalen Komponente erweitert wird.

#### 5.1.4.2 Leverage-Salience-Theorie

Die Leverage-Salience-Theorie systematisiert zunächst Merkmale, deren empirische Relevanz bisher in der Literatur bestätigt wurde, aber auch darüber hinaus Variablen, die für die Erklärung sinnvoll erscheinen. Insgesamt lässt sich von vier Hauptfaktoren ausgehen, die das Auftreten von Unit Nonresponse beeinflussen (Groves et al. 2004, S.176; Groves und Couper 1998): Die Soziale Umgebung, persönliche Merkmale der Zielperson, das Survey Design und die Merkmale des Interviewers. Diese Hauptfaktoren sind alle direkt aus dem Survey Lifecycle entlehnt. Sie machen je nach Ausprägung die Verweigerung wahrscheinlicher oder unwahrscheinlicher.
Die Faktoren auf der linken Seite können vom Forscher nicht verändert werden (Abbildung 50). Diese Merkmale haften am potentiellen Befragten oder bilden seine Umwelt.[104] Anders verhält es sich mit den beiden rechten Faktoren, welche durchaus im Rahmen der Erhebungsausgestaltung beeinflusst und gesteuert werden können. Ähnlich wie Schnell erkennen die Vertreter der Leverage-Salience-Theorie, dass die Faktoren soziale *Umgebung* und *Befragtenmerkmale* sehr unterschiedliche Evidenz in verschiedene Studien aufweisen, und die Vermutung nahe liegt, dass eine geeignete Theorie die Varianz der Entscheidungsgründe für oder gegen die Teilnahme aushalten und systematisieren muss. Zudem sind diese beiden Faktoren allein kaum in der Lage, etwas über die persönlichen Gründe auszusagen.

[104] Wobei die Trennung zwischen Merkmalen, die den Befragten direkt anhaften und Umgebungsmerkmalen in der Literatur nicht immer gleich gehandhabt werden. Deshalb wird in Abbildung 50 auch von Befragtenmerkmalen im weitesten Sinn gesprochen.

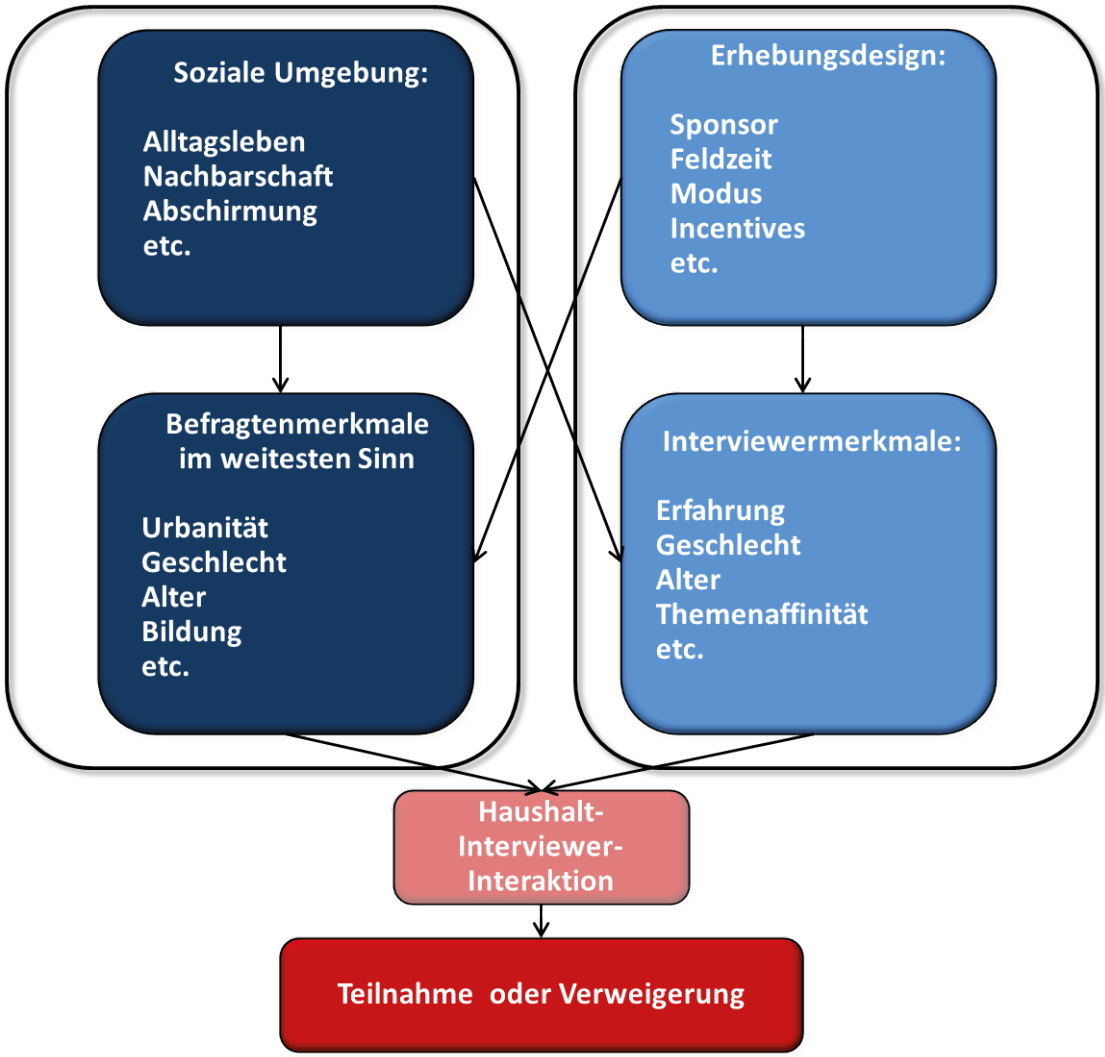



Abbildung 50: Modell zur Verweigerung nach Groves und Couper (1998)

Die eigentlichen Gründe für eine Verweigerung werden deshalb folgendermaßen zusammengefasst: Zunächst lässt sich ein Teil der Gründe unter *Opportunitätskosten* subsumieren. Dies knüpft nahtlos an diverse RC-Ansätze an, da in diesem Rahmen gerade Zeitmangel als häufiger Grund für Verweigerung sogar explizit angegeben wird (Stoop 2007). Gerade Personen, die aufgrund persönlicher oder beruflicher Verpflichtungen wenig Zeit haben, verweigerten häufiger die Teilnahme (Groves et al. 2004, S.176). Zweitens nennen Groves und Couper die „Exchange Hypotheses“, die einmal bereits von Dillmann (1978) und von Groyder (1983) in verschiedenen Varianten diskutiert wurden. Dabei geht es um die Kalkulation, die der mögliche Befragte aufstellt, wenn er in eine soziale Interaktion mit dem Interviewer tritt.[105] Als dritter Grund für Verweigerung wird Soziale Isolation – oder Deprivation – angeführt. Die Befragung wird dann zum Ausbruch aus

[105] In der Regel wird die Hypothese mit sozio-ökonomischen Indikatoren, Bildung und Empfang von Transferleistungen operationalisiert, vgl. Groves und Couper (1998), S.127ff.

der Isolation, der jedoch von bestimmten Gruppen nicht wahrgenommen wird, da die eigene Zugehörigkeit zur Gesellschaft bezweifelt wird. Viertens lässt sich das Interesse oder die Zentralität eines Themas als Teilnahmegrund (Groves et al. 2004, S.176; Schnell 1997, S.181ff) benennen; und schließlich – fünftens – als auch in der Literatur neuer Trend, kann das Oversurveying für die Verweigerung verantwortlich sein. Oversurveying verursache bei den potentiellen Befragten Interviewmüdigkeit, auch wenn dieser Punkt teilweise bestritten wird (z.B. von Schnell 1997, S.171ff). Die hier genannten Ausfallgründe, die im Wesentlichen auf größeren sozialwissenschaftlichen Theorien fußen, können teilweise empirisch belegt werden, teils stehen sie im Einklag mit existierenden RC-Ansätzen. Die Erklärungskraft der vier Hauptfaktoren (persönliche Merkmale, Design-Merkmale, Interviewmerkmale, soziale Umgebung) kann auch mit dem betrachteten Datensatz des ALLBUS 2008 in Teilen getestet werden (Abschnitt 5.3.1.2).

Die eigentliche Leverage-Salience-Theorie lässt sich durch eine Metapher beschreiben:

> „Consider a sacle with multiple hooks on which to place weights, each hook representing some attribute of the request that could be judged relevant to the decision. The distance from the fulcrum to the hook measures the importance the sample person assigns to the attribute in the decision to participate [...]." (Groves et al. 2000, S.300)

Der nun von Groves et al. (2000) vorgeschlagene Rahmen ist eine Systematisierung des bereits 1998 ausgeführten Konzepts des „tailorings" (Groves und Couper 1998), das Interviewern als Richtschnur dienen soll, bei der Kontaktaufnahme vermutlich wichtige Faktoren der Verweigerung zu erkennen und entgegenzusteuern. Verbildlicht sieht dieser Rahmen so aus:

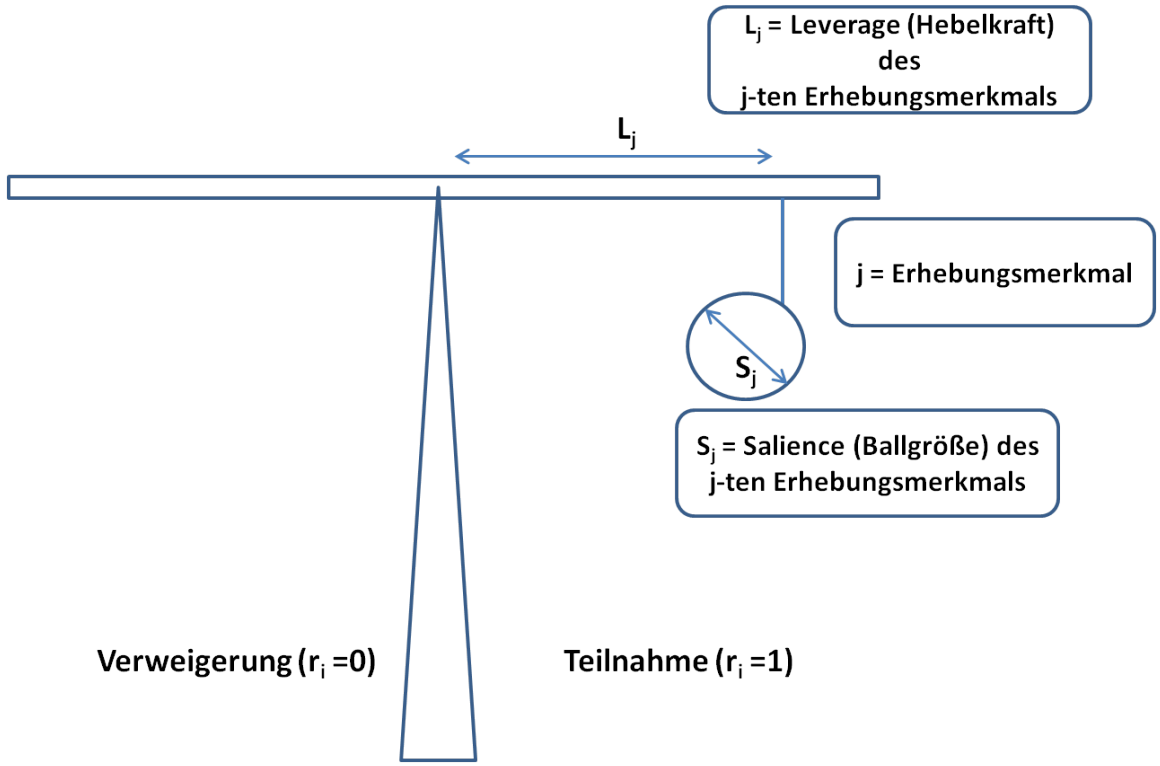



Abbildung 51: Schaubild zur Leverage-Salience Theorie nach Groves et al. 1998

Die Distanz vom Drehmoment bis zum Kurvball in Abbildung 51 wird als „Leverage“ *L* bezeichnet und beschreibt das Gewicht, das ein Merkmal der Erhebung (Erhebungsmerkmal) für die i-te Zielperson besitzt. Diese Distanz wird beeinflusst von persönlichen Erfahrungen und Charaktereigenschaften sowie vom gesellschaftlichen Umfeld, wie sie im obigen Abschnitt vorgestellt wurden. Die Größe der Kugel verdeutlicht das Hervorspringen („Salience“) *S* der Erhebungsmerkmale im Zuge der Kontaktanbahnung, die mehr oder weniger für die Zielperson wahrnehmbar sind. Diese sind direkt von Designentscheidungen und dem Konzept der Kontaktanbahnung abhängig (Broschüre des Sponsors, Incentives usw.). Man kann also von designindogenen und -exogenen Erhebungsmerkmalen sprechen.
Die Wahrscheinlichkeit an der Erhebung teilzunehmen, ließe sich dann als Funktion von *L* und *S* darstellen, wenn die Faktoren, die *L* bestimmen, bekannt wären. Dies ist aber unrealistisch. Groves et al. (2000) verweisen darauf, dass Proxyvariablen zur Verfügung stünden, die als Indikatoren für *L* dienen können. Nach ihrer Einschätzung kann die Leverage-Salience Theorie nur über einen Umweg empirisch gefasst werden. Wie dieser Umweg skizziert wird, stellt einen wichtigen Fortschritt gegenüber anderen Theorien zum Teilnahmeverhalten dar.[106]

Irgendein Merkmal *C* bildet dann zusammen mit *S* die Funktion

[106]Dennoch ist die empirische Überprüfung sehr schwierig und wird bei Groves et al. auch mit einem sehr komplizierten Verfahren getestet, vgl. Groves et al. (2000), S.303ff.

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 C_{ij} + \beta_2 S_{ij} + \beta_3 C_{ij} S_{ij} + \varepsilon$$

Die Teilnahmewahrscheinlichkeit lässt sich dann als eine Funktion von *C* und *S* bzw. der Interaktion von *C* und *S* auffassen. *C* und *S* können wiederum für den i-ten potentiellen Befragten beim j-ten Erhebungsmerkmal stets anders ausfallen.

Abschließend lassen sich zwei Vorteile der Leverage-Salience-Theorie benennen: Zum einen stellt sie „a rich set of deductions for methodological research" dar (Groves et. al 2000, S.307). Eine Vielzahl von Studien, die sich dieses theoretischen Rahmens bedienen, spricht hierfür. Der Vorteil für die Methodenlehre scheint zum anderen darin zu liegen, dass sie grundlegenden Annahmen der RC-Theorie nicht widerspricht, jedoch das situationsgebundene Handeln des potentiellen Befragten auf eine praktische Ebene bringt, die es sogar erlaubt, Schlüsse für zukünftige Erhebungen zu ziehen. Dabei geht es vor allem darum, dass Interviewer positive Aspekte einer Befragung für die Zielperson sichtbar machen und negative Aspekte und Befürchtungen, die die Befragung bei der Kontaktaufnahme auslösen, zu minimieren versuchen.

#### 5.1.4.3 Konzept einer wertrationalen Erklärung für die Teilnahme

Die Motivation für Engel et al. für eine Erweiterung handlungsrationaler zu einer wertrationalen Theorie der Teilnahme liegt zunächst in der geringen Evidenz der „Opportunity Cost"-Hypothese. Daneben gäbe es empirische Hinweise für altruistische Teilnahmegründe (Engel et al. 2004, S.57f).[107] Das Konzept einer wertrationalen Theorie baut auf zwei theoretischen Fundamenten auf. Es fließen zum einen austauschtheoretische Überlegungen (Exchange-Theorie) als auch integrationstheoretische Überlegungen (Soziale Isolation) in das Modell ein. Daneben sollen aber auch rationale Komponenten, Wertorientierungen, Handlungsmotive und die Perspektive der Kollektivgutproblematik eingebracht werden (Engel et al. 2004, S.59).

[107] Zitiert wird eine Studie von Porst und von Briel (1995) über die Motivationsgründe zur weiteren Teilnahme an einer Panelerhebung.

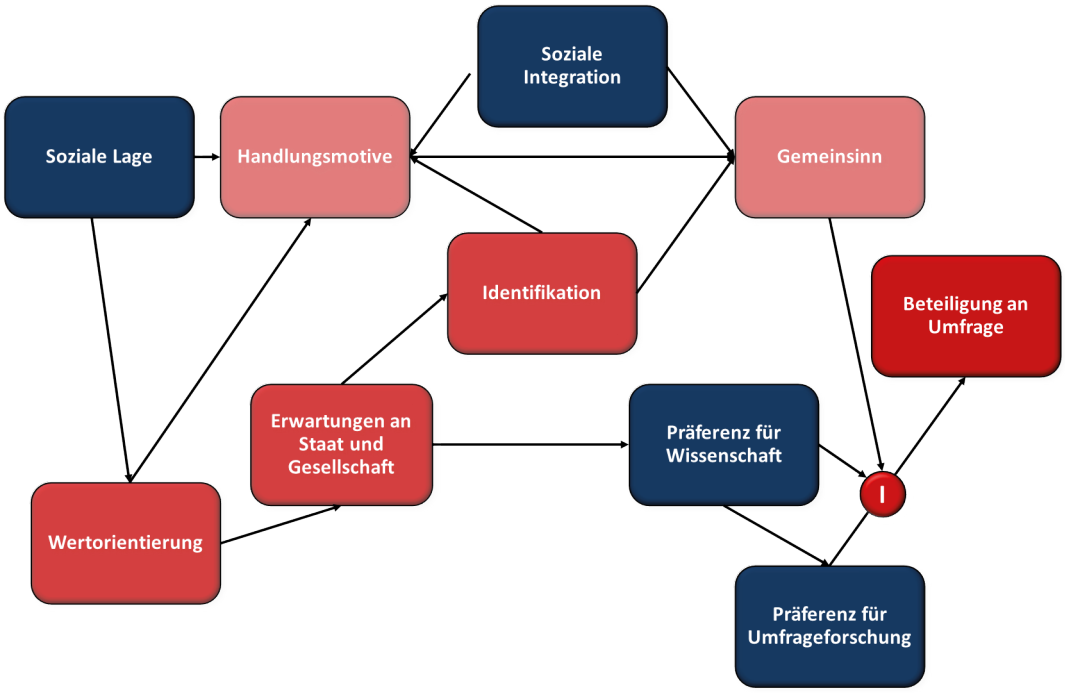



Abbildung 52: Wertrationales Erklärungsmodell der Verweigerung nach Engel et al. 2004

Engel et al. gehen in ihrem Modell von zwei Kausalketten aus (Abbildung 52). In deren Hauptglieder befinden sich einerseits „Handlungsmotive“ und „Gemeinsinn“ (Kausalkette 1) und andererseits „Wertorientierung“, „Erwartungen an Staat und Gesellschaft“ sowie die „Identifikation mit den Institutionen des Systems“ (Kausalkette 2).
Teilweise sind die Begriffe synonym für bereits eingeführte Begriffe aus anderen Theorien. So meint „Erwartungen“ an eine Erhebung und „Identifikation“ mit dem Sponsor Ähnliches wie der Begriff „authority of the sponsor“; auch „topic interest“ (Präferenz für Umfrageforschung) oder „social integration“ (soziale Integration) finden sich bei Engel et al. mehr oder weniger ähnlich wieder (beide Begriffe bei Groves et al. 2000).[108] Der Unterschied zwischen den Modellen besteht in der Betonung tatsächlich wertrationalen Handelns. Indem die Umsetzung einer Erhebung als öffentliches Gut interpretiert (überhöht) wird, lassen sich unterschiedliche Einstellungsmuster dreier Typen (Individualisten, Kollektivisten, Egalitarier) zu öffentlichen Gütern als Determinanten der Teilnahmebereitschaft anwenden (Engel et al. 2004, S.62ff).
Das Bemerkenswerte des von Engel et al. entwickelten Modells liegt tatsächlich in der Übertragung von Begriffen, die in der politischen Soziologie („civic duty“) bereits einen festen Platz in der empirischen Analyse besitzen (vgl. hierzu etwa das Konzept der „Wahlnorm“, Lüeße 2007). Zudem schließt das Konzept rationales Handeln aus, sondern gibt einen Antwortversuch auf die empirische Tatsache, dass ein Teil der Befragten tatsächlich aus „altruistischen“ Motiven wertrational an Erhebungen teilnimmt.
In diesem Abschnitt wurden zwei Modelle der Teilnahme/Verweigerung – so kurz wie möglich

[108] Genauso wie bei Groves et al.(2000) befinden sich die Anknüpfungspunkte gerade im Bereich des empirisch schwer Überprüfbaren.

– skizziert. Soweit es die Datenlage zulässt, werden Teile bei der Analyse des hier verwendeten Datensatzes berücksichtigt (Abschnitt 5.3.1.2).

### 5.1.5 Überlegungen zum Ausfallmechanismus bei Unit Nonresponse

In Abschnitt 3.2.4 wurden einige Überlegungen zusammengefasst, die sich aus der Theorie über die Entstehung von Item Nonresponse ergeben haben und Aussagen über den wahrscheinlichen Ausfallmechanismus bei Item Nonresponse zulassen. Auch für Unit Nonresponse existieren – wie soeben dargestellt – Theorien zu ihrer Entstehung. Dabei ist offenkundig, dass sich die Theorie eng an die Unterscheidung der Ausfallgründe in Nichterreichbarkeit, Befragungsunfähigkeit und Verweigerung anlehnt.
Die Überlegungen zum Ausfallmechanismus bei Unit Nonresponse werden dadurch nicht leichter. Die Hauptherausforderung liegt aber in der doppelten Unsicherheit zwischen der Person, die aus gewissen Gründen nicht an der Befragung teilnimmt, und den deshalb unbeantworteten Items der Erhebung. Der einfachste Fall ist dabei, wenn der individuelle Ausfallgrund mit dem generellen Erhebungsthema in Beziehung steht. Bei Mehrthemenerhebungen ist dies jedoch nicht eindeutig. Im Zweifelsfall müsste für die jeweilige Analyse gefragt werden, ob der Ausfallgrund für Unit Nonresponse einen verzerrenden Einfluss auf die Schätzung der Parameter haben könnte oder ob dies eher unwahrscheinlich ist. Die Überlegungen, ob der entsprechende Ausfall eher MCAR, MAR oder NMAR ist, sind dabei wesentlich spekulativer als bei Item Nonresponse.

## 5.2 Messung von Unit Nonresponse

Bereits in den ersten Abschnitten wurde deutlich, dass in Teilen der Literatur die Diskussion, was Unit Nonresponse ist und wie Unit Nonresponse gezählt werden soll, breiten Raum einnimmt. Bankhofer beschreibt beispielsweise auch Berechnungs- und Darstellungsmöglichkeiten von Item Nonresponse, die allerdings bisher kaum eine Rolle spielen (Bankhofer 1995, S.30ff). Ganz anders verhält es sich mit Unit Nonresponse, die als Ausfallrate, Rücklaufquote, Responserate, Antwortrate, Verweigerungsrate etc. gemessen und publiziert werden. Die Bestrebungen bei der Messung von Unit Nonresponse lässt sich etwa so zusammenfassen: die Ausfallursachen sollten möglichst zu den *neutralen* Ausfällen zählen (vgl. Kapitel 2.1 und 2.2). Dieses Vorgehen liegt darin begründet, dass die sogenannte Ausschöpfungsquote immer als Quotient einer bereinigten Bruttostichprobe (Abzug der *neutralen* Ausfälle von der Bruttostichprobe) und aller anderen Ausfälle bzw. der realisierten Interviews gebildet wird – hier bestehen kaum Abweichungen im wissenschaftlichen Diskurs. Gerade in der kommerziellen Sozial- und Marktforschung wird mit diesem Quotienten als Ausweis für die Qualität der Erhebung geworben, was er nicht ist bzw. sein muss. Kritisieren lässt sich jedoch noch Grundlegenderes. Da keine verbindliche Definition dieser Quoten existiert, „ergeben sich unter Umständen bei gleicher Anzahl von Ausfällen dramatische Unterschiede in den Ausschöpfungsquoten“ (Schnell 1997, S.23). In den folgenden Kapiteln wird deshalb zunächst die Veränderung der Ausschöpfungsquote vor dem Hintergrund dieser Kritikpunkte betrachtet (5.2.1), ehe die Konzepte und Veränderungen von Unit Nonresponse für den ALLBUS analysiert werden (5.2.2).

### 5.2.1 Problematik der Ausschöpfungsquote: generelle Trends

Zwei Veröffentlichungen aus der zweiten Hälfte der 1990er Jahre vertreten unterschiedliche Positionen bezüglich der Frage, ob die Ausschöpfungsquoten sinken bzw. die Nonresponseraten steigen.
Schnell konstatiert für den internationalen Trend, dass gerade das vermeintliche Ansteigen der Verweigerung „empirisch bislang zweifelhaft“ sei (Schnell 1997, S.43).[109] Betrachtet man die bundesdeutschen Verhältnisse (vor der Wiedervereinigung), so stellt Schnell insgesamt die These von steigenden Ausfallraten in Frage. Bei differenzierter Betrachtung tun sich stattdessen große Unterschiede in der Methoden- und Planungssorgfalt auf, die maßgeblich für die Höhe der Nonresponserate verantwortlich seien (Schnell 1997, S.131f). De Heer (1999) kommt zum gegensätzlichen Befund: Er bezieht sich auf mehrere, auch internationale Studien aus den USA, die eine eindeutige Zunahme von Unit Nonresponse feststellen. Aufgeschlüsselt nach Modus blieben allein postalische Befragungen stabil, während Telefon und persönliche Interviews steigende Nonresponseraten aufwiesen (De Heer 1999, S.130).
So erstaunlich es klingt, die beiden Befunde müssen sich nicht zwangsläufig widersprechen. Denn hinter der Frage nach der Entwicklung der Ausschöpfung steht zunächst die Frage, wie diese Ausschöpfung jeweils überhaupt zu berechnen ist. Die Bandbreite der Varianten ist zum einen ungeheuer groß, zum anderen dominiert die Vorstellung, dass nur eine hohe Ausschöpfungsquote

[109] Auch in einer neueren Veröffentlichung wird von „not unambiguous“ gesprochen, vgl. Stoop 2010, S.2.

eine gute realisierte Stichprobe darstellt (Stoop 2010, S.2). Gerade die kommerzielle Markt- und Sozialforschung erhebt dieses Kriterium als Qualitätsmerkmal zu „Heilige[n] Kühe[n]"(Sommer 1987, S.300f), während der Umgang mit Item Nonresponse eher von fahrlässiger Ignoranz geprägt ist. Dies ist, im Gegensatz zu fehlenden verbindlichen Regeln, wohl eher der Grund für die Berechnung der Nonresponsequote (Wiseman und Billington 1984, S.337). Da für Unit Nonresponse ein Wert – die Ausschöpfungsquote – berechnet werden kann, für Item Nonresponse zwar Messkonzepte vorhanden sind (Bankhofer 1995), diese aber in der Praxis kaum angewendet werden, verstärkt sich diese Fixierung. Es interessiert, wenn überhaupt, nur die Item Nonresponse der für die konkrete Analyse relevanten Variablen. Als zweites Problem tritt die Vergleichbarkeit über die Zeit hervor (Stoop 2010, S.2), da – wie sich auch beim ALLBUS zeigen wird – immer wieder Elemente des Erhebungsprozesses geändert werden. Schließlich – drittens – wird erst durch eine genaue Analyse nach den Ausfallgründen (und deren Definition) eine Abschätzung möglich, wie die Entwicklung der Ausschöpfung tatsächlich verläuft (Stoop 2010, S.2).
Wenn in der Literatur immer wieder beklagt wird, dass es keine verbindlichen Standards für Berechnungen gibt, ist das sicherlich richtig (Schnell 1997; Smith 2002). Eine der ersten Definitionen der Nonresponserate stellte 1980 der Council of American Survey Research Organisations (CASRO) auf.[110] Die Responsequote wird hier als Quotient von realisierten Interviews (mit einer genauen Definition) geteilt durch die Anzahl der in der Grundgesamtheit befindlichen Personen errechnet. Zu Recht weisen Schnell (1997) und Wiseman et al. (1984) auf die Umsetzungsprobleme hin, die diese scheinbar einfachst mögliche Definition mit sich bringt. Anhand zweier Beispielrechnungen lässt sich zeigen, dass bei der häufig verwendeten Vereinbarung fester Interviewzahlen, wie sie in der Marktforschung üblich, jedoch auch der Sozialforschung nicht fremd ist, erhebliche Probleme auftreten können (Wiseman at al. 1984, S.337).[111] Aufgrund nicht dokumentierter Ersetzungen von Ausfällen wird eine Quotenberechnung jeglicher Art gar nicht mehr möglich (Schnell 1997, S.20, Fußnote 29). Ein weiterer Kritikpunkt ist die Berechnung der tatsächlichen Grundgesamtheit, die berücksichtigt, dass nicht alle Fälle von Unit Nonresponse echt sind (Wiseman et al. 1984, S.337).[112]
Wenn man bereits Probleme hat, sich auf eine Definition zu einigen, und selbst allgemein anerkannte Definitionen unter massiven Umsetzungsproblemen leiden, dann erscheinen Bemühungen, wenigstens gewisse Standards in der Dokumentation zu berücksichtigen, sinnvoll. Maas und de Heer (1995) nehmen in ihren Fragebogen zu Nonresponse Kriterien auf, die eine große Bandbreite an Informationen über den gesamten Erhebungsprozess abdecken. Dabei sollten möglichst weitgehende Informationen über das Sampling Design, über die Durchführung der Erhebung – speziell die Feldarbeit – , über die Interviewer und über das Antwortverhalten zur Verfügung stehen. Diese Informationen ließen wenigstens eine Vergleichbarkeit zu und lieferten nachvollziehbare Gründe, warum sogar in den amtlichen Erhebungen international so große Unterschiede bestehen (De Heer 1999, S.140). Damit wird auch die Vergleichbarkeit größer und ein Trend in der Entwicklung von Nonresponse kann zuverlässiger abgeschätzt werden.
Für die hier verwendete Erhebung – den ALLBUS – steht eine gute Dokumentation des Erhebungsablaufs zur Verfügung, die es auch zulässt, Trends abzulesen (vgl. Abbildung 53).

[110] http://www.casro.org/index.cfm, Council of American Survey Research Organizations 1982.
[111] In Deutschland besonders bei Telefoninterviews, vgl. Schnell (1997), S.19.
[112] Hierzu die Abschnitt 5.1.1 und 5.1.3.

### 5.2.2 Ausschöpfungsquote beim ALLBUS

Die Feldarbeit einer ALLBUS-Erhebung wird mittlerweile sehr gut dokumentiert und lässt Schlüsse auf das genaue Zustandekommen der Ausschöpfungsquote zu (vgl. GESIS-TR 2010/04, S.52f). Bereits durch die Ankündigungsschreiben kann festgestellt werden, welche Adressen neutral ausfallen (indem das Anschreiben in der Regel mit Angabe von Gründen zurückgesandt wurde). Für diese Fälle existieren Ersatzadressen als Aufstockungsstichprobe. Ziel ist eine vordefinierte Mindestanzahl von realisierten Interviews. Die Interviewer sind während der Erhebungszeit genau instruiert, wie (persönlicher Erstkontakt) und wann (unterschiedliche Tageszeiten und Wochentage) der Kontakt hergestellt werden sollte. Zur Kontrolle müssen die Interviewer ein detailliertes Kontaktprotokoll führen. Von hohem Informationswert ist neben den dort notierten Zeiten die mögliche Ausfallursache (z.T. offene Abfrage). Eindeutig wird beispielsweise definiert, dass erst nach vier erfolglosen Kontaktierungsversuchen (an verschiedenen Tagen und zu verschiedenen Uhrzeiten) eine Zielperson als nichterreichbar gilt. Eben beschriebener Vorgang wird als Hauptbearbeitung bezeichnet, die als Kalkulationsgrundlage für eine mögliche Aufstockungsstichprobe herangezogen wird.[113] Die sogenannte Aufstockungsstichprobe durchläuft dann unter gleichen Bedingungen oben beschriebenen Prozess. Fasst man alles nun allgemein für die Berechnungsformel zusammen, so ergibt sich folgende Rechnung:

| | |
|---|---|
| ursprüngliche Bruttostichprobe<br>+ zusätzliche Adressen für neutrale Ausfälle | = Bruttostichprobe |
| − stichprobenneutrale Ausfälle | = bereinigter Stichprobensatz |
| − systematische Ausfälle | = auswertbare Interviews |

Tabelle 33: Berechnung der Ausschöpfungsquote bei ALLBUS-Erhebungen

Zu den stichprobenneutralen Ausfällen zählen: Anschreiben nicht zustellbar, Adresse falsch oder nicht mehr existent, Zielperson verstorben oder verzogen, Zielperson lebt nicht in diesem Privathaushalt (GESIS-TR 2010/04, S.56). Diese Fälle als *neutral* zu bezeichnen, erscheint prinzipiell zulässig, solange man von einer sauberen Dokumentation ausgehen kann. Als systematische Ausfälle gelten: im Haushalt niemanden angetroffen (Nichterreichbarkeit), Zielperson nicht angetroffen (Nichterreichbarkeit), Zielperson nicht befragungsfähig (Unfähigkeit), Zielperson spricht nicht hinreichend gut deutsch (Unfähigkeit), Zielperson aus Zeitgründen nicht bereit (Verweigerung), Zielperson generell nicht zum Interview bereit (Verweigerung) und sonstige Gründe (Bear-

[113] Dabei geht es um die Erfüllung des Interviewersolls, das in der Planung vorab definiert wird.

beitungsfehler, Interviewfälschungen etc.). Auch diese Einteilung erscheint sinnvoll und praktikabel, wenn man von einer sorgfältigen Bearbeitung ausgeht. Die Antwortrate wird beim ALLBUS als Ausschöpfungsquote bezeichnet und als Quotient der oben ausgeführten Größen berechnet:

$$\frac{\text{auswertbare Interviews}}{\text{bereinigter Stichprobensatz}}$$

Dieser Quotient wird auch in der folgenden Analyse der Ausschöpfungsquote aller ALLBUS-Erhebungen seit 1980 verwendet.

#### 5.2.2.1 Veränderung Ausschöpfungsquote beim ALLBUS

Zunächst wird die Ausschöpfungsquote (als Anteil an der bereinigten Bruttoausschöpfung) betrachtet. Die blaue Linie trägt die Werte für Westdeutschland ab 1980 ab, die rote Linie die Werte für Ostdeutschland ab 1991.[114]

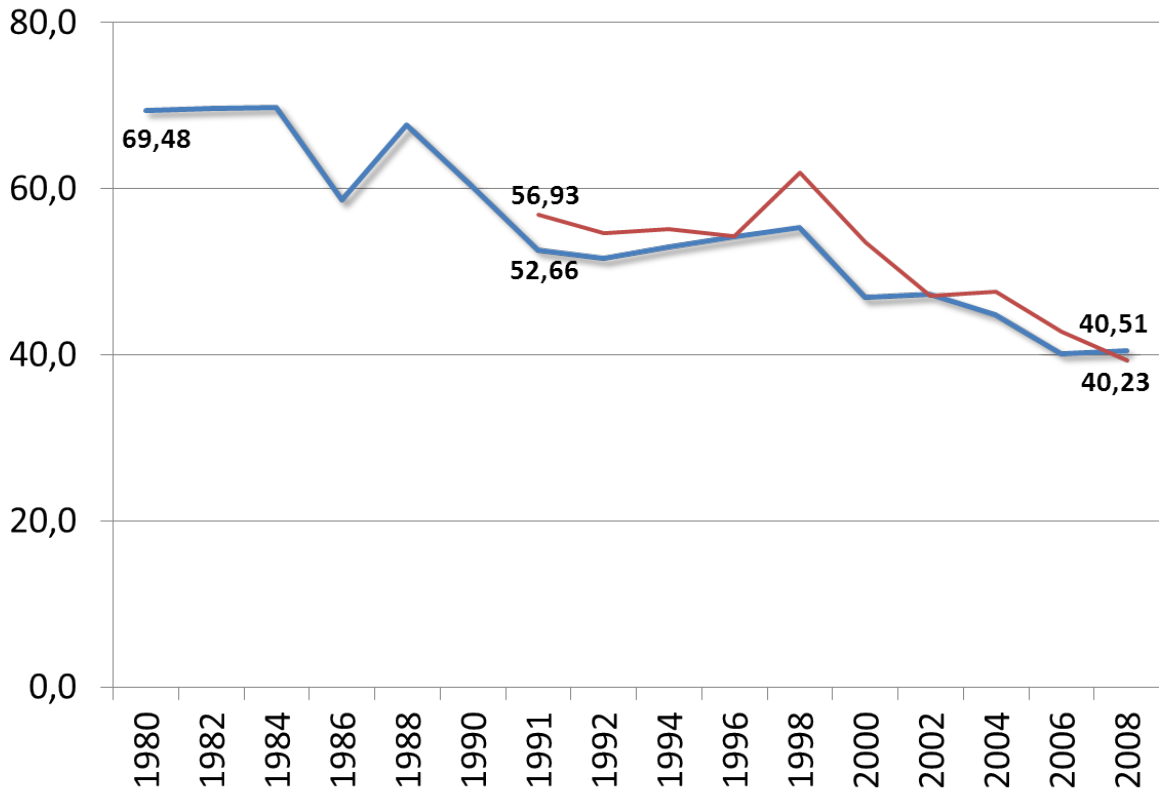



Abbildung 53: Veränderung der Ausschöpfungsquote des ALLBUS 1980 - 2008

Auf dem ersten Blick erscheint in Abbildung 53 die Entwicklung der Ausschöpfungsquote eindeutig. Von 1980 bis 2008 fällt sie in Westdeutschland um 28,97 %-Punkte, in Ostdeutschland von 1991 bis 2008 um immerhin 17,14 %-Punkte, wobei das Ausgangsniveau mit gut 56 % relativ niedrig ist (zum Vergleich: Westdeutschland 1980 fast 69,48 %). Bei näherer Betrachtung stechen einige Werte jedoch ins Auge – und zwar die Werte für 1988 West und 1998 Ost. Diese ALLBUS-Erhebungen haben im Vergleich zu jeweils vorhergehenden Erhebungen eine bedeutend höhere Ausschöpfung (9,02 bzw. 7,78 %-Punkte). Dagegen sacken die Werte für den ALLBUS 1990 West sowie den ALLBUS 2000 in beiden Landesteilen ab (7,32 bzw. 8,46 und 8,45 %-Punkte). Im Kontext der gesamten Zeitreihe scheinen die Ausreißer nach oben aber nur den Gesamttrend kurzzeitig

[114] Zum Zeitpunkt der Analyse lag der exakte Wert für die Ausschöpfungsquote 2010 noch nicht vor; vorläufig wurde hier mit etwa 35 % Rücklauf für Westdeutschland und etwa einem Drittel für Ostdeutschland gerechnet. Dies würde die in Abbildung 53 dargestellten Verlauf nach unten noch deutlicher in eine Abwärtskurve drücken.

zu konterkarieren, denn die folgenden Werte schließen an die Vorausreißerwerte in etwa an. Somit ergibt sich ein doch linearer Abwärtstrend in Ost- und Westdeutschland für den ALLBUS. Es lässt sich nun aber weiter fragen, wie es zu Besonderheiten kommt und welche Faktoren den Abwärtstrend beschleunigt oder verlangsamt haben.

#### 5.2.2.2 Analyse der Ausschöpfungsquote

Wohl wenige Erhebungen sind so langfristig und detailliert geplant wie der ALLBUS. Zeugnis davon geben die immer detaillierteren Methodenreporte des ALLBUS (zum Vergleich: Umfang des Methodenreports 1980: 43 Seiten und 2008: 81 Seiten). Hier sind auch Änderungen in der Stichprobenziehung oder Erhebungsmethode verzeichnet und Konsequenzen diskutiert, die teilweise als Erklärung für außergewöhnliche Ausschöpfungsquoten dienen könnten.
Tabelle 34 ergänzt die Werte aus Abbildung 53 um die Feldzeit in Tagen und um den Namen des ausführenden Ehrebungsinstituts.

| Jahr | West | | Ost | | |
|---|---|---|---|---|---|
| | Quote | Feldzt. | Quote | Feldzt. | Erhebungsinst. |
| 1980 | 69,48 | 62 | - | - | GETAS |
| 1982 | 69,70 | 117 | - | - | GETAS |
| 1984 | 69,89 | 105 | - | - | GETAS |
| 1986 | 58,67 | 53 | - | - | Infratest |
| 1988 | 67,69 | 68 | - | - | GFM-GETAS |
| 1990 | 60,37 | 90 | - | - | Infas |
| 1991 | 52,66 | 48 | 56,93 | 55 | Infratest |
| 1992 | 51,61 | 48 | 54,67 | 33 | Infratest |
| 1994 | 53,02 | 105 | 55,21 | 76 | Infratest |
| 1996 | 54,22 | 124 | 54,23 | 119 | Infratest |
| 1998 | 55,38 | 140 | 62,01 | 114 | GFM-GETAS |
| 2000 | 46,92 | 196 | 53,56 | 195 | Infratest |
| 2002 | 47,33 | 179 | 47,15 | 179 | Infas |
| 2004 | 44,89 | 134 | 47,58 | 134 | Infratest |
| 2006 | 40,23 | 157 | 42,82 | 157 | TNS-Infratest |
| 2008 | 40,51 | 176 | 39,79 | 176 | TNS-Infratest |

Tabelle 34: Ausschöpfungsquote, Feldzeit und Erhebungsinstitut nach Landesteil

Jedoch haben sich auch andere Faktoren des Survey Lifecycle für den ALLBUS geändert: **Zielgesamtheit/Grundgesamtheit**, **Stichprobenziehung** (Sample Frame) und **Erhebungsmodus**.
1990 wurde die Grundgesamtheit ausgeweitet. Befragt werden sollen seitdem nicht nur alle erwachsenen Personen mit deutscher Staatsangehörigkeit, sondern alle erwachsenen Personen – auch

ausländische Personen – in Privathaushalten in der Bundesrepublik Deutschland. Einzige Voraussetzung ist eine ausreichende Sprachkenntnis in der Erhebungssprache Deutsch; dies wird vom jeweiligen Interviewer vor Ort beurteilt. Diese Umstellung wird im Methodenreport 1991 für das Absacken der Ausschöpfungsquote verantwortlich gemacht (Bandilla et al. 1992). Auch beim Auswahlverfahren wurde mehrmals gewechselt.
Zwischen 1980 bis 1992 sowie beim ALLBUS 1998 wurde das ADM-Design mit dreistufiger Auswahl[115] verwendet, 1994, 1996 sowie ab 2000 wurde mit Stichproben aus den Einwohnermelderegistern gearbeitet. Die Stichprobenziehungen waren hierbei nur zweistufig.[116] Der Anstieg 1998 in Ostdeutschland und das Absacken in beiden Landesteilen 2000 wird mit dieser Umstellung begründet (Koch 1997; Koch et al. 2001, S.57f; Koch et al. 1999, S.42). Koch nennt die Ausschöpfungsquoten für ADM-Stichproben „weniger zuverlässig"(Koch 1997, S.111). Bei ADM-Stichproben sei es nämlich wesentlich einfacher, statt der eigentlichen Zielperson eine andere, der Befragung offen gegenüberstehende Person zu befragen. Hingegen können bei Einwohnermelderegisterstichproben Interviewer einfacher kontrolliert werden, da die Adressen ja im Vorfeld gezogen werden und bekannt sind (Albers 1997, S.123). Ob die Verwendung einer ADM-Stichprobe tatsächlich die vergleichsweise hohen Ausschöpfungsquoten erklärt, ist nicht ganz eindeutig, da zum einen gerade in Ostdeutschland mit der auffällig hohen Quote laut dem Erhebungsinstitut kontrolliert wurde, für Westdeutschland seltsamerweise aber überhaupt keine Informationen über Kontrollen vorliegen. In Westdeutschland war die Quote jedoch unauffällig (Koch et al. 1998, S.45f).
Für den ALLBUS 2000 wurde erstmals CAPI verwendet. Die Umstellung wurde wissenschaftlich genau beobachtet und ausgewertet. Empirisch gibt es keine Hinweise, dass diese Umstellung einen Einfluss auf die Ausschöpfungsquote hatte; anders als dies bei Item Nonresponse der Fall ist (Wasmer und Koch 2002).
Aus der Tabelle 34 lässt sich mit Hilfe der zusätzlichen Informationen bereits erkennen, dass neben dem Faktor Zeit auch andere Faktoren eine Rolle spielen. Zuerst fällt auf, dass ein Institut wesentlich höhere Ausschöpfungsquoten hervorbringt als alle anderen. Die Länge der Feldzeit variiert stark und zeigt oberflächlich gesehen keinen eindeutigen Trend. Im Weiteren wird mit diesen Variablen ein OLS-Modell berechnet, das neben der Zeit auch noch andere Regressoren umfasst. Insgesamt soll geklärt werden, inwieweit das Institut (als Dichotome Variable mit 1 = GETAS), der Landesteil (mit Ost=1), die Länge der Feldzeit und schließlich die Art der Stichprobenumsetzung eine Rolle spielen. Bei der Stichprobenziehung lassen sich zwei Verfahren unterscheiden: eine Registergestützte Stichprobenziehung und eine ADM-Stichprobe (mit 1 = Einwohnermelderegister). Die Variablen werden jeweils schrittweise eingefügt, da insgesamt nur 26 Beobachtungen (ALLBUS in Ost oder Westdeutschland) vorliegen. Das Modell 5 lautet dann:

$$quote_t = \beta_1 + \beta_2 zeit_t + \beta_3 landesteil_t + \beta_i institut_t + \beta_5 stichprobe_t + \beta_6 feldzeit_t + \varepsilon_t$$

mit $t = 1, \ldots, 26$

[115] Wobei die erste Stufe den Bezirk auswählt (mit Probabilities Proportional to Size), die zweite Stufe bestimmt die Adresse (in der Regel mit Random Route Methoden) und die dritte Stufe wählt die Zielperson im Haushalt aus (in der Regel durch „Schwedenschlüssel").

[116] Die erste Stufe besteht in der zufälligen Auswahl von Gemeinden. In der zweiten Stufe werden aus diesen Gemeinden zufällig Adressen gezogen.

Die Ergebnisse lauten:

| | Modell 1 | Modell 2 | Modell 3 | Modell 4 | Modell 5 |
|---|---|---|---|---|---|
| Konstante | $2105{,}60^{***}$ | $2176{,}54^{***}$ | $1755{,}35^{***}$ | $1883{,}64^{***}$ | $1976{,}76^{***}$ |
| Zeit | $-1{,}03^{***}$ | $-1{,}07^{***}$ | $-0{,}86^{***}$ | $-0{,}92^{***}$ | $-0{,}97^{***}$ |
| Landesteil | - | $1{,}83^{ns}$ | $2{,}40^{*}$ | $2{,}42^{*}$ | $2{,}59^{*}$ |
| Institutseffekt | - | - | $7{,}92^{***}$ | $8{,}44^{***}$ | $7{,}73^{***}$ |
| Stichprobenziehung | - | - | - | $1{,}64^{ns}$ | $0{,}81^{ns}$ |
| Feldzeit | - | - | - | - | $0{,}01^{ns}$ |
| korrigiertes $R^2$ | 0,826 | 0,828 | 0,936 | 0,937 | 0,936 |
| n=26 | | | | | |

Tabelle 35: Einflüsse auf die Entwicklung der Ausschöpfungsquote (Beta-Koeffizienten)

Geht man nach den Koeffizienten und der Erklärungskraft der einzelnen Modelle, ist die Zeit immer noch der wichtigste, stets signifikante Einflussfaktor. Mit Abstand kommt an zweiter Stelle der Institutseffekt, der ebenfalls hochsignifikant ist und auf hohem Niveau zusätzliche Erklärungskraft für das Modell liefert. Die Stichprobenziehung an sich hat dann keinen signifikanten Einfluss mehr; ebenso wenig die Feldzeit.
Zusammengefasst: Zwei langfristige Faktoren können zur Erklärung der sinkenden Ausschöpfungsquoten beim ALLBUS beitragen. Der eine Faktor lässt sich sehr gut mit gerade berechnetem Modell bestätigen: das Erhebungsinstitut. In Westdeutschland zeichnet das Forschungsinstitut GETAS für die fünf höchsten Ausschöpfungsquoten (und für die mit Abstand höchste Quote in Ostdeutschland) verantwortlich. Dies hat sich auch in den Koeffizienten niedergeschlagen. Der zweite Faktor kann mit diesen Informationen noch nicht empirisch bestätigt werden: Die Grundgesamtheit wurde durch eine Personengruppe, die sich durch mangelnde Sprachfähigkeit (damit Befragungsunfähigkeit) und eine allgemein höhere Verweigerungshaltung auszeichnet, vergrößert. Trotz guter Dokumentation bleibt die Frage, wie Institutseffekte genau die Ausschöpfungsquote beeinflussen (Schnell 1997). Faktoren, die sich auf das allgemeine Befragungsklima auswirken, wurden nicht untersucht; was nicht bedeutet, dass sie sich nicht ausgewirkt haben könnten. Beispielsweise könnte die Ausschöpfung des ALLBUS 1986 im Zuge des äußerst umstrittenen Zensus 1987 gelitten haben.
Um der Frage nach der Veränderung der Ausfallgründe nachzugehen, wird in Tabelle 36 der Verlauf der drei Ausfallgründe abgetragen:

| Jahr | West | | | Ost | | |
|---|---|---|---|---|---|---|
| | Verwg. | Nichterb. | Unfähigkt. | Verwg. | Nichterb. | Unfähigkt. |
| 1980 | 16,2 | 10,5 | 1,1 | – | – | – |
| 1982 | 18,1 | 9,4 | 1,2 | – | – | – |
| 1984 | 18,2 | 8,5 | 1,1 | – | – | – |
| 1986 | 25,7 | 12,2 | 3,3 | – | – | – |
| 1988 | 13,8 | 11,9 | 0,9 | – | – | – |
| 1990 | 19,5 | 13,9 | 2,3 | – | – | – |
| 1991 | 24,9 | 20,3 | 2,1 | 24,7 | 16,2 | 2,2 |
| 1992 | 26,5 | 18,3 | 3,3 | 25,3 | 18,4 | 1,6 |
| 1994 | 37,8 | 2,8 | 5,0 | 35,8 | 3,4 | 4,6 |
| 1996 | 35,6 | 4,7 | 4,7 | 38,3 | 3,4 | 3,8 |
| 1998 | 30,0 | 8,3 | 2,6 | 29,1 | 5,9 | 2,7 |
| 2000 | 40,6 | 7,4 | 3,0 | 39,8 | 2,4 | 1,9 |
| 2002 | 30,9 | 4,3 | 6,2 | 32,9 | 5,9 | 3,7 |
| 2004 | 42,3 | 7,1 | 5,0 | 42,5 | 6,0 | 3,6 |
| 2006 | 46,5 | 6,6 | 5,0 | 46,5 | 5,9 | 3,7 |
| 2008 | 47,9 | 6,4 | 4,3 | 48,9 | 7,2 | 3,1 |

Tabelle 36: Veränderung des Anteils der Verweigerer, Nichterreichten und Befragungsunfähigen nach Landesteilen

Zwar wurden schon für einige hervorstechende Ausschöpfungsquoten in den Methodenreporten auffällige Verschiebungen in den Gründen als ursächlich genannt, analysiert man die systematischen Ausfälle genauer, treten jedoch weitere Fragen und Probleme hervor. Prinzipiell fällt bei der Dokumentation auf, dass die Kategorien für die Ausfallgründe verschieden gebildet werden.[117] Es lässt sich nicht genau sagen, ob die unterschiedlichen Einteilungen institutsbedingt sind oder von der unterschiedlichen Autorenschaft der Methodenreporte herrühren, oder beides. Daraus folgt jedenfalls eine nur bedingte Vergleichbarkeit der einzelnen Ausfallgründe, die oben nach den in der Literatur hauptsächlich verwendeten Ausfallgründen für Ost- und Westdeutschland dargestellt sind (Verweigerung, Nichterreichbarkeit, Unfähigkeit).
Trotz der Brüche in der Dokumentation lassen sich folgende drei Feststellungen treffen: Erstens bilden die Verweigerer immer die größte Gruppe und diese Gruppe wird tendenziell gewichtiger und gewinnt spätestens seit den 2000ern mit immer größeren Abstand zu den anderen Ausfallgründen an Bedeutung. Zweitens lässt sich bei dieser genauen Beobachtung über die Zeit erkennen, dass der Anteil der Nichterreichbaren ab 1992 stabil erscheint, wie dies die Verweigerungsrate ceteris paribus tut. Der wesentlich interessantere Aspekt ist das Verhältnis zwischen der Verweigerungsgruppe und der Gruppe der Nichterreichbaren, denn dies legt die Vermutung nahe, dass

[117] Alle Änderungen sind in Anhang 8 beschrieben, einzige durchgehende Kategorie ist „im HH niemanden angetroffen“.

eventuell interviewerbedingt (und damit wohl institutsbedingt) eine Verschiebung in der Angabe der Gründe geschieht. Ohne tieferen Einblick in die Feldarbeit bleibt dies jedoch nur Vermutung. Drittens scheint die Gruppe derjenigen, die nicht die sprachliche, körperliche oder geistige Fähigkeit zur Teilnahme an der Befragung besitzen, stabil zu sein, wenn man den leichten Anstieg, den die Vergrößerung der Grundgesamtheit (auch Ausländer bei entsprechenden Sprachkenntnissen) ab 1990 mit sich gebracht hat, ausnimmt.
Die tieferen Einblicke könnte aber eine Differenzierung nach den Ausfallgründen liefern. Deshalb werden – wiederum schrittweise – die Modelle als OLS-Regression für Befragungsunfähige, Nichterreichbare und Verweigerer berechnet, die die ersten Ergebnisse aus der Zeitreihe (Tabelle 34) vertiefen.

Befragungsunfähige:

$$unfaehig_t = \beta_1 + \beta_2 zeit_t + \beta_3 landesteil_t + \beta_i institut_t + \beta_5 stichprobe_t + \beta_6 feldzeit_t + \varepsilon_t$$

mit $t = 1, \ldots, 26$

Die Ergebnisse lauten:

| | Modell 1 | Modell 2 | Modell 3 | Modell 4 | Modell 5 |
|---|---|---|---|---|---|
| Konstante | $-225{,}60^{***}$ | $-256{,}48^{***}$ | $-184{,}98^{**}$ | $-81{,}23^{ns}$ | $174{,}98^{*}$ |
| Zeit | $0{,}12^{***}$ | $1{,}31^{***}$ | $0{,}95^{**}$ | $-0{,}92^{ns}$ | $0{,}89^{*}$ |
| Landesteil | – | $-0{,}79^{ns}$ | $-0{,}89^{*}$ | $-0{,}87^{*}$ | $1{,}03^{**}$ |
| Institutseffekt | – | – | $1{,}35^{*}$ | $0{,}93^{ns}$ | $-0{,}21^{ns}$ |
| Stichprobenzhg. | – | – | – | $1{,}33^{*}$ | $2{,}17^{**}$ |
| Feldzeit | – | – | – | – | $0{,}01^{*}$ |
| korrigiertes $R^2$ | 0,390 | 0,437 | 0,545 | 0,620 | 0,676 |
| n=26 | | | | | |

Tabelle 37: Einflüsse auf die Entwicklung des Verweigerungsanteils (Beta-Koeffizienten)

Aufgrund der geringen Fallzahl und des geringen Anteils an Befragungsunfähigen sind die Modelle nicht sonderlich stabil. Mit der Hinzunahme zusätzlicher Variablen verliert der Faktor Zeit an Bedeutung oder ist nur noch schwach signifikant. Schließlich lässt sich nur noch für den Landesteil und die Art der Stichprobenziehung ein Einfluss auf die Höhe des Anteils an Befragungsunfähigen konstatieren. Der Landesteileffekt wird durch die unterschiedliche Höhe des Ausländeranteils plausibel, während sich der signifikante positive Einfluss der Stichprobenziehung auf den Anteil der Verweigerer durch die wesentlich strengere Restriktion der Zielperson bei der Registerauswahl begründen lässt. Bei den drei Arten der ADM-Stichprobe bleibt ja dem Interviewer eine wesentlich größere Freiheit, die Zielperson, die befragungsunfähig ist, durch eine andere zu ersetzen.

Damit bestätigen sich die auch in den Methodenreporten erwähnten Effekte auf den Anteil der Nichtbefragungsfähigen.

Nichterreichbare:

$$nichterb_t = \beta_1 + \beta_2 zeit_t + \beta_3 landesteil_t + \beta_i institut_t + \beta_5 stichprobe_t + \beta_6 feldzeit_t + \varepsilon_t$$

mit $t = 1, \ldots, 26$

Die Ergebnisse lauten:

| | Modell 1 | Modell 2 | Modell 3 | Modell 4 | Modell 5 |
|---|---|---|---|---|---|
| Konstante | $611{,}35^{*}$ | $590{,}86^{*}$ | $779{,}04^{**}$ | $-185{,}66^{ns}$ | $-311{,}55^{*}$ |
| Zeit | $-0{,}30^{*}$ | $-0{,}29^{*}$ | $-0{,}39^{**}$ | $0{,}11^{ns}$ | $0{,}17^{ns}$ |
| Landesteil | – | $-0{,}52^{ns}$ | $-0{,}78^{ns}$ | $-0{,}98^{ns}$ | $-1{,}20^{ns}$ |
| Institutseffekt | – | – | $-3{,}54^{ns}$ | $-7{,}41^{***}$ | $-6{,}45^{***}$ |
| Stichprobenzhg. | – | – | – | $-12{,}34^{***}$ | $-11{,}21^{***}$ |
| Feldzeit | – | – | – | – | $0{,}19^{ns}$ |
| korrigiertes $R^2$ | 0,229 | 0,164 | 0,200 | 0,823 | 0,825 |
| n=26 | | | | | |

Tabelle 38: Einflüsse auf die Entwicklung des Nichterreichbarenanteils (Beta-Koeffizienten)

Die Modelle für die Nichterreichbaren sind etwas stabiler und zeigen ebenfalls einen Bedeutungsverlust der Zeit bei Hinzunahme weiterer Determinanten. Der größte Sprung ereignet sich zwischen Modell 3 und 4, wenn die Stichprobenziehungsart eingefügt wird. Es gilt das bereits oben Genannte: Nichterreichbare können bei ADM-Stichproben aufgrund des eigenmächtigen Handelns des Interviewers durch andere Befragte ersetzt werden. Bei Registerstichproben erhöht sich deshalb auch die Quote der Nichterreichbaren. Das Institut hat darüber hinaus noch einen eigenständigen, signifikanten Einfluss auf die Höhe der Nichterreichtenrate.

Verweigerer:

$$verweiger_t = \beta_1 + \beta_2 zeit_t + \beta_3 landesteil_t + \beta_i institut_t + \beta_5 stichprobe_t + \beta_6 feldzeit_t + \varepsilon_t$$

mit $t = 1, \ldots, 26$

Die Ergebnisse lauten:

| | Modell 1 | Modell 2 | Modell 3 | Modell 4 | Modell 5 |
|---|---|---|---|---|---|
| Konstante | $-2322{,}83^{***}$ | $-2301{,}74^{***}$ | $-2067{,}64^{**}$ | $-1465{,}97^{***}$ | $-1617{,}39^{***}$ |
| Zeit | $1{,}18^{***}$ | $1{,}17^{***}$ | $1{,}05^{***}$ | $0{,}75^{***}$ | $0{,}82^{***}$ |
| Landesteil | – | $0{,}54^{ns}$ | $0{,}22^{ns}$ | $0{,}34^{ns}$ | $0{,}08^{ns}$ |
| Institutseffekt | – | – | $-4{,}40^{ns}$ | $-1{,}91^{ns}$ | $-0{,}83^{ns}$ |
| Stichprobenzhg. | – | – | – | $-7{,}70^{**}$ | $-9{,}05^{**}$ |
| Feldzeit | – | – | – | – | $-0{,}02^{ns}$ |
| korrigiertes $R^2$ | 0,805 | 0,815 | 0,866 | 0,823 | 0,863 |
| n=26 | | | | | |

Tabelle 39: Einflüsse auf die Entwicklung des Befragungsunfähigenanteils (Beta-Koeffizienten)

Anders als bei Befragungsunfähigen und Nichterreichbaren spielt die Zeit bei den Verweigerern auch unter Kontrolle mit anderen Variablen eine signifkante Rolle. Der Effekt schwächt sich aber leicht ab. Alle zwei Jahre nimmt am Ende die Rate der Verweigerer um durchschnittlich etwa 0,8 % zu. Einzig noch signifikanten Einfluss hat wiederum die Stichprobenziehung. Hier ist der Effekt ebenfalls deutlich und lässt sich mit denselben Mechanismen zu erklären wie bei der Rate der Befragungsunfähigen und Nichterreichten.

### 5.2.3 Zusammenfassung

Im Vergleich zu den Ergebnissen, die Schnell (1997) präsentiert, scheint eher nach 1990 und verstärkt ab den 2000ern generell die Ausschöpfungsquote beim ALLBUS zu sinken. Dies steht keinesfalls im Widerspruch zu Schnells Befunden, sondern setzt als Entwicklung erst nach dessen Untersuchungszeitraum ein (in der Regel bis 1990). Bezieht man in die Analyse den möglichen Einfluss der Institute und die qualitative wie quantitative Dokumentation der ALLBUS-Erhebungen mit ein, so gehen die Ergebnisse bezüglich der Nichterreichbaren und Befragungsunfähigen mit dem von Schnell Konstatierten konform: hier lässt sich zeitlich kaum eine Veränderung feststellen, jedoch spielt die Stichprobenziehung eine entscheidende Rolle. Anders als bei

Schnell kann bei den Verweigerern davon ausgegangen werden, dass ihre Zahl auch unter Kontrolle des Institutseffekts und der Stichprobenziehung über die Zeit hin steigt. Selbst bei immer detaillierteren und besser dokumentierten Erhebungen stört für eine Analyse über die Zeit die inkonsistente Kategorisierung.[118]
Es lässt sich aber abschließend konstatieren, dass mit Blick auf die fast zwanzigjährige Geschichte des ALLBUS mit dem ALLBUS 2008, der als Datengrundlage später verwendet wird, ein vorläufiger Tiefstand der Ausschöpfungsquote für Gesamtdeutschland (40,29 %) erreicht wurde.[119]

## 5.3 Praxis: ALLBUS 2008

Für die Untersuchung potentieller Korrekturmöglichkeiten von Unit Nonresponse bedarf es spezieller Daten, die normalerweise nur auf Anfrage für den Anwender zur Verfügung gestellt werden. Der Anwender hat in der Regel nur mit der Nettostichprobe zu tun und Zugang zu Informationen über Unit Nonresponse in Form einer Ausschöpfungsquote und von dem, was in offiziellen Dokumentationen veröffentlicht ist. Für diese Untersuchung stellte GESIS einen Pre_release des Auszugs des Bruttobandes 2008 zur Verfügung. Die Version beinhaltet sowohl alle neutralen als auch alle systematischen Ausfallgründe. Bei der nun folgenden Analyse wird zunächst die Zusammensetzung der Ausfälle für den ALLBUS 2008 Pre_release und die Definition, auf der die Zählung basiert, dargestellt. Anschließend wird die Ausschöpfungsquote noch einmal berechnet. Genau wie bei der Analyse der Item Nonresponse, wird für Unit Nonresponse ein Erklärungsmodell geschätzt. Hierbei erscheint es am geschicktesten, nach Ausfallgrund getrennte Modelle zu berechnen. Die Determinanten wurden zum großen Teil bereits im Theorieteil diskutiert und werden nur noch in ihrer Operationalisierung erläutert.

### 5.3.1. Ausfallgründe und Ausschöpfungsquote

In der ALLBUS-Dokumentation wird konsistent nach ursprünglicher Bruttostichprobe, Bruttostichprobe und bereinigtem Stichprobensatz unterschieden. Die linke Darstellung von Abbildung 54 unterscheidet nach dem Anteil der realisierten Interviews, der *systematischen* und der *neutralen* Ausfälle. Als *neutrale* Ausfälle gelten für diese Erhebung:

- Anschreiben nicht zustellbar,
- Adresse falsch, existiert nicht (mehr),

[118] Dabei mangelt es mittlerweile nicht mehr an Sensibilität für das Problem, vgl. Menold und Züll (2010).
[119] Die vorläufige Schätzung für den ALLBUS 2010 lässt sogar noch befürchten, dass sich der Trend beschleunigt.

- Zielperson verstorben,

- Zielperson verzogen,

- Zielperson lebt nicht in Privathaushalten.

Insgesamt machen diese Ausfälle gut 10 % der Ausfälle aus (West: 13,02 %; Ost: 11,16 %). Da sie als *neutral* definiert werden, jedoch gleichzeitig eine Mindestzahl an Interviews durchgeführt werden soll, werden für diese Ausfälle während der Feldphase zusätzliche Adressen ausgeben (dies wird in der Berechnung der Ausschöpfungsquote ausgewiesen). Ob alle Ausfälle in diesen Kategorien tatsächlich *neutral* im Sinn von zufällig sind, erscheint als sehr starke Annahme.
Ein weiterer Teil der linken Säule visualisiert den Anteil der als *systematisch* bezeichneten Ausfälle. Folgende Kategorien bilden *systematische* Ausfälle:

- Im Haushalt niemanden angetroffen,

- Zielperson nicht angetroffen,

- Zielperson nicht befragungsfähig,

- Zielperson aus Zeitgründen nicht zum Interview bereit,

- Zielperson generell nicht zum Interview bereit,

- Zielperson spricht nicht hinreichend gut deutsch,

- Adresse nicht abschließend bearbeitet,

- Interviews als (Teil-)Fälschung identifiziert.

Der Block der systematischen Ausfälle umfasst etwa 53 % der Bruttostichprobe, wobei sich nach Ost- und Westdeutschland nahezu keine Unterschiede ergeben. Einzige Ausnahme stellt die Kategorie des Ausfalls wegen mangelnder Sprachfähigkeit dar. Hier liegt der Anteil in den alten Bundesländern bei 2,1 %, in Ostdeutschland hingegen nur bei 0,6 %. Der Unterschied ergibt sich plausibel aus der Bevölkerungsstruktur. In dunkelrot ist abschließend der Anteil tatsächlich realisierter Interviews gefärbt. Dieser beträgt für Gesamtdeutschland knapp 36 %.

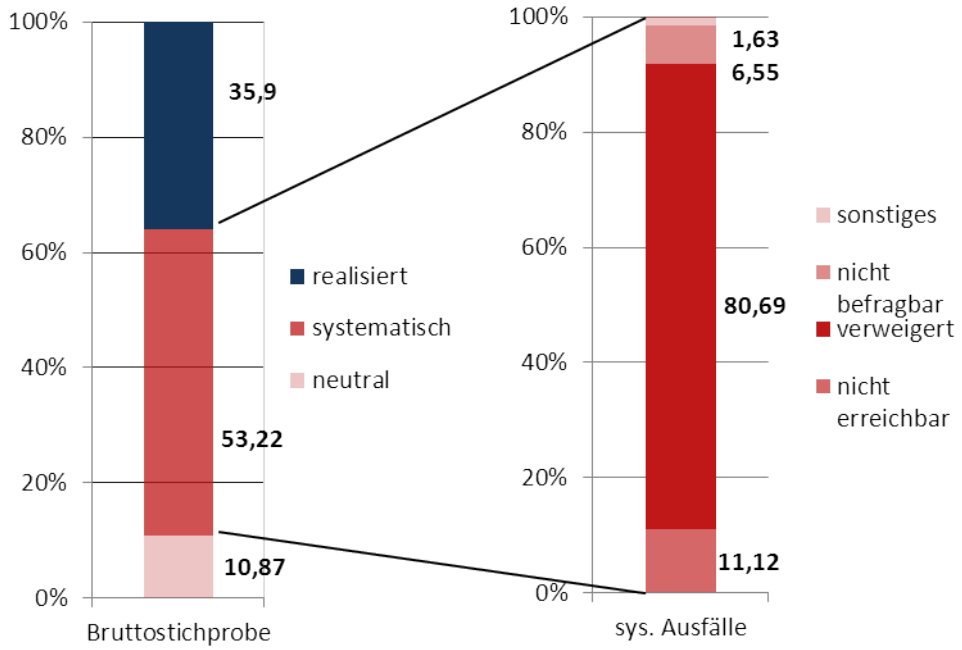



Abbildung 54: Ausfallgründe im ALLBUS 2008

Betrachtet man die systematischen Ausfälle genauer (Abbildung 54 rechter Teil) und fasst man diese in die in der Literatur gebräuchlichen Ausfallgründe zusammen, dominieren quantitativ die Verweigerer, die über 80 % der systematischen Ausfälle auf sich vereinigen. Am zweithäufigsten ist der Anteil der nicht erreichbaren Stichprobenelemente, die nicht interviewt werden konnten, mit 11,12 %. Nichtbefragbare machen lediglich 6,55 % der systematischen Ausfälle im ALLBUS 2008 aus. Bei den Ausfallgründen gibt es nur sehr marginale Unterschiede zwischen Ost- und Westdeutschland. Während die Verweigerer den höchsten Anteil seit Beginn des ALLBUS aufweisen, bleiben die Anteile für Nichterreichbare und Befragungsunfähige im Rahmen des bereits Aufgetretenen.
Die bereits in der Theorie vorgestellte Formel der Ausschöpfungsquote wird hier für den ALLBUS 2008 berechnet:

| | |
|---|---|
| urspr. Bruttostichprobe (8.735)<br>+ zus. Adressen für neutrale Ausfälle (1.100) | = Bruttostichprobe = 9.661 |
| − stichprobenneutral Ausfälle (1.050) | = bereinigter Stichprobensatz (8.611) |
| − systematische Ausfälle (5.142) | = auswertbare Interviews = 3.469 |

Tabelle 40: Berechnung der Anzahl auswertbarer Interviews für den ALLBUS 2008

Damit ergibt sich für den ALLBUS 2008 nach dem Pre_release eine Ausschöpfungsquote von

$$\frac{\text{auswertbare Interviews}}{\text{bereinigter Stichprobensatz}}$$

$$\frac{3.469}{8.611} = 0,4029$$

Mit Blick auf die Unterscheidung in *systematische* und *neutrale* Ausfälle lässt sich auch im konkreten Fall an die Kritik, die in der Literatur zu finden ist (Schnell 1997; Birkelbach 1998; Porst 1996), anschließen. Einige der Kategorien liegen eng beieinander: so erscheint es plausibel, dass ein Teil der nun Verstorbenen (*neutraler* Ausfall) im Falle der Interviewanfrage vor dem Tod zu den nicht Befragungsfähigen gezählt hätten (*systematisch*). Kritisch lässt sich hinterfragen, ob alle Verzogenen *neutrale* Ausfälle darstellen oder nicht wenigstens teilweise in Zusammenhang mit eventuell interessierenden Variablen stehen könnten. Dies wirft ferner die Frage auf, inwieweit die zusätzlich ausgegebenen Adressen, die die als *neutral* definiertem Ausfälle kompensieren sollen, vielleicht verzerrend wirken (Diekmann 2006); auf der anderen Seite lassen sich diese Situationen in der Praxis kaum anders handhaben, wenn damit keine massive Kostensteigerung einhergehen soll.

### 5.3.2 Erklärungsmodell für Unit Nonresponse

Eine statistische Analyse von Unit Nonresponse stellt ganz andere Herausforderungen an die Analyse als Item Nonresponse. Während die Item Nonresponse-Analyse eine sehr spezielle Modellschätzung notwendig macht (Abschnitt 3.3.3), da jeder Befragte von keinem bis k (Anzahl der Items) Item Nonresponse aufweisen könnte, kann eine Person nur einmal nicht an der Erhebung teilnehmen. Nachdem zwar konkret die Abgrenzung der Unit Nonresponse in *systematische* und *neutrale* Ausfälle, und diese wiederum in *nichterreichbar*, *nicht befragungsfähig* und *verweigert*, strittige Fälle zur Folge haben kann, erscheint die Beibehaltung der Unterscheidung generell für die Analyse sinnvoll. Auf diese Weise lässt sich auf den theoretischen Rahmen, der in Abschnitt 3.2 aufgespannt wurde, zurückgreifen. Eine Testung gerade der wertrationalen Theorien zur Befragungsteilnahme ist mit diesen Daten im Nachhinein allerdings nicht mehr durchführbar.
Zunächst soll umrissen werden, welche Unterschiede zwischen den Interviewten und Ausfällen aus unterschiedlichen Gründen bestehen. Menold und Züll grenzen die persönlichen und demografischen Merkmale der **Verweigerer** wie folgt ein: Geschlecht, Alter, soziale Lage, Bildung, Gemeindegröße, Lebensstil, soziale Kontakte und gesellschaftliche Partizipation (Menold und Züll 2010). Diese Variablen werden immer wieder als evidente Determinanten für Teilnahmeverweigerung genannt und überschneiden sich mit den mutmaßlichen Gründen für **Befragungsunfähigkeit** und **Nichterreichbarkeit** – soziale Lage, Lebensstil und Partizipation dienen ab und an als entsprechende Indikatoren für wertrationale Einstellungen (Engel et al. 2004). Mit dem ALLBUS 2008 stehen solche Variablen nicht zur Verfügung. Dennoch kann das im Theorieteil skizzierte Modell nach Groves und Couper als Gerüst für die Analyse auch dieses Datensatzes dienen (Groves und Cooper 1998): Damit lassen sich die Determinanten in Interviewermerkmale, Erhebungsmerkmale, Befragtenmerkmale und Umgebungsmerkmale einteilen. Die für die Analyse herangezogenen Variablen werden zur Vergleichbarkeit jeweils für alle Ausfallgründe – Verweigerung, Nichterreichbarkeit, Befragungsunfähigkeit und *neutrale* Ausfallgründe – gleichermaßen in die Modelle eingefügt. Da das Surveydesign, also die Erhebungsmerkmale, bei einer Erhebung in der Regel in weiten Teilen konstant ist (Ausnahme ist beispielsweise das Modeexperiment des ALLBUS 2000), wird an dieser Stelle die einzig verfügbare Variable zu den Interviewumständen herangezogen. Darüber hinaus wird ein Block mit Interaktionsvariablen eingeführt, der Interaktionen zwischen Befragten und Interviewer abdeckt, soweit dies der Variablenbestand zulässt:[120]

[120] In der letzten Spalte der Tabelle 41 sind zudem die Ausprägungen der Variablen vermerkt.

| Faktoren | Variablen | Ausprägungen |
|---|---|---|
| Umgebung | Landesteil | dichotom |
| | Gemeindegröße | 10 Ausprägungen |
| Befragtenmerkmale | Geschlecht | dichotom |
| | Alter | 4 Kategorien |
| | Nationalität | dichotom |
| Interviewermerkmale | Geschlecht | dichotom |
| | Bildung | 4 Ausprägungen |
| | Alter | 3 Kategorien |
| | Erfahrung | 4 Kategorien |
| Surveymerkmal | Anzahl der Kontakte | |
| Interaktionen | Int.Alter-Bef.Alter | 4 Kategorien |
| | Int.Geschlecht-Bef.Geschlecht | dichotom |

Tabelle 41: Beschreibung der Variablen für das Erklärungsmodell für Unit Nonresponse

Für die Variablen im Bereich persönliche Merkmale lassen sich folgende Hypothesen bilden:[121]

*Hypothese 1: Frauen* verweigern eher als Männer die Teilnahme an der Erhebung.

In der Literatur gibt es hierfür eine einhellige Erklärung. Frauen gewährten Fremden weniger gern Zutritt in den Privatbereich (DeMaio 1980; Esser 1973; Pickery und Loosveldt 2002; Porst und Schneid 1989; Rogelberg und Luong 1998; Stoop 2004; Williams et al. 2007; Zeh 1976).

*Hypothese 2:* Das *Alter* dürfte vor allem positiv mit der Erreichbarkeit korrelieren. Jüngere Zielpersonen sind wohl eher schlechter zu erreichen. Auch die Befragungsunfähigkeit dürfte mit dem *Alter* zunehmen.

Meistens wird beim *Alter* mit der Berufstätigkeit argumentiert; dies geht einher mit Attributen wie kinderlos, alleinstehend und mobil (Stoop 2004). Bei den befragungsunfähigen Personen gilt als Ausfallgrund eine Krankheit der Zielperson, die eher bei älteren Zielpersonen auftreten wird (Reuband 2006).

*Hypothese 3:* Bei nichtdeutscher *Nationalität* kann vermutet werden, dass die Person mit größerer Wahrscheinlichkeit nicht die entsprechenden Sprachkenntnisse zur Teilnahme an der Erhebung aufweist und deshalb als nicht befragungsfähig gilt.

[121] Der Übersichtlichkeit halber werden nur erwartete Effekte auf die abhängigen Variablen Befragungsunfähigkeit, Nichterreichbarkeit und Verweigerung formuliert; für die neutralen Ausfälle müsste die Globalhypothese lauten, dass es generell überhaupt keinen Einfluss gibt.

Für den Block der Umgebungsvariablen lassen sich folgende Hypothesen aufstellen:

*Hypothese 4:* Mit zunehmender *Gemeindegröße* sinkt sowohl die Wahrscheinlichkeit der Teilnahme als auch der Erreichbarkeit.

Im Zusammenhang mit der *Gemeindegröße* wird häufig mit der Anonymität und damit verbundenen Angst (u.a. Kriminialitätsangst, vgl. Schnell und Rässler 2004) argumentiert – bezogen auf die Verweigerung. Genauso zutreffend könnte aber auch die höhere Mobilität von Stadtbewohnern als Korrelat zur Nichterreichbarkeit angesehen werden (Steeh 1981). Überdies könnte man die Hypothese auch noch auf neutrale Ausfälle erweitern. Dies könnte damit zusammenhängen, dass bei steigender Größe der Gemeinde die Einwohnermelderegister tendenziell größere Fehler aufweisen (Fitzgerald und Fuller 1982).

Zudem steht der *Landesteil* (Ost- oder Westdeutschland) als Kontrollvariablen in den Modellen zur Verfügung.

Für den Bereich der Erhebungsmerkmale konnte nur eine Variable in das Modell eingefügt werden – nämlich die Anzahl der Kontaktversuche. Aufgrund des vorgeschriebenen Ablaufs während der Feldarbeit kann die folgende Hypothese formuliert werden (Wasmer et al. 2010, S.52ff):

*Hypothese 5:* Verweigerung und Nichterreichbarkeit dürfte mit einer höheren Zahl von Kontaktversuchen einhergehen, da entweder versucht wird, doch noch jemanden von der Teilnahme zu überzeugen oder jemanden im Haushalt anzutreffen.

Ebenfalls bereits angesprochen wurde die Rolle des Interviewers bei der Kontaktanbahnung mit dem potentiellen Befragten (Groves und Cooper 1998, S.191ff). Hier stehen, gleich wie bei der Analyse von Item Nonresponse, für alle *Interviewer*alter, *-geschlecht*, *-erfahrung* und *-bildung* als Merkmale zur Verfügung.

*Hypothese 6, 7 und 8:* Mit steigender *Erfahrung*, steigender *Bildung* und höherem *Alter* können eher vermindernde Einflüsse auf die Wahrscheinlichkeit von Verweigerung und Nichterreichbarkeit erwartet werden.

Bildung, Alter und damit wohl korrelierend die Erfahrung des Interviewers dürften bei der Überwindung von Hürden eine Rolle spielen und auch positiv mit einer Überredung zur Teilnahme korrelieren (Pickery und Loosveldt 2002; Steeh 1981).

*Hypothese 9:* Des Weiteren dürften weibliche Interviewer eine geringere Verweigerungshaltung auslösen.

Aus der Theorie lässt sich dieser Geschlechtereffekt damit erklären, dass Frauen offensichtlich für die Zielperson eine weniger unangenehme Störung bedeuten als männliche Interviewer; zudem dürften sie weniger als Bedrohung wahrgenommen werden.

Der letzte Block, der in die Modelle eingefügt wird, betrifft die *Interaktionsvariablen* und ist eine Wiederholung der bereits in Abschnitt 3.3.3 aufgestellten ‚Homogamie'-These (Koch 2002; Nealon 1983; Groves und Couper 1998, S.34). Die Wirkung dieser Merkmale kann sich nur auf die Gruppen beschränken, bei denen tatsächlich Kontakt zwischen Interviewer und Befragten zustande kommt.

*Hypothese 10:* Für die Teilnahme förderlich ist ein Interviewer, der ähnliche Merkmale aufweist wie der Befragte.

Für die abhängigen dichotomen Variablen[122] werden mit den oben genannten und beschriebenen Variablen Logitmodelle geschätzt:

[122] Die Zugehörigkeit zur Gruppe mit dem jeweiligen Ausfallgrund ist als 1 codiert. Mit 0 ist die Gruppe der Befragten codiert, bei denen das Interview realisiert wurde.

| | n. Ausfälle | nichterreichb. | verweigert | befragungsunf. |
|---|---|---|---|---|
| Konstante | 0,336** | 0,0042** | 0,336*** | 0,057*** |
| Soziale Umgebung | | | | |
| Landesteil (1=Ost) | 1,052$^{ns}$ | 1,189$^{ns}$ | 0,970$^{ns}$ | 0,763$^{ns}$ |
| Gemeindegröße | 1,171*** | 1,547*** | 1,048** | 0,971$^{ns}$ |
| Befragtenmerkmale | | | | |
| Geschlecht (1=weibl.) | 0,318* | 0,827$^{ns}$ | 1,181** | 1,218$^{ns}$ |
| Ausländer (1=nich dt.) | 2,060*** | 1,195$^{ns}$ | 0,575*** | 8,392*** |
| 18-29 Jahre | 3,032*** | 3,161** | 0,996$^{ns}$ | 0,677$^{ns}$ |
| 30-44 Jahre | 1,876*** | 3,179*** | 1,248$^{ns}$ | 0,767$^{ns}$ |
| 45-59 Jahre | 0,731$^{ns}$ | 1,811** | 1,114$^{ns}$ | 0,765$^{ns}$ |
| 60- Jahre | - | - | - | - |
| Interviewermerkmale | | | | |
| Geschlecht (1=weibl.) | 0,994$^{ns}$ | 0,813$^{ns}$ | 0,801** | 0,733$^{ns}$ |
| Bildung | 0,992$^{ns}$ | 0,998$^{ns}$ | 0,966$^{ns}$ | 1,211** |
| -10 Jahre tätig | 0,933$^{ns}$ | 2,050** | 1,371** | 0,754$^{ns}$ |
| 11-20 Jahre tätig | 0,731$^{ns}$ | 1,548$^{ns}$ | 1,214$^{ns}$ | 0,687$^{ns}$ |
| 21-30 Jahre tätig | 0,835$^{ns}$ | 2,050* | 1,264** | 0,811$^{ns}$ |
| 31- Jahre tätig | - | - | - | - |
| 29-44 Jahre | 1,335$^{ns}$ | 0,459** | 0,865$^{ns}$ | 0,338** |
| 45-59 Jahre | 1,020$^{ns}$ | 1,091$^{ns}$ | 0,704$^{ns}$ | 1,008$^{ns}$ |
| 60- Jahre | - | - | - | - |
| Surveymerkmal | | | | |
| Anzahl der Kontakte | 0,749*** | 1,195*** | 1,138*** | 1,051** |
| Interaktion Bef.-Int. | | | | |
| Geschlecht (0=gleich) | 1,119$^{ns}$ | 1,116$^{ns}$ | 1,144** | 1,253$^{ns}$ |
| -5 Jahre jünger | 1,066$^{ns}$ | 1,502$^{ns}$ | 1,405* | 5,654** |
| 6 jünger-11 älter | 0,594$^{ns}$ | 1,332$^{ns}$ | 1,094$^{ns}$ | 1,566$^{ns}$ |
| 12-25 Jahre älter | 0,896$^{ns}$ | 1,006$^{ns}$ | 1,051$^{ns}$ | 1,094$^{ns}$ |
| 26- Jahre älter | - | - | - | - |
| Nagelkerke $R^2$ | 0,21 | 0,28 | 0,06 | 0,18 |
| | n=3.879 | n=3.474 | n=6.532 | n=3.243 |

Tabelle 42: Ergebnisse der Logitmodelle nach Ausfallgründen

Zusammenfassend lässt sich zunächst feststellen: In den vier Modellen wirken die verwendeten Determinanten sehr unterschiedlich.

Auch die Erklärungskraft der Modelle variiert beträchtlich. So kann die Nichterreichbarkeit einer ausgewählten Person für empirische Verhältnisse relativ gut erklärt werden, die sogenannten neutralen Ausfälle und Ausfall aufgrund von Befragungsunfähigkeit wenigstens befriedigend (ca. 20 % Nagelkerke $R^2$), während das Modell bei Verweigerung insgesamt eher enttäuscht.

Beim Faktor **soziale Umgebung** weist die Variable *Landesteil* in keinem der Modelle einen signifikanten Beitrag. Dagegen erscheint die in der Literatur häufig zitierte Variable *Urbanisierung* (hier in Form der Gemeindegröße) mit einer Ausnahme als signifikanter Erklärungsfaktor, und zwar so wie es in der Hypothese 4 für Nichterreichbarkeit und Verweigerung formuliert wurde.

Die Variablen des Faktors **Befragtenmerkmale** zeigen mit Ausnahme der Determinante *Ausländer* keine signifikanten Ergebnisse für die Modelle Verweigerung und Befragungsunfähigkeit. Bei Befragungsunfähigkeit wird Hypothese 3 damit bestätigt. Dennoch stößt die quantitative Analyse ohne weitere Merkmale wohl hier an ihre Grenzen. Der Umstand, dass eine mögliche Zielperson Ausländer ist, erhöht die Wahrscheinlichkeit der Befragungsunfähigkeit stark, was angesichts der expliziten Definition nicht wundert. Auf der anderen Seite vermindert sich die Wahrscheinlichkeit für Verweigerung durch diesen Umstand. An dieser Stelle könnte vermutet werden, dass ein systematischer Ausfall (Verweigerung) durch einen neutralen Ausfall (angebliche Befragungsunfähigkeit) substitutiert wird. Ob dies tatsächlich zutrifft, lässt sich anhand dieser Daten kaum mit Sicherheit sagen. Für Verweigerung und Befragungsunfähigkeit spielt das *Alter* des Befragten in den Modellen keine Rolle – anders als es die Hypothese nahe legt. Das Modell Nichterreichbarkeit zeigt dagegen deutlich: je jünger die Zielperson ist, desto größer ist die Wahrscheinlichkeit diese nicht anzuteffen. Eine Bestätigung der oben formulierten Hypothese 2.

Die Ergebnisse für die Variable *Geschlecht* bestätigen die Hypothese 1 bezüglich einer erhöhten Verweigerungswahrscheinlichkeit von Frauen.

Sehr unterschiedlich zeigt sich die Wirkung der **Interviewermerkmale**, von denen eine ganze Reihe für die Modelle zur Verfügung stand. Erwartungsgemäß haben die Interviewermerkmale keinen Einfluss auf die Wahrscheinlichkeit eines neutralen Ausfalls, ansonsten könnte auch dies auf Probleme in der Feldphase hinweisen. Die komplexe Interaktion von Interviewer und Befragten, die entweder zu Verweigerung oder zur Realisierung eines Interviews führen kann, hinterlässt auch bei einem vereinfachten Modell wie diesem Spuren. So sinkt tatsächlich bei weiblichen Interviewern die Wahrscheinlichkeit der Verweigerung, wie es nach der Hypothese 9 zu erwarten ist. Die Variable *Bildung des Interviewers* zeitigt alleine im Modell Befragungsunfähigkeit eine signifikante Wirkung, allerdings dahingehend, dass mit steigender Bildung die Wahrscheinlichkeit zunimmt, dass die Zielperson eher als befragungsunfähig eingestuft wird. Der Einfluss der Bildung des Interviewers scheint damit entgegen der Hypothese 7 für Erreichbarkeit und Verweigerung keine allzu große Rolle zu spielen. Ob dies an der besseren Einschätzung der Situation bei Kontaktaufnahme liegt, lässt sich wiederum nicht endgültig feststellen.

Die am wenigsten *erfahrenen Interviewer* weisen bei Verweigerung und Nichterreichbarkeit eine signifikant höhere Wahrscheinlichkeit eines Ausfalls auf als die erfahrenen Interviewer – im Falle der Nichterreichbarkeit sogar noch deutlich stärker – der Verlauf ist aber keinesfalls so linear, wie es aufgrund der Hypothese 6 erwartbar war. Neben der *Erfahrung des Interviewers* konnte in die Modelle auch das *Alter des Interviewers* eingefügt werden.[123]

Das *Alter* zeigt allerdings nur für Nichterreichbarkeit und für Nichtbefragungsfähigkeit einen signifikanten Einfluss. Der Einfluss liegt in der Gestalt vor, dass die jüngste Alterskategorie der Interviewer stark senkend auf die Wahrscheinlichkeit von Nichterreichbarkeit und Nichtbefragbarkeit wirkt. Dies muss nicht im Widerspruch zu den Ergebnissen der Erfahrungsvariablen stehen – es lässt allerdings an der Hypothese 8 zweifeln. Gerade bei der Befragungsfähigkeit, die zu Recht als sehr subjektive Kategorie in der Kritik steht, könnten sich jüngere Interviewer mehr Mühe geben als ältere Interviewer. Eventuell ist dies ein Hinweis auf die Ambivalenz der Interviewererfahrung.

Das einzige Merkmal im Faktor **Survey Design**, die *Anzahl der Kontaktversuche*, verhält sich wie nach Hypothese 5 erwartet. Nachdem eine Zielperson als neutraler Ausfall registriert wurde, finden konsequenterweise keine Kontaktversuche mehr statt.

Die wenigen signifikanten Werte der **Interaktionsvariablen** finden sich im Verweigerungsmodell und bei Befragungsunfähigkeit. Bei der Interaktionsvariablen für das Geschlecht von Interviewer und Befragten weist die Kombination Mann-Mann bzw. Frau-Frau eine signifikant niedrigere Wahrscheinlichkeit zur Verweigerung auf. Die Altersinteraktion zeigt ebenfalls eine deutliche Richtung. In der Kombination jüngerer Interviewer erhöht sich die Verweigerungswahrscheinlichkeit, die Wahrscheinlichkeit, dass die Zielperson als nichtbefragungsfähig vermerkt wird, steigt jedoch noch wesentlich deutlicher an. Hier könnte ein gewisses Autoritätsproblem jüngerer Interviewer gegenüber älteren Befragten der Grund für diese eindeutigen Werte sein. Die ‚Homogamie'-These scheint sich damit teilweise zu bestätigen (Hypothese 10).

[123] Da für den ALLBUS nur erfahrene Interviewer zum Zug kommen sollen, ist der jüngste Interviewer bereits 27 Jahre alt (Technical Report 2010,4).

Wie schon beim Erklärungsmodell für Item Nonresponse wird auch für Unit Nonresponse ein Teil der in der Literatur diskutierten Hypothesen bestätigt. Die teilweise geringe Erklärungskraft der Modelle – besonders bei Verweigerung – kann vielleicht als Indikator gelten, dass sich viel ungeklärte Varianz in Merkmalen und Dispositionen versteckt, die für die Analyse nicht zur Verfügung standen. Deutlich wurde jedoch die sinnvolle Trennung der Analyse nach den Hauptausfallgründen.

# 6 Unit Nonresponse: Korrekturmethoden im Vergleich

Nach der Identifikation von Einflussfaktoren auf die Wahrscheinlichkeit von Unit Nonresponse werden wie bei Item Nonresponse nunmehr Korrekturmethoden verglichen. Das nun für Unit Nonresponse modifizierte Verfahren under der Stresstest aus Abschnitt 4.3 wird in Abschnitt 6.1 noch einmal beschrieben. Auch die ausgewählten Korrekturmethoden sind andere: für Unit Nonresponse wird die Gewichtung mit der Multiplen Imputation verglichen. Die Ergebnispräsentation geschätzter uni- und multivariater Parameter umfasst danach die Abschnitte 6.3.1 bis 6.3.6.

## 6.1 Verfahren zum Vergleich von Korrekturmethoden: Modifikation für Unit Nonresponse

Wie in der grafischen Darstellung (Abbildung 55) zu sehen ist, ändert sich von der Reihenfolge der Schritte nichts. Die Datenlage erzwingt allerdings Anpassungen und wirkt sich auch auf die Aussagekraft des Korrekturmethodenvergleichs aus.
Anders als im Kapitel über Item Nonresponse wird nur ein Datensatz verwendet, der ALLBUS 2008 Pre_lease; auch die uni- oder multivariaten Schätzungen werden aus Stichproben einer Grundgesamtheit geschätzt, die nur aus diesem einen Datensatz erzeugt wurde.

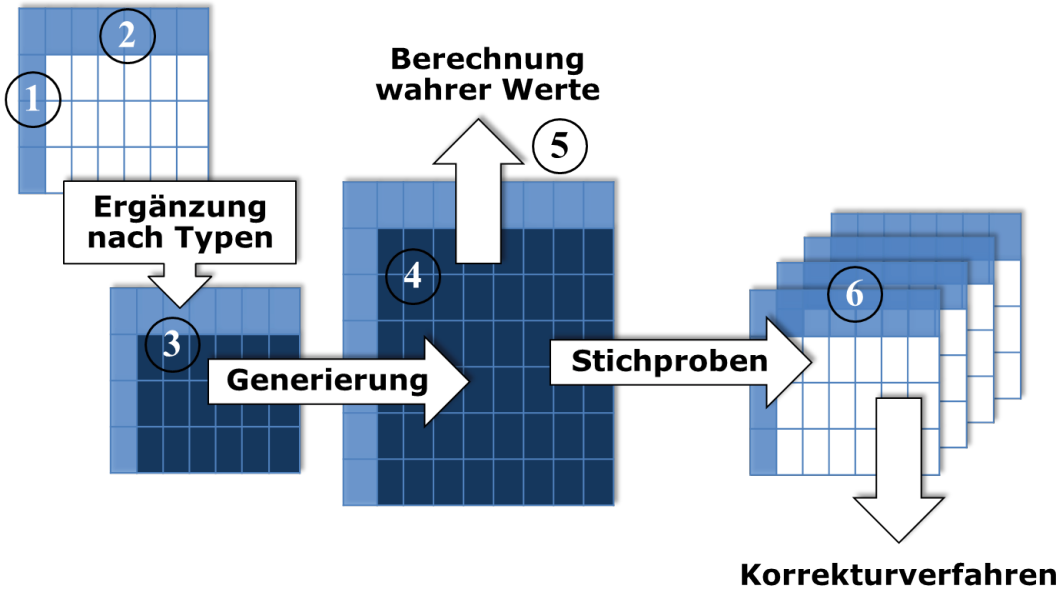



Abbildung 55: Schema Verfahren zum Vergleich von Unit Nonresponse-Korrekturmethoden

Wie der ALLBUS 2002, 2004 und 2006 besitzt auch dieser ALLBUS einige Variablen, die vollständig sind, unabhängig davon, ob der mögliche Befragte teilgenommen hat, nicht erreicht wurde, nicht befragungsfähig war oder verweigert hat:

- Alter
- Geschlecht
- Erhebungsgebiet
- Nationalität
- Größe der politischen Gemeinde
- Geschlecht des Interviewers
- Alter des Interviewers
- Erfahrung des Interviewers
- Anzahl der persönlichen oder telefonischen Kontaktversuche
- Interviewerwechsel

Diese Variablen wurden auch für die Analyse von Unit Nonresponse herangezogen. Nun bilden sie den Streifen vollständiger Informationen am linken Rand des Datensatzes (Abbildung 55, Punkt 1). Allen anderen Variablen fehlen aufgrund von Nichterreichbarkeit, Befragungsunfähigkeit oder Verweigerung gleichermaßen etwa 60 % der Antworten. Dieser Ausfall übertrifft bei weitem die Ausfälle, die die Beispiele für die Item Nonresponse aufweisen. Die restlichen 40 % enthalten zunächst auch noch Item Nonresponse. Um nun die Leistungsfähigkeit von Korrekturverfahren allein für Unit Nonresponse zu messen, wird das Item Nonresponse-Problem vollständig ausgeblendet, indem zunächst nur Variablen für die Parameterschätzung ausgewählt werden, die vergleichsweise wenig Item Nonresponse enthalten würden. Zweitens werden die verbliebenen Item Nonresponse imputiert.[124] Dies bedeutet, es liegt, anders als bei Item Nonresponse, ein idealisiert monotoner Ausfall vor.
Die Auffüllung der Unit Nonresponse-Zellen in Punkt 3 der Abbildung 55 beschert dem in Abschnitt 4.3 beschriebenen Vorgehen allerdings ein doppeltes Problem. Doppelt meint in diesem Zusammenhang, dass die Konstruktion von Extremtypen doppelt hypothetisch ist: man vermutet für Items ein Antwortverhalten von Befragten, die zu diesem Item nie befragt wurden, weil sie entweder nicht erreichbar oder nicht befragungsfähig waren oder die Teilnahme an der gesamten Befragung verweigert haben. Ein weiterer Diskussionspunkt sind die Konsequenzen, die sich aus der Anwendung der Extremtypenbildung ergeben. Aus der Monotonie des Ausfalls sowie des Ausfallsumfangs könnte eine Überforderung jeglicher Korrekturmethoden resultieren, die für den Korrekturenvergleich bei Item Nonresponse in der Form kaum Relevanz zeigte. Da die Extremtypen allein zur Berechnung „wahrer“ Werte herangezogen werden, später dann aber entsprechend der Ausfallmuster der Originaldaten gelöscht werden (also nicht in der Stichprobe vor dem Einsatz der Korrekturmethode zur Verfügung stehen, Punkt 6 in Abbildung 55), ist die Gefahr groß, dass die Extremtypen außerordenlich extreme Parameter zur Folge haben. Eine ähnliche Diskussion findet sich bei Horowitz und Manski (2000), die mit dem Konzept der „Bounds“ Kriterien für die Güte von Parameterschätzungen unter Ausfall entwickelt haben.[125] Als zusätzliche Alternative zu Extremtypen kommt neben gemäßigten Typen (wie er beim Vergleich von Item Nonresponse-Korrekturmethodenvergleich zum Einsatz kam) zudem jeweils ein „wahrer“ Wert in Frage, der für die gegebenen Daten die Bedingung MAR erfüllt (hierzu Abschnitt 6.2.2).
Für Punkt 4 gilt überdies: Aufgrund des Verhältnisses der vorgegebenen absoluten Zahlen hat die erzeugte Grundgesamtheit einen wesentlich größeren Umfang als bei Item Nonresponse. Zum Vergleich: für Unit Nonresponse wird tausend mal eine Stichprobe von der Originalgröße 9.661 gezogen.
Da die Konstruktion von Typen bedeutend schwieriger ist, gibt es auch Anpassungen an die Gegebenheiten von Unit Nonresponse in Hinblick auf Darstellungsmöglichkeiten der Ergebnisse. Bei Punkt 6 sollte noch ergänzt werden, dass hier unterschiedlich große Stichproben gezogen werden, um die Ausfallhöhe für MI zu variieren. Der Effekt wird für die uni- und multivariaten Parameterschätzungen jeweils grafisch aufbereitet.

[124] Hierbei handelt es sich um eine Single Imputation bei Punkt 1 und 2, Abbildung 55.

[125] Das Vorgehen unterscheidet sich in einigen Punkte erheblich. So sind die Bänder von Horowitz und Manski (2000) als extreme Parameter konstruiert, die selbst bei geringerm Datenausfall im multivariaten Raum eine beträchtliche Breite erhalten. Dies kritisiert in einem Kommentar auch Raghunatan (2000).

## 6.2 Ausgewählte Parameter

Die sehr spezifische Datenlage verhindert die Replikation publizierter Analysen, wie es beim Methodenvergleich von Item Nonresponse der Fall war. Dennoch wurde bei der Auswahl der Parameter auf eine breite Mischung uni- und multivariater Parameter geachtet. In Abschnitt 6.2.1 werden als univariate Parameter ein Anteilswert- und Mittelwertbeispiel vorgestellt; multivariate Parameterschätzungen beinhalten ein OLS- und ein Logitmodell. Die Modifikationen bezüglich des Stresstests, die der Methodenvergleich für Unit Nonresponse notwendig macht, werden in Abschnitt 6.2.2 erläutert. Die für Unit Nonresponse zum Vergleich angewendeten Korrekturverfahren sind wie die Verfahren bei Item Nonresponse Postsurvey Adjustments:[126] verglichen wird Multiple Imputation mit zwei Gewichtungen verschiedener Komplexität (6.2.3).

### 6.2.1 Uni- und multivariate Parameter

Alle hier vorgestellten Parameter und Modelle besitzen einen monotonen Datenausfall von etwa 60 %. Alle Parameter und Modelle wurden zwar in dieser Form nicht veröffentlicht, sie könnten aber so oder in ähnlicher Form als Analyse vorkommen. Der erste univariate Parameter ist der Anteilswert von Personen, die die *SPD bei der Bundestagswahl wählen* würden. Der Mittelwert einer zehnstufigen Skala (der *Linksrechtsselbsteinstufung*) bildet den zweiten geschätzten univariaten Parameter.
Dann folgen insgesamt zwei Modelle mit einer abhängigen Variablen und einer Anzahl von unabhängigen Merkmalen. Das erste Modell bildet eine OLS-Regression auf die abhängige Variable *Zufriedenheit mit Regierungsleistung* (kurz *RZuf*). OLS-Regressionen erfreuen sich trotz Annahmenverletzungen in der Sozialwissenschaft aufgrund der einfachen Anwendbarkeit und Interpretierbarkeit großer Beliebtheit. Auch hier sind viele Voraussetzungen für eine OLS-Schätzung nicht geben.[127] Dennoch sollen die Paremter zunächst mittels dieser Regression als Repräsentant einer populären Schätzpraxis geschätzt werden. Die Determinanten dieses Modells lauten:

- Linksrechtsselbsteinstufung (*Linksrechts*),
- Berufstätigkeit der Frau oder Karriere des Mannes (*Beruf*),
- Wahrscheinlichkeit CDU zu wählen (*Wahrcdu*),
- Wahrscheinlichkeit SPD zu wählen (*Wahrspd*).

[126] Wobei noch einmal betont werden sollte, dass CC keine Korrekturmethode, jedoch eine Verfahrensweise für Nonresponse ist.

[127] Mit viel Fantasie lässt sich die abhänige Variable *Zufriedenheit mit Regierungsleistung* mit fünf Ausprägungen als metrische Variable auffassen.

Das Modell lautet dann:

$$ZReg_i = \beta_1 + \beta_2 Linksrechts_i + \beta_3 Beruf_i + \beta_4 Wahrcdu_i + \beta_5 Wahrspd_i$$

mit $i = 1, \ldots, n$

Die *Linksrechtsselbsteinstufung* ist bereits aus der Schätzung für den Mittelwert bekannt; das Frauenberufstätigkeitsitem ist eine dichotome Variable, während die beiden Wahlwahrscheinlichkeiten elfstufige Skalen aufweisen. Als nächstes folgt ein Logitmodell, dessen abhängige dichotome Variable die *Wahl der Unionsparteien* (kurz: *Wahlcdu*) ist. Auch das Logitmodell ist ein typischer Vertreter einer häufig verwendeten Schätzpraxis für multivariate Analysen in den Sozial- und Wirtschaftswissenschaften. Die unabhängigen Variablen sind:

- Wahrscheinlichkeit CDU zu wählen (*Wahrcdu*),
- Berufstätigkeit der Frau oder Kindererziehung (*Kind*),
- Nationalstolz (*Stolz*).

Das Modell mit lautet dann:

$$logit(Wahlcdu) = \beta_1 + \beta_2 Wahrcdu_i + \beta_3 Kind_i + \beta_4 Stolz_i$$

mit $i = 1, \ldots, n$

Wiederum besitzt die Wahlwahrscheinlichkeit elf Ausprägungen, die Frauenvariable ist dichotom und die Variable *Nationalstolz* weist fünf Ausprägungen auf.

### 6.2.2 Stresstest

Wie schon zuvor angesprochen, unterscheidet sich die Datenlage bei Unit Nonresponse erheblich im Umfang und dem Muster des Ausfalls von der Situation bei Item Nonresponse. Auch die Überlegungen zu möglichen Ausfallmechanismen im Abschnitt 5.1.5 signalisieren eine andere Ausgangslage als bei Item Nonresponse. Darauf sollte auch der Stresstest reagieren und entsprechend modifiziert werden, soweit es möglich ist. Auf alle Fälle ist die Konstruktion eines Stresstests für Unit Nonresponse im Einzelnen von größerer hypothetischer Natur als bei Item Nonresponse. Neben den Extremtypen, die aus der Verwendung für den Korrekturenvergleich bei Item Nonresponse beibehalten werden, muss als Konsequenz für die besonderen Gegebenheiten bei Unit Nonresponse jeweils ein „wahrer“ Wert herangezogen werden, der auf einer Schätzung unter MAR beruht.

Die gebildeten Extremtypen tendieren wiederum mehr oder weniger stark zu NMAR. Für den Anteilswert kann davon ausgegangen werden, dass beim ersten Extremtyp keiner der Nichtrespondenten SPD wählen würde und beim zweiten Extremtyp alle Nichtrespondenten SPD wählen würden. Wie in Tabelle 45 zu sehen ist, führt dies zu einer großen Spannweite im Anteilswert von knapp 7 % für Extremtyp 1 und knapp 80 % für Extremtyp 2.[128] Ein großer Abstand liegt auch beim Mittelwert der *Linksrechtsselbsteinstufung* für die Extremtypen vor.
Für die multivariaten Modelle gibt es zwei auch theoretisch fundierte Typen. Typ 1 verschweigt seine *Zufriedenheit mit der Regierung*, die einhergeht mit einer rechten politischen Selbsteinstufung (*Linksrechtsselbst.*). Die Frau soll für diesen Typus die *Karriere des Mannes unterstützen* anstatt selbst berufstätig zu sein. Die *Wahrscheinlichkeit CDU zu wählen* ist maximal, die *Wahrscheinlichkeit SPD zu wählen* hingegen minimal. Typ 2 ist genau entgegengesetzt. Die Tabellen 42 und 43 zeigen noch einmal zusammengefasst die Ausprägungen der beiden Typen und jeweils alle Ausprägungen der abhängigen und der unabhängigen Variablen:

| Variable | Ausprägungen | Extremtyp 1 | Extremtyp 2 |
|---|---|---|---|
| Zuf. mit Reg. | 1-6 | 1 | 6 |
| Linksrechtsselbst. | 1-10 | 10 | 1 |
| Frau unterstützt | 1-4 | 1 | 4 |
| Wkt. CDU-Wahl | 1-10 | 10 | 1 |
| Wkt. SPD-Wahl | 1-10 | 1 | 10 |

Tabelle 43: Bildung von Typen für die OLS-Regression

Für das Logitmodell werden zur Gewinnung der Parameter, die den „wahren“ Wert darstellen, die fehlenden Werte zunächst durch zwei Extremtypen ersetzt.

| Variable | Ausprägungen | Extremtyp 1 | Extremtyp 2 |
|---|---|---|---|
| Wahl Union | 0-1 | 1 | 0 |
| Wkt. CDU-Wahl | 1-10 | 10 | 1 |
| Frau berufstätig | 1-4 | 1 | 4 |
| Nationalstolz | 1-5 | 5 | 1 |

Tabelle 44: Bildung von Typen für das Logitmodell

[128] Die Tabelle 45 verzeichnet alle möglichen „wahren“ Werte, die als Vergleichswerte und zur Berechnung des MSE dienen. Die Werte werden nach Möglichkeit auch in der grafischen Umsetzung verwendet.

Die zwei Extremtypen des Logitmodells lassen sich als klassischer, konservativer CDU-Wähler einerseits und dessen Antityp andererseits beschreiben. Der CDU-Wähler von Extremtyp 1 wählt plausiblerweise sehr *wahrscheinlich CDU*, möchte nicht, dass die *Frau berufstätig* ist, wenn sie kleine Kinder hat und ist ausgesprochen *stolz, Deutscher zu sein*. Entsprechend entgegengesetzt sind die Ausprägungen für den Extremtyp 2 zugewiesen.

| Parameter | Bezeichnung | unter MAR | Extremtyp 1 | Extremtyp 2 |
|---|---|---|---|---|
| Anteilswert | *Wahl SPD* | 0,2550 | 0,0683 | 0,7901 |
| Mittelwert | *Linkrechtsselbsteinst.* | 5,1900 | 2,5010 | 8,2651 |
| OLS | $\beta_1$ *ZReg* | 4,4310 | 3,0396 | 4,7086 |
| | $\beta_2$ *Linksrechts* | −0,0300 | −0,2106 | −0,3833 |
| | $\beta_3$ *Beruf* | −0,0740 | 0,5211 | 0,2529 |
| | $\beta_4$ *Wahrcdu* | −0,1010 | −0,1619 | −0,1806 |
| | $\beta_5$ *Wahrspd* | −0,0700 | 0,0269 | 0,0675 |
| Logit | $\beta_1$ *Wahlcdu* | −2,2330 | −4,7832 | −5,7566 |
| | $\beta_2$ *Wahrcdu* | 0,1888 | 0,7591 | 0,6194 |
| | $\beta_3$ *Kind* | 0,0053 | −0,7494 | −0,1611 |
| | $\beta_4$ *Stolz* | 0,0700 | 0,3683 | 0,3399 |

Tabelle 45: Werte der Typen für den Anteilswert, Mittelwert, Parameter der OLS-Regression und des Logitmodells

Die so gewonnenen Werte dienen im Weiteren wieder als Vergleichswerte für die mit den Korrekturverfahren geschätzten Werte und werden von einem dritten Wert in Spalte 3 ergänzt, der die Bedingung MAR erfüllt. Soweit es in der grafischen Darstellungen praktikabel erscheint, werden sie auch visuell umgesetzt.

### 6.2.3 Ausgewählte Korrekturverfahren

Die Auswahl der Korrekturverfahren richtet sich zum einen wieder nach der Verbreitung. Das bei weitem gängigste Korrekturverfahren in den Sozialwissenschaften ist in diesem Bereich immer noch die Gewichtung. Diese Korrekturmethode soll wiederum verglichen werden mit MI, die bisher kaum zur Korrektur in diesem Bereich eingesetzt wird. Da die Gewichtung nicht nur zur Korrektur des Nonresponsefehlers eingesetzt wird, sondern auch für

1. Design-Erfordernisse,
2. den Versuch der Reduktion des Zufallsfehlers durch Schichtung a posteriori und
3. die Sicherstellung der Vergleichbarkeit herangezogen wird,

ist die Literatur zu diesem Thema entsprechend umfangreich (Rösch 1994, S.9ff). Literatur für die Lösung des Unit Nonresponse-Problems mit Hilfe von MI sucht man dagegen fast vergebens. Es gibt in der Forschung zu diesem Thema nur wenige Veröffentlichungen im Rahmen sehr spezifischer Datensätze und dazu noch auf akademische Erhebungen begrenzt (z.B. Schnell und Rässler 2004). Abschnitt 6.2.3.1 beschreibt das Prinzip der Gewichtung als Korrekturverfahren von Unit Nonresponse und stellt die beiden Beispielgewichtungen für den späteren Methodenvergleich vor. Die Idee, MI als Korrekturmethode für Unit Nonresponse heranzuziehen, wird in Abschnitt 6.2.3.2 skizziert.

#### 6.2.3.1 Gewichtung

In der Regel meint man mit Gewichtung Zellgewichtungen; der Name Zellgewichtung ist allerdings etwas irreführend, da Gewichtung so definiert ist: diese ist zunächst die „Vergabe von positiven reellen Zahlen, die Gewichtungsfaktoren, an die Merkmalsträger, d.h. an die die befragten Personen repräsentierenden Merkmalsvektoren“ (Rösch 1994, S.7). Gewichtet werden die Merkmalsträger mit ihrer oder ihren Ausprägungen eines, mehrerer oder aller Merkmale (Globalgewicht): z.B.

$$w_i = \frac{n_j d_j N/n}{N_j}$$

Der individuelle Gewichtungsfaktor einer Zelle ist also ein Quotient aus der Anzahl der Beobachtungen in der Zelle multipliziert mit dem Designgewicht und der Sollvorgabe.[129] Natürlich

[129] Die Sollvorgabe wird als Restriktion bezeichnet; zur mathematischen Formulierung des daraus entstehenden Anpassungsproblems siehe Rösch (1994), S.11ff.

sind auch andere Verfahren denkbar und in der Praxis weit verbreitet: mehrere einfache Zellgewichtungen hintereinander, der Iterative Proportional Fitting Algorithmus, Verfahren minimaler Varianz der Gewichtungsfaktoren, Verfahren nach dem Prinzip des minimalen Informationsverlustes (Rösch 1994, S.10) usw.
Die wichtige Frage bei der Berechnung der Gewichtung ist die Frage der Gewichtungsmerkmale. Deren Auswahl liegen häufig Plausibilitätsüberlegungen zu Grunde. Generell gilt: je höher die Korrelation zwischen den Gewichtungsmerkmalen und der Zielmerkmale ist, desto besser die Ergebnisse. Im Fall von Unit Nonresponse werden teilweise auch Designgewichte eingeflochten, die Wahrscheinlichkeiten für Erreichbarkeit und für Verweigerung in Subgruppen einbeziehen. Ziel der Gewichtung ist die Anpassung des unter Unit Nonresponse leidenden Datensatzes bezüglich der ausgewählten Gewichtungsmerkmale an die Grundgesamtheit. Rösch spricht hierbei allerdings von „blinder Substitution“ (Rösch 1994, S.10) – eine Kritik, die später noch einmal aufgegriffen wird. In der Tat muss einerseits die Verteilung des potentiellen Gewichtungsmerkmals (Redressmentmerkmal) in der Gesamtheit überhaupt bekannt sein, andererseits ist die Wirkung der Gewichtung selbst unter Umständen verzerrend.[130]
In unserem Fall sind die Verteilungen der Gewichtungsvariablen idealisierterweise bekannt. Dies liegt an der künstlichen Grundgesamtheit, deren Erzeugung in Abschnitt 6.1 beschrieben wurde, und auf der der Methodenvergleich basieren wird. In der Realität wird die externe Information aus anderen Erhebungen entnommen, wie z.B. aus dem Mikrozensus (Rothe und Wiedenbeck 1994, S.51).
Die Kritikpunkte an Gewichtungen entfalten sich unter anderem am in Kauf genommen Effektivitätsverlust (Rösch 1994, S.12f; Stenger 1994, S.44). Ist die Korrelation zwischen Gewichtungs- und gewichteten Merkmalen nicht sonderlich groß oder gar gering, erfährt die Stichprobenvarianz eine Vergrößerung.
Der zweite Kritikpunkt bezieht sich direkt auf den Umgang mit dem Ausfallmechanismus. Ein nachlässiger Umgang mit dem Ausfallmechanismus verzerre bei Gewichtung noch zusätzlich, wie Rothe und Wiedenbeck anhand eines Beispiels, das sich an Oh und Scheuren (1983) anlehnt, illustrieren. Bliebe man bei einer einfachen Zellgewichtung, müsste man sich fragen, wie komplex diese Gewichtung werden soll, damit der Ausfallmechanismus adäquat abgefangen werden könnte.[131] Die Wirkung der Gewichtung auf die gewichteten Merkmale lässt sich in der Regel nicht wirklich vorhersagen: „eine allgemeine Aussage über Gewichtungseffekte ist nicht möglich“ (Rothe 1994, S.64). Schnell und Rässler (2004) interpretieren die Gewichtung als Korrekturmethode von Unit Nonresponse unter Umständen als bedingte Mittelwert-Imputation – allerdings müssen hierfür dann auch die entsprechenden Annahmen zutreffen, was in der Praxis fraglich ist, nämlich die Homogenität der Ausfallwahrscheinlichkeiten in den durch die Gewichtung gebildeten Gruppen (Schnell 1993; Rothe und Wiedenbeck, S.54). Anders als bei CC für den Umgang mit Item Nonresponse wird für Gewichtung bereits von MAR als Aufallmechanismus bezüglich der Gewichtungsmerkmale und der Merkmale, die unter Datenausfall leiden, ausgegangen.

[130] Die Literatur unterscheidet prinzipiell zwischen Transformationsgewichtung und Redressmentgewichtung; bei ersterer wird ein Gewicht proportional zum Kehrwert der Auswahlwahrscheinlichkeit eines Elements im Sinne des Stichprobenplans berechnet (Horvitz-Thompson); Redressmentgewichtungen passen die bestehende Stichprobenmatrix an externe Strukturen (Bundesland x Altersgruppe x Geschlecht usw.) als Kombination von Merkmalen (aktive Merkmale) an, vgl. Rothe und Wiedenbeck (1994); Rösch (1994).

[131] Es dürften aber aufgrund der Eigenschaften des Horvitz-Thompson-Schätzers gar nicht sonderlich komplizierte oder zu feine Gewichtungen verwendet werden, vgl. Rothe und Wiedenbeck (1994), S.52.

Für den Methodenvergleich werden zwei Gewichtungen berechnet und die Analysevariablen, deren uni- oder multivariate Parameter geschätzt werden sollen, damit jeweils gewichtet.[132]

G1: Geschlecht x Bildung

G2: Alter x Geschlecht x Bildung

*Alter* und *Geschlecht* sind häufig verwendete Gewichtungsmerkmale (Rothe 1994, S.64). Die Variable *Bildung* (hier gemessen als höchster formaler Schulabschluss) hält als Determinante in der sozialwissenschaftlichen Theorie eine herausragende Stellung (Rothe 1994, S.65). Auch für das Teilnahmeverhalten sowie für die Erreichbarkeit ist sie von tragender Bedeutung. Leider steht die Bildungsvariable für Unit Nonresponse bei Querschnittsdaten in der Regel nicht zur Verfügung – anders als Alter und Geschlecht, die häufig im Sample Frame gegeben sind (z.B. durch die Ziehung in der Einwohnermelderegisterstichprobe). Im konkreten Fall ist die Verteilung der Bildungsvariablen in der künstlichen Grundgesamtheit exakt bekannt. Beim Vergleich der Gewichtungen als Korrekturmethode müsste bei der empirischen Überprüfung die komplexere Gewichtung G2 der weniger komplexen G1 überlegen sein.

#### 6.2.3.2 Multiple Imputation

Vom Prinzip ändert sich für die Verwendung von MI bei Unit Nonresponse nichts am Verfahren – es werden wie bei den Item Nonresponse-Beispielen Standard-MI-Verfahren angewendet (ebenfalls mit MICE in R). Das Ausfallmuster ist nunmehr monoton. Die auch für Item Nonresponse verwendeten vollständigen Variablen zählen zu Paradaten und sind nicht-reaktiv, die im Verlauf der Feldphase oder schon zu einem früheren Zeitpunkt im Survey Lifecycle vorhanden sind oder gesammelt werden. Diese Daten können auch interessante Informationen über den Ausfall der Merkmalsträger enthalten, wie dies in der Analyse zu Unit Nonresponse in Abschnitt 5.3.2 (auch für die Item Nonresponse-Analyse in Abschnitt 3.3.3) der Fall war. Unglücklicherweise ist der Umfang dieser Paradaten sehr begrenzt; auch ihre Erklärungskraft z.B. für das Auftreten von Verweigerung scheint eher gering zu sein; ein Teil dieser Variablen steht auch nur auf Nachfrage in speziell editierten Datensätzen zur Verfügung. So verwundert es nicht, dass die Verwendung von Paradaten bei der Korrektur von Unit Nonresponse erst am Anfang steht (Kreuter und Olson 2011; Sackshaug und Kreuter 2011).
In einem weiteren Punkt unterscheidet sich das Vorgehen bei MI im Vergleich zu den Item Nonresponse-Beispielen: Der Umfang der Ausfälle ist noch einmal bedeutend größer. Zudem war die Information in den Variablen, die von Ausfällen betroffen waren, im Analysemodell nicht unerheblich, während bei Unit Nonresponse nur Informationen aus Paradaten (aufgelistet auf Seite 150f) vorhanden sind. Die Variablen, für die eine oder mehrere Parameter geschätzt werden (Abschnitt 6.2.1), und die Paradaten bilden im Folgenden zusammen wiederum das Imputationsmodell.
Aufgrund des massiven Datenausfalls von etwa 60 % soll beim folgenden Methodenvergleich der

[132] Eine ähnliche Konstruktion macht Rothe (1994), S.70f.

Ausfall schrittweise erhöht werden. Das bedeutet, dass zwar im Originalstichprobenumfang aus der Grundgesamtheit gezogen wird. Die Stichprobe wird dann aber geteilt in einen Teil ohne Ausfälle und einen Teil mit den Ausfällen. Dem Teil ohne Ausfälle wird dann ein Prozentsatz des Teils mit Ausfällen hinzugefügt. Der Übersichtlichkeit halber wird dies in sieben Stufen ausgeführt.

### 6.3.4 Ergebnisse

Die folgenden Kapitel stellen die Ergebnisse der Korrekturmethoden am Beispiel verschiedener Parameterschätzungen dar. Einige Änderungen im Vergleich zur Ergebnispräsentation der Item Nonresponse-Korrekturmethoden ergeben sich aus der unterschiedlichen Datenlage; die Problematik wurde in den vorhergenden Abschnitten diskutiert. Anders als bei Item Nonresponse wird es hier von Anfang wenig aussagekräftig sein, Coveragequoten zu berechnen; die Stichprobenvarianzen für MI sind durch die Größe der Stichprobe (9.661) extrem schmal. Die Problematik für MI, die bereits von den Ergebnissen in Abschnitt 4.3.3 vor allem im univariaten Bereich aufgetreten ist, verstärkt sich dadurch in einem Maße, das das Coveragekriterium wertlos macht. Als Hauptanalysemethode erscheint eine Visualisierung der mit MI bzw. mit den Gewichtungen geschätzten Werte für die jeweiligen Parameter sinnvoller. Schließlich kann noch der jeweilige MSE mit Hilfe der „wahren“ Werte berechnet werden. Die jeweilige Methodenperformanz wird nach der Höhe des MSE in eine Reihenfolge gebracht, um so einen Überblick über mögliche Muster zu erhalten. Mit Verzicht auf die Coverageberechnung entfällt auch der Vergleich der Intervalllängen. Dafür wird möglichst anschaulich die Veränderung der MI-Schätzwertverteilung bei unterschiedlich hohen Graden des Ausfalls analysiert. Dabei werden die geschätzten Parameterwerte unter minimalem und maximalem Ausfall grafisch ausführlich dargestellt, die anderen Ergebnisse in etwas kürzerer Form.

#### 6.3.4.1 Ergebnis 1: Anteilswert

Für den Vergleich von Gewichtung und MI sollte zunächst der *Anteilswert der SPD-Wähler* bei der *Sonntagsfrage* geschätzt werden. Insgesamt wird der *Anteil der SPD-Wähler* weit unter 50 % liegen, sodass die Verteilung zwischen SPD-Wählern und Nicht-SPD-Wählern ziemlich asymmetrisch ist. In den folgenden Histogrammen sind einmal die Werte der MI-Schätzung für einen Ausfall von 12,5 % und für den vollen Ausfall von 100 % dargestellt. Da der einzig sinnvoll verwendbare „wahre“ Wert der Wert unter Richtigkeit von MAR ist, wird nur ein Strich (hier und im Folgenden an der grünen Farbe zu erkennen) in die Verteilungen eingezeichnet. Unter MAR beträgt der Wert der SPD-Wählerschaft gut 25 % (0,255). Wie bei der Konstruktion des Stresstests gezeigt wurde, sind alle anderen Werte nicht plausibel herleitbar oder nicht mehr im vernünftigen Rahmen visualisierbar.

Die Verteilung der Anteilsschätzwerte der SPD-Wählerschaft bei der Sonntagsfrage liegen bei einem Ausfall von 12,5 %, genau wie bei 100 % Ausfall, nur leicht unter dem „wahren“ Wert (grüne Linie) (Abbildung 55). Im Mittel schätzt MI den Anteil auf 0,2453 (mit 12,5 % Ausfall) bzw. 0,2413 (mit 100 % Ausfall). In der Variation der Schätzwerte gibt es große Unterschiede: während bei 100 % Ausfall die Standardabweichung 0,02463 beträgt, reduziert sie sich bei einem Ausfall von 12,5 % auf 0,00974. Wie verhalten sich dagegen die Schätzungen unter der Gewichtung (Abbildung 57)? Da die Achsendimensionierung keinen verzerrenden Einfluss auf die Darstellung der Verteilung haben soll, wirken die Unterschiede zur MI-Schätzung noch einmal wesentlich schärfer.

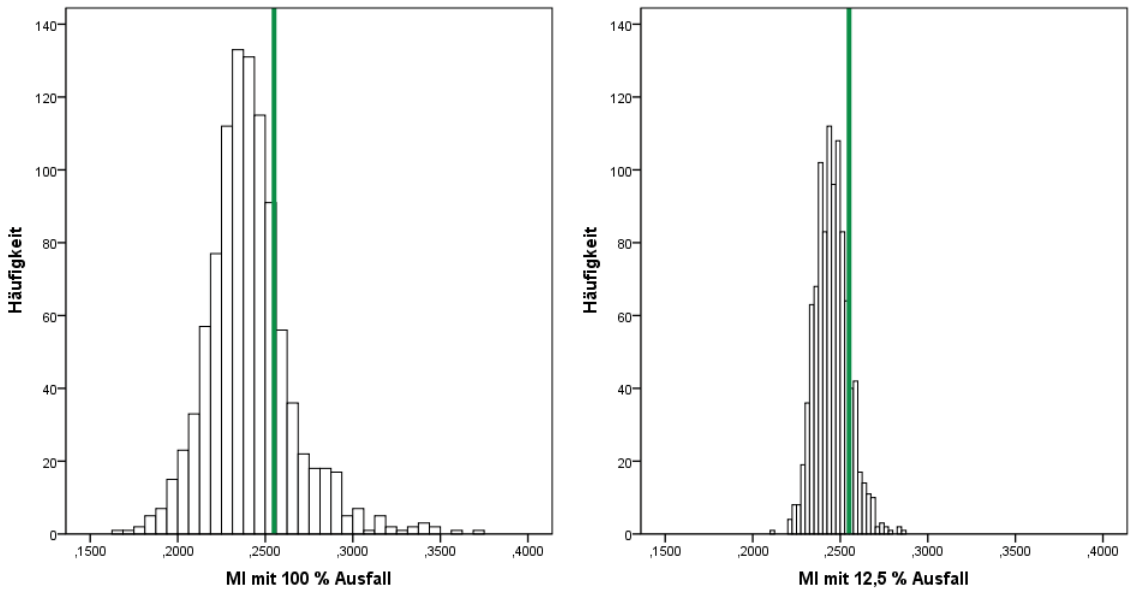



Abbildung 56: MI Histogramm des Anteilswerts zur Variablen *Wahl der SPD*

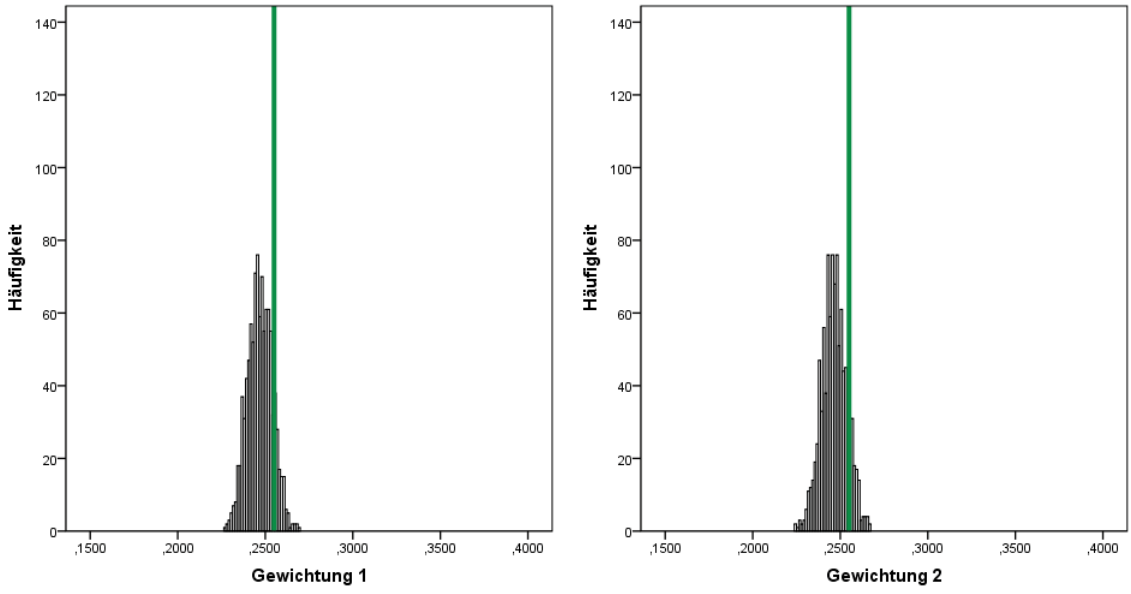



Abbildung 57: Gewichtung 1 und 2: Histogramm des Anteilswerts zur Variablen *Wahl der SPD*

Durchschnittlich wird bei Gewichtung 1 ein Wert von 0,2469, bei Gewichtung 2 ein geringfügig niedriger Wert von 0,2467 geschätzt. Die Standardabweichung beträgt 0,00708 für Gewichtung 1 und 0,00727 für Gewichtung 2. Beide Beispiele liegen damit unter dem Umfang der Variation bei der MI-Schätzung – selbst beim geringsten Ausfall von 12,5 %.
Abbildung 58 veranschaulicht über alle Ausfallstufen die Entwicklung der MI-Schätzwertverteilung.

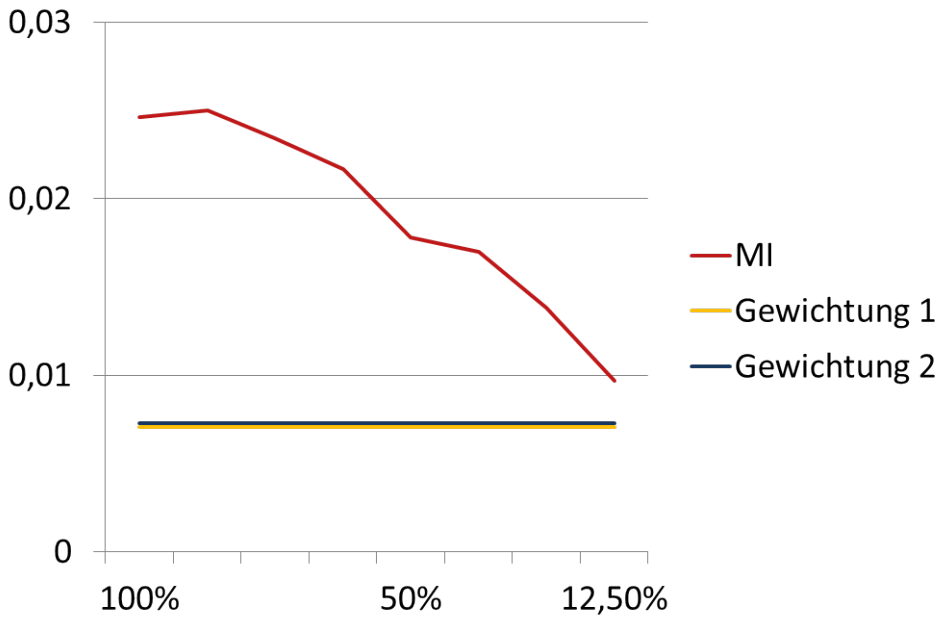



Abbildung 58: Veränderung der Standardabweichung der geschätzten Werte des Anteilswerts unter MI mit sinkendem Ausfall

Die durchschnittlichen Schätzwerte bei MI sind konstant und sehr nah an den durchschnittlichen Schätzwerten der beiden Gewichtungsbeispiele. Mit abnehmendem Umfang der Ausfälle verringert sich auch die Abweichung vom durchschnittlich geschätzten Anteilswert bei MI. Sie ist aber bei der Schätzung des Anteilswerts – wie oben bereits erwähnt – stets größer als bei den Gewichtungen. Die generelle Entwicklung lässt sich durch die verminderte Unsicherheit plausibel erklären (da der Ausfall geringer wird, sinkt die Unsicherheit).
Alle Verteilungen unterschätzen den „wahren“ Wert leicht. Im Stresstest weisen dann die beiden Extremtypen (die zu NMAR tendieren) „wahre“ Parameterwerte für den Anteilswert von ca. 8 % bzw. 80 % auf; beide Korrekturmethoden würden hier deutlich daneben liegen bzw. würden, wie schon in Abschnitt 6.1 angedeutet, überfordert sein.

#### 6.3.4.2 Ergebnis 2: Mittelwert

Auch in den Histogrammen für die Mittelwertschätzwerte findet sich nur eine einzige Markierung eines „wahren“ Wertes. Wiederum mit der Farbe Grün, die den „wahren“ Wert unter MAR abträgt. Die Gründe, sich auf diesen einen „wahren“ Wert zu beschränken, sind dieselben wie für die Analyse des Anteilswertes, also die Sprenngung der Darstellbarkeit dieser Extremwerte. Begonnen wird wiederum mit den geschätzten Mittelwerten der Variable *Linksrechtsselbsteinstufung* durch MI (Abbildung 59).

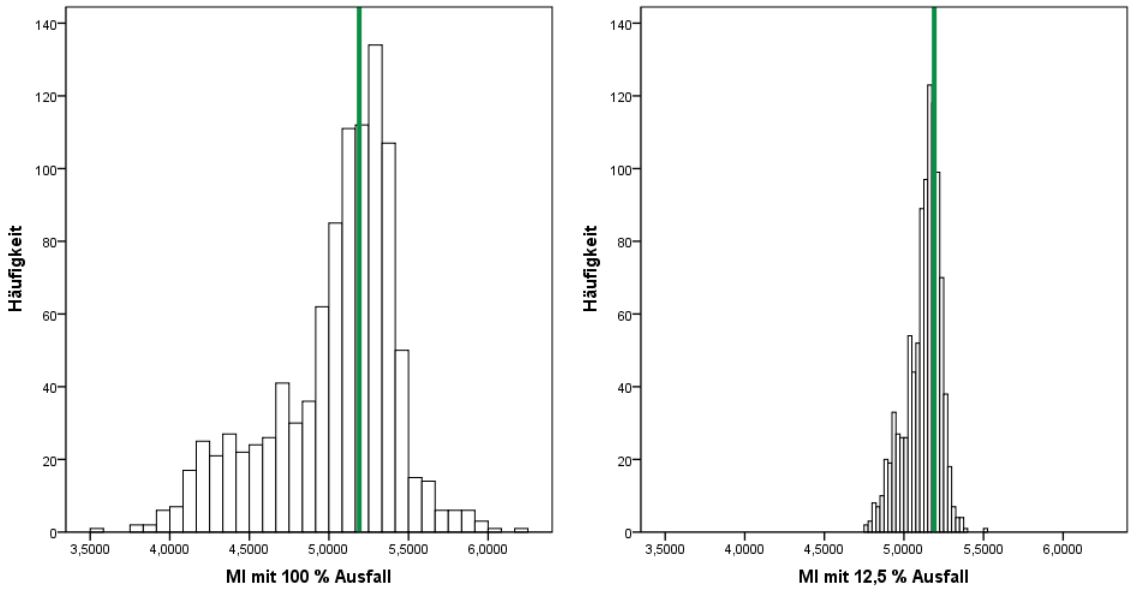



Abbildung 59: MI Histogramm des Mittelwerts zur Variablen *Linksrechtsselbsteinstufung*

Die Verteilungen einmal unter einem Ausfall von 12,5 % und einmal von 100 % sind nicht ganz symmetrisch um den „wahren“ Wert. Im Durchschnitt wird ein Mittelwert von 5,03 (100 % Ausfall) und 5,12 (12,5 % Ausfall) geschätzt. Wie schon bei den Anteilswerten sinkt die Standardabweichung mit sich reduzierendem Ausfall von 0,39100 auf 0,11114 ganz erheblich, während sie bei den Gewichtungen konstant bleibt (Abbildung 60):

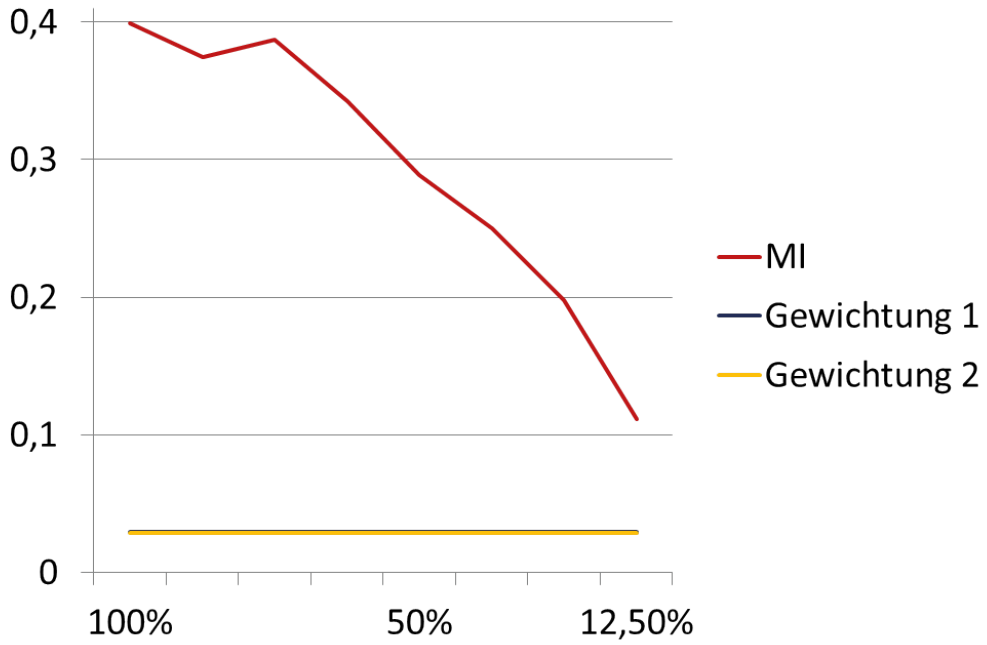



Abbildung 60: Veränderung der Standardabweichung der geschätzten Werte des Mittelwerts unter MI mit sinkendem Ausfall

Die geschätzten Mittelwerte bei den Gewichtungen verteilen sich folgendermaßen (Abbildung 61):

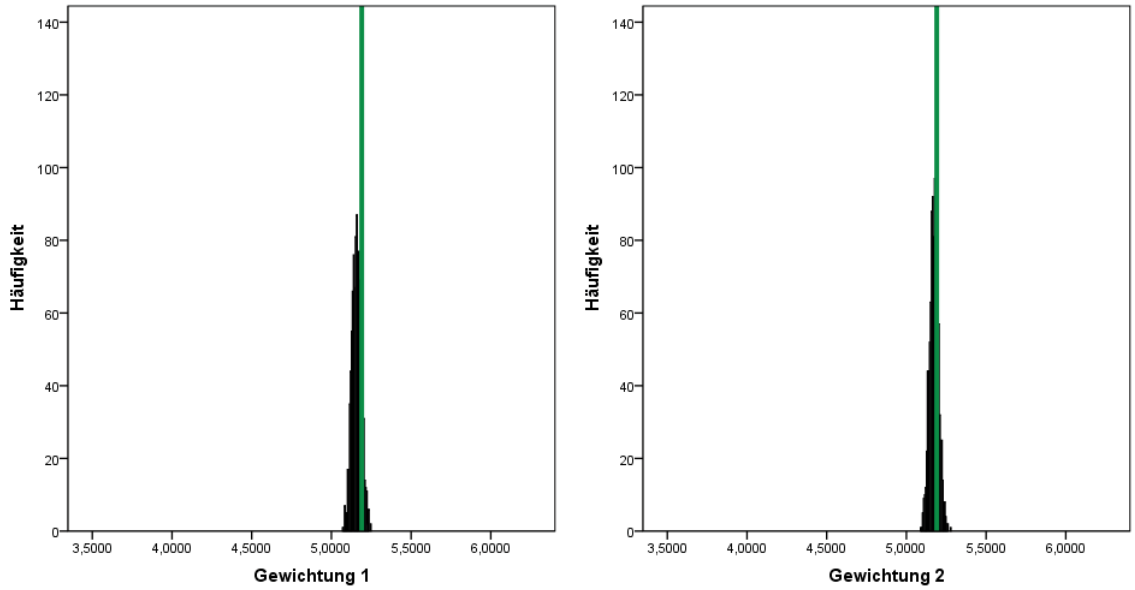



Abbildung 61: Gewichtung 1 und 2: Histogramm des Mittelwerts zur Variablen *Linksrechtsselbst-einstufung*

Durch die einheitliche Skalierung fällt zunächst wieder auf, wie vergleichsweise schmal beide

Schätzwerteverteilungen der Gewichte sind. Sowohl Gewichtung 1 als auch Gewichtung 2 unterschreiten in der Standardabweichung den Wert der MI mit 0,02926 bzw. 0,02845. Betrachtet man die Lage der beiden Verteilungen und setzt sie in Bezug zum „wahren“ Wert, so wird deutlich, dass wiederum nur eine leichte Unterschätzung stattfindet. Versucht man die Ergebnise mittels des MSE zusammenzufassen, ergibt sich folgende Reihenfolge:

Gew. 2: 0,0011 $<$ Gew. 1: 0,0019 $<$ MI 12,5 %: 0,0170 $<$ MI 100 %: 0,1765

Diese Werte zeigen, dass die MI mit der Imputation des vollständigen Ausfalls zu einem größeren MSE führt. Bei geringerem Datenausfall sinkt die Variation der MI-Schätzwerte allerdings schnell (Abbildung 60). Bei den Ergebnissen für die Gewichtungen verwundert es nicht, dass Gewichtung 2 bei der Analysevariable *Linksrechtsselbsteinstufung* eine bessere Performanz zeigt als Gewichtung 1, da diese nicht die für politische Einstufungen sehr relevante Variable Bildung einschließt.
Für die möglichen Extremwerte nach der Methode des Stresstests würden, wie beim Anteilswertbeispiel, alle Korrekturmethoden überfordert, da die „wahren“ Werte extreme 2,5 bzw. 8,3 annehmen.
Die Ergebnisse der univariaten Parameter machen bereits die Probleme deutlich, die eine spezifische Gewichtung sowie die eingeschränkte Auswahl an Prädiktoren für die MI mit sich bringen. An dieser Stelle sei allerdings daran erinnert, dass auch beim Korrekturmethodenvergleich bei Item Nonresponse die Stärke von MI nicht bei der Schätzung von univariaten Parametern, sondern bei multivariaten Schätzungen lag. Deren Ergebnisse werden nun vorgestellt.

#### 6.3.4.3 Ergebnis 3: OLS-Modell

Die Schätzung der Parameter des OLS-Modells beginnt den Reigen multivariater Parameter. Der erste hier betrachtete Parameter ist der Achsenabschnitt. In den ersten beiden grafischen Darstellungen werden die Verteilungen von MI-Schätzwerten abgetragen (Abbildung 62). Für den Achsenabschnitt erschienen drei „wahre“ Werte als Vergleichspunkte praktikabel und plausibel. Es sind die beiden Extremtypen (rote und blaue Linie) sowie wiederum die grüne Linie, die den „wahren“ Wert unter MAR symbolisiert.

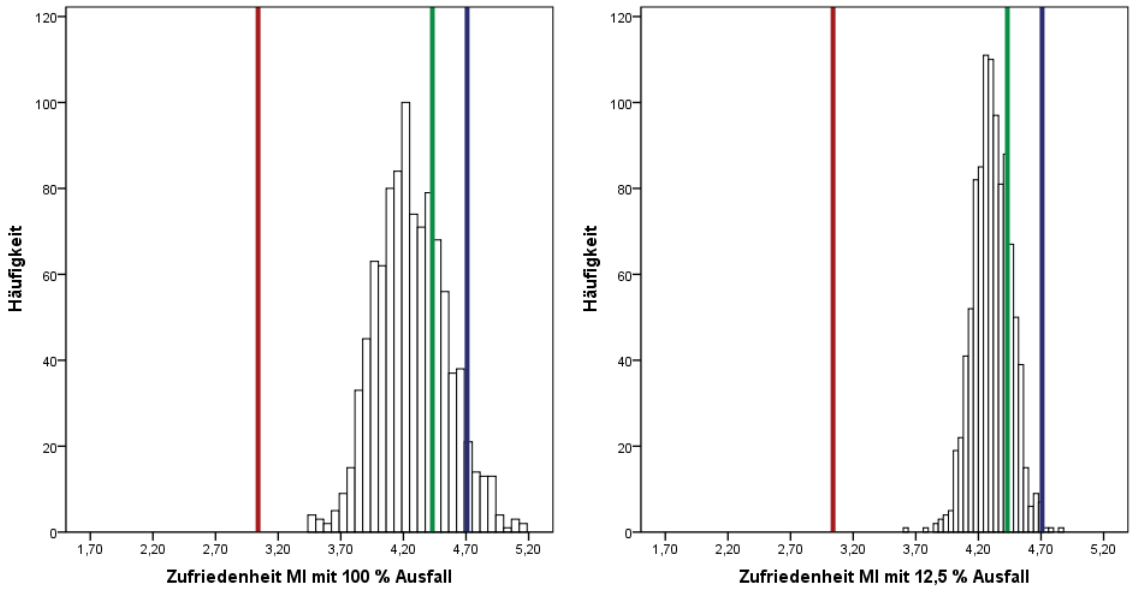



Abbildung 62: MI Histogramm des Achsenabschnitts der OLS-Regression

Die Verteilungen der mit MI geschätzten Werte unter verschiedenen Graden von Ausfällen ergeben für den Achsenabschnitt ein gewohntes Bild. Die Standardabweichung der Schätzwerte wird mit abnehmendem Ausfall deutlich geringer. So weisen die Schätzwerte bei einem Ausfall von 100 % eine Standardabweichung von 0,28309 auf, bei einem Ausfall von 12,5 % nur noch 0,15138. Die Verteilungen liegen zwischen den „wahren“ der Extremtypen bei durchschnittlich 4,2670 (mit 100 % Ausfall) und 4,3102 (mit 12,5 % Ausfall).

Die Gestalt der Schätzwertverteilung der beiden Gewichtungen unterscheidet sich merklich im Verhältnis zu den MI-Verteilungen (Abbildung 63).

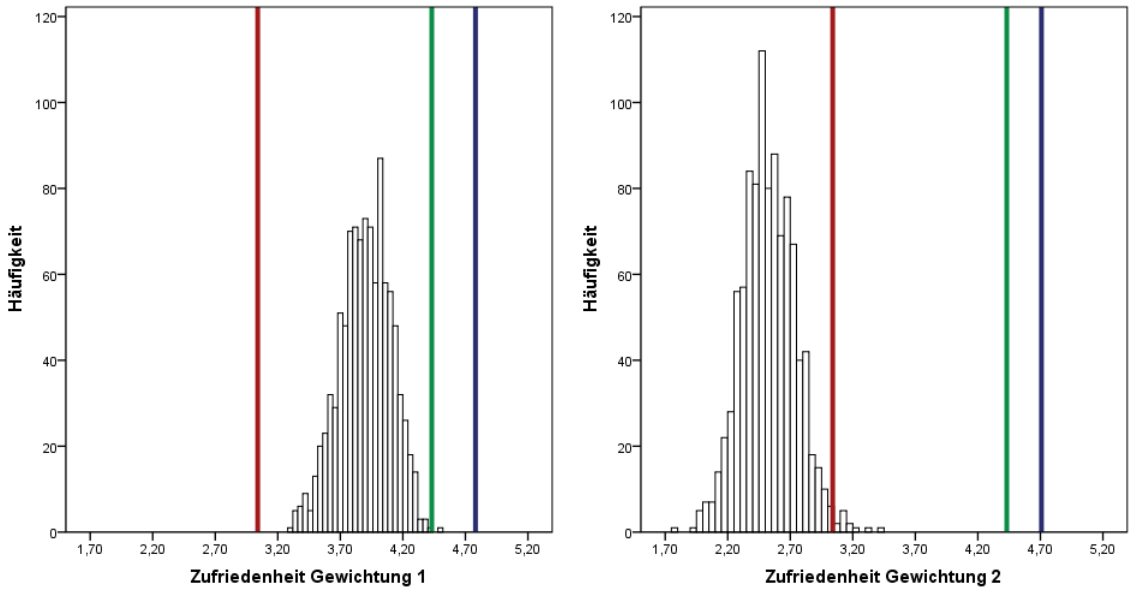



Abbildung 63: Gewichtung 1 und 2: Histogramm des Achsenabschnitts der OLS-Regression

Anders als bei den univariaten Schätzungen sind hier die Verteilungen deutlich breiter als die MI-Verteilung mit 12,5 % Ausfall: 0,20609 (Gewichtung 1) bzw. 0,21520 (Gewichtung 2). Noch auffälliger ist jedoch die Lage der beiden Verteilungen, die sich bedeutend unterscheiden (Abbildung 63). Während die Gewichtung 1 (3,9004) eine Verteilung noch in der Nähe des „wahren“ Wertes unter MAR aufweist, verschiebt sich die Verteilung der Schätzwerte von Gewichtung 2 nach links (2,5279).
Um eine eindeutige Aussage über das Abschneiden der Korrekturmethoden zu machen, wird hier noch der MSE bezüglich der drei „wahren“ Werte für die grafisch dargestellten MI-Verteilungen und die beiden Gewichtungen berechnet:

**Extremtyp 1**

Gew. 2: 0,4770 < Gew. 1: 0,9471 < MI 100 %: 1,5867 < MI 12,5 %: 1,7658

**Extremtyp 2**

MI 12,5 %: 0,1816 < MI 100 %: 0,2752 < Gew. 1: 0,6957 < Gew. 2: 4,8017

**wahrer Wert unter MAR**

MI 12,5 %: 0,0375 < MI 100 %: 0,1070 < Gew. 1: 0,324 < Gew. 2: 3,6681

Die Ergebnisse fallen je nach Typ unterschiedlich aus: Extremtyp 1 zeigt für MI jeweils einen wesentlich höheren MSE als für die beiden Gewichtungen. Allerdings schneidet hier Gewichtung 2 um einiges besser ab. Bei Extremtyp 2 hat MI bessere Ergebnisse vorzuweisen – mit niedrigerem Ausfall ist der MSE sogar noch einmal deutlich geringer. Gewichtung 1 zeigt einen wesentlich höheren MSE und Gewichtung 2 den schlechtesten MSE über alle Typen. Vergleicht man die Ergebnisse der Korrekturmethoden mit dem „wahren“ Wert unter MAR hat MI entsprechend der Ausfallehöhe (also MI mit 12,5 % besser als mit MI 100 %) jeweils einen wesentlich niedrigeren MSE.
Auch wenn der Effekt im Vergleich zu den univariaten Parametern wesentlich geringer ausfällt, führt auch hier die konstatierte niedrigere Standardabweichung bei sinkendem Ausfall zur Frage, ab welchem Ausfallanteil denn die Variation der geschätzten MI-Werte zumindest im Falle des Achsenabschnitts unter die der Gewichtungen fällt. Folgende grafische Darstellung beantwortet die Frage:

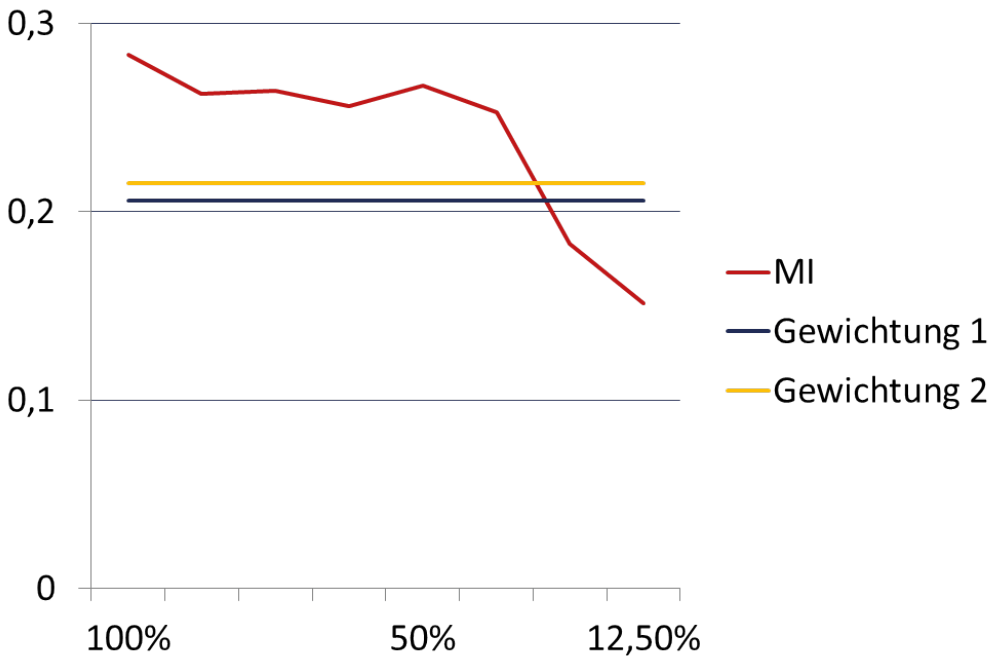



Abbildung 64: Veränderung der Standardabweichung der geschätzten Werte des Achsenabschnitts unter MI mit sinkendem Ausfall

Gut zu sehen ist in Abbildung 64, wie bei einem Ausfall von etwa 20 % und 30 % die Standardabweichung der mit MI geschätzten Werte unter die der beiden Gewichtungen fällt.

Der nun analysierte Parameter gibt den Einfluss der *Linksrechtsselbsteinstufung* auf die abhängige Variable an. Als „wahre“ Werte werden die Werte des Extremtyps 1 (rote Linie) und ein Wert unter MAR (grüne Linie) verwendet. Begonnen wird wiederum mit den MI-Schätzwerten unter unterschiedlichem Ausfall (Abbildung 65).

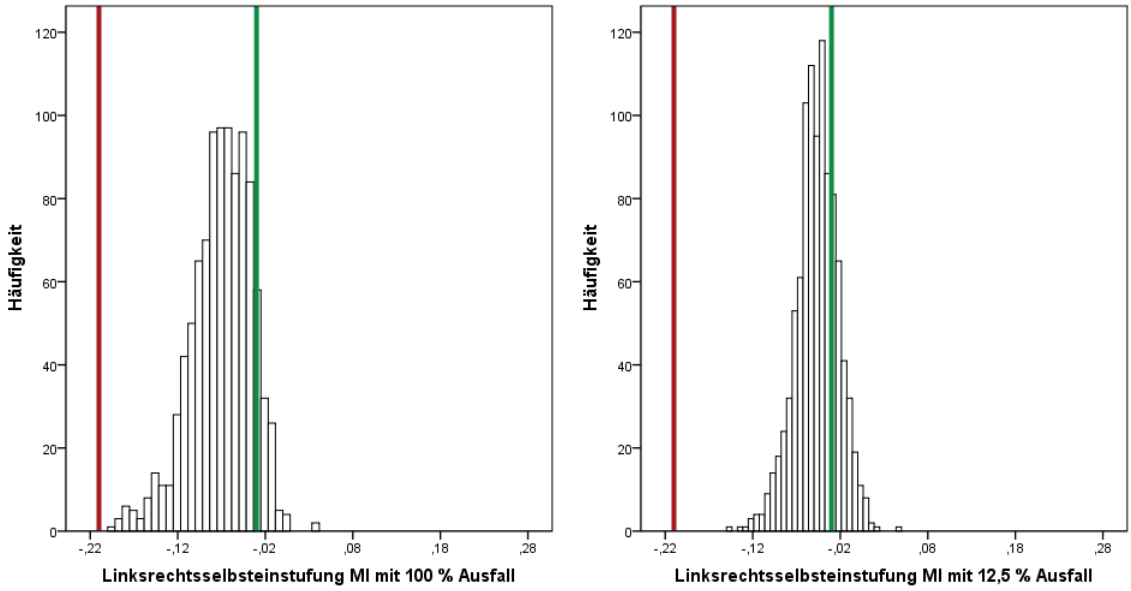



Abbildung 65: MI Histogramm des Parameters der Variablen *Linksrechtsselbsteinstufung*

Die Verteilungen der geschätzten Parameterwerte ähneln in ihrer Gestalt denen des Achsenabschnitts. Die Lage der Verteilungen unterscheidet sich kaum (MI 100 % Ausfall: -0,0703; MI 12,5 % Ausfall: -0,0467); die Streuung verringert sich von 0,03471 auf 0,02424.
Wie beim Achsenabschnitt besitzen die Verteilungen der nach den Gewichtungen geschätzten Werte teilweise eine größere Streuung (Abbildung 66):

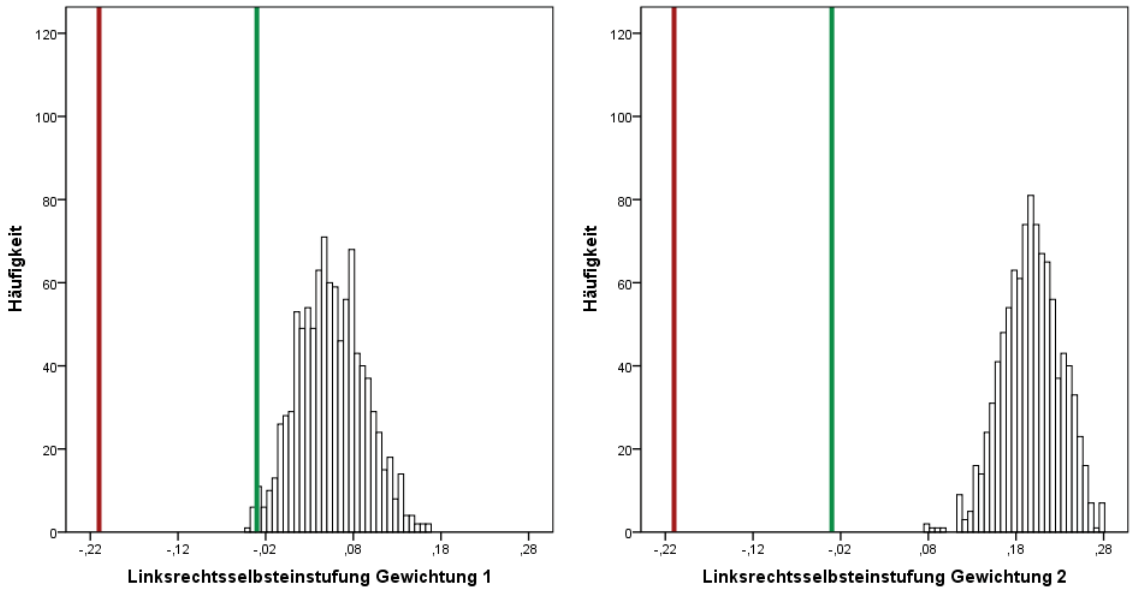



Abbildung 66: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Linksrechtsselbsteinstufung*

Die Standardabweichung bei Gewichtung 1 beträgt 0,03814 und bei Gewichtung 2 0,03355 und ist damit stets größer als bei MI mit 100 % Ausfall – Abbildung 67 veranschaulicht dies:

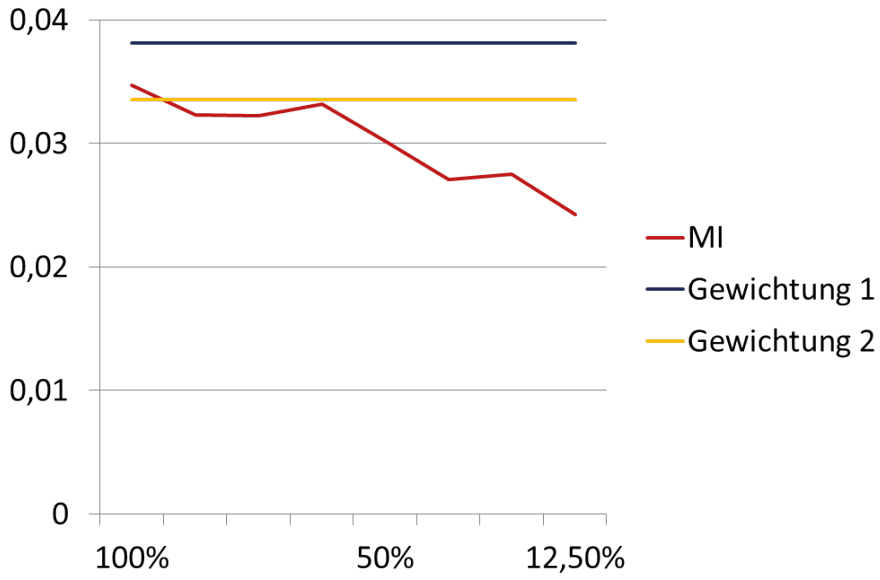



Abbildung 67: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Linksrechtsselbsteinstufung* unter MI mit sinkendem Ausfall

Während die Verteilung der mit Gewichtung 1 geschätzten Werte ähnlich wie die Verteilungen bei MI liegt, schert die Verteilung von Gewichtung 2 massiv aus (Abbildung 66). Die zwei zum Vergleich herangezogenen „wahren“ Werte führen zu folgenden MSE:

**Extremtyp 1**

MI 100 %: 0,0209 < MI 12,5 %: 0,0275 < Gew. 1: 0,0718 < Gew. 2: 0,1678

**wahrer Wert unter MAR**

MI 12,5 %: 0,0009 < MI 100 %: 0,0028 < Gew. 1: 0,0086 < Gew. 2: 0,0529

Die Ergebnisse der jeweiligen MSE für die Variable *Linksrechtsselbsteinstufung* sind eindeutig: MI (mit vertauschten Plätzen bei Ausfall von 12,5 % bzw. MI 100 %) besitzen den deutlich niedrigeren MSE und zwar diesmal sowohl für den Extremtyp 1 als auch für den „wahren“ Wert unter MAR.

Die dritte Determinante im OLS-Modell, die Variable „Frau sollte Karriere des Mannes unterstützen“, zeigt für die beiden grafisch umgesetzten MI-Schätzwertverteilungen:

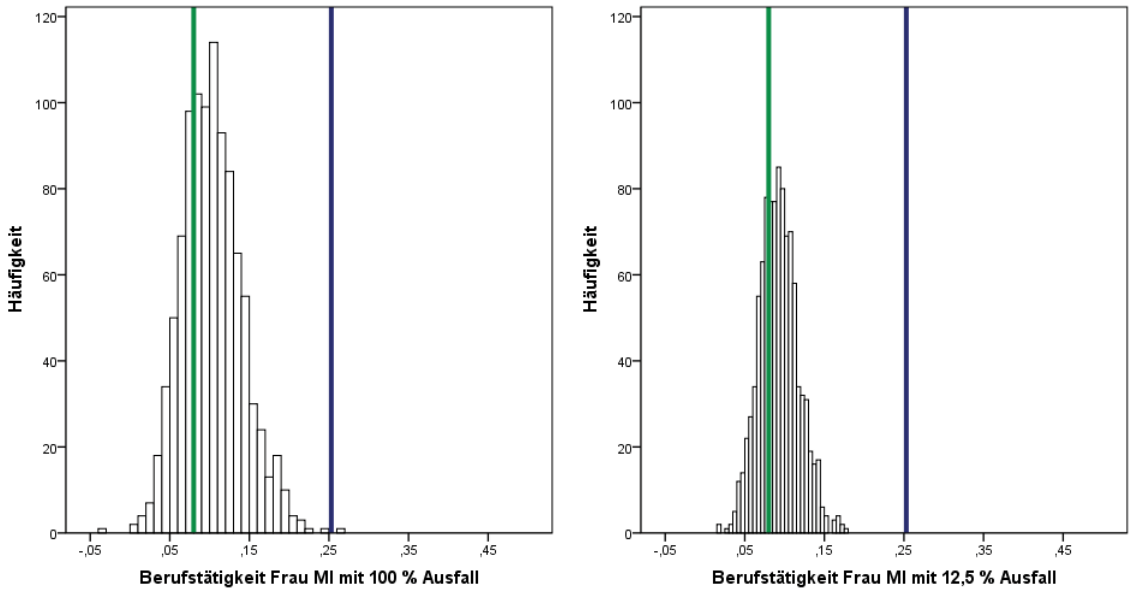



Abbildung 68: MI Histogramm des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen*

An der Lage der Schätzwertverteilungen gibt es kaum einen Unterschied zu erkennen (MI mit 100 %: 0,1033 und MI mit 12,5 %: 0,0927), jedoch ist die Streuung wiederum mit sinkendem Ausfall wesentlich geringer: 0,03773 (100 % Ausfall) zu 0,02474 (12,5 % Ausfall). Die folgende Grafik zeigt sodann den schon ähnlich gesehenen Verlauf (Abbildung 69):

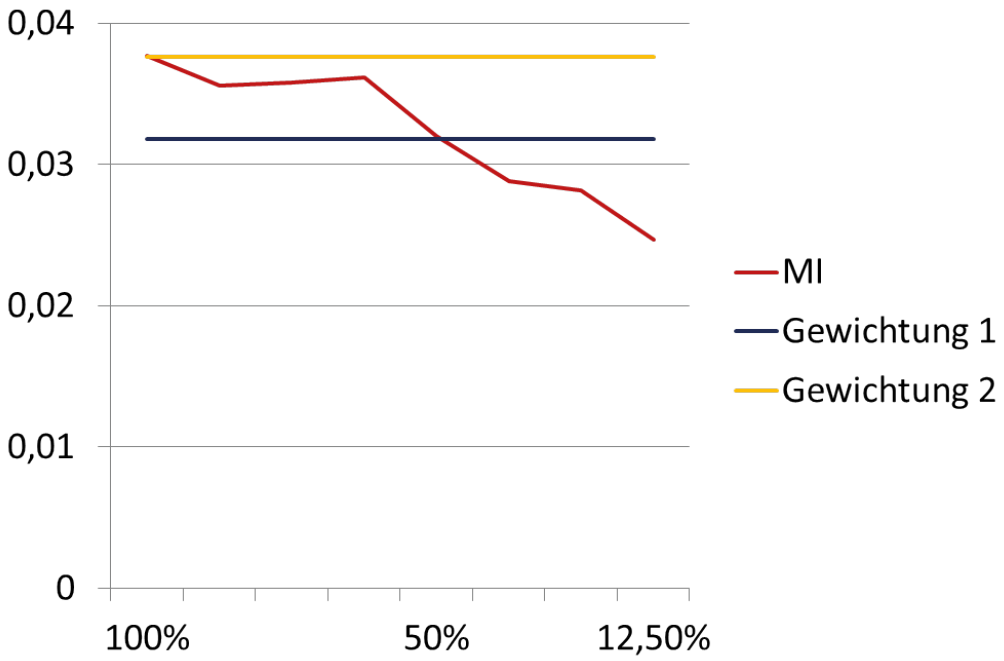



Abbildung 69: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen* unter MI mit sinkendem Ausfall

Bei etwa 40 % bis 50 % Ausfall unterschreitet die Standardabweichung der mit MI geschätzten Parameterwerte in Abbildung 68 die Standardabweichung der Schätzwerte der Gewichtung 1. Die Standardabweichung bei Gewichtung 2 wird schon mit ca. 90 % unterschritten. Die Schätzwerte von Gewichtung 1 und Gewichtung 2 besitzen folgende Verteilungen:

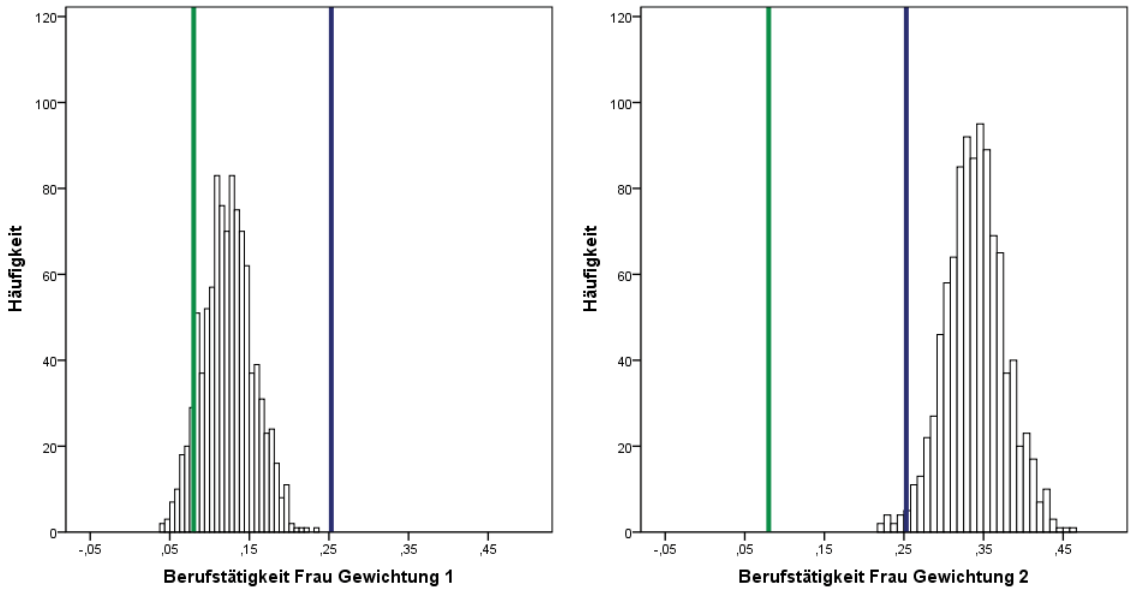



Abbildung 70: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Frau soll Karriere des Mannes unterstützen*

Schnell wird auch hier ersichtlich, dass die Lage der Verteilung im Vergleich variiert. Während Gewichtung 1 eine Verteilung um den durchschnittlichen Schätzwert von 0,1238 aufweist, liegt der durchschnittlich geschätzte Wert nach Gewichtung 2 bei 0,3396 (Abbildung 70).
Was bedeutet das für die möglichen MSE, für die hier zwei Vergleichspunkte als realistisch erachtet wurden – nämlich Extremtyp 2 und der gemäßigte Typ?

**Extremtyp 1**

Gew. 1: 0,0177 < MI 100 %: 0,0220 < MI 12,5 %: 0,0263 < Gew. 2: 0,0893

**wahrer Wert unter MAR**

MI 12,5 %: 0,0008 < Gew. 1: 0,0029 < MI 100 %: 0,0112 < Gew. 2: 0,0689

Für diese unabhängige Variable würde Gewichtung 2 stets die schlechtesten Schätzwerte nach sich ziehen. Gewichtung 1 hingegen schafft für den Extremtyp 1 den geringsten MSE. Allerdings ist der Abstand zu den beiden dicht beieinanderliegenden MSE der MI nicht allzu groß. Insgesamt liegen für den „wahren“ Wert unter MAR die jeweiligen MSE eng beieinander mit einem Vorteil für MI unter einem Ausfall von 12,5 %.
Das bisher einheitliche Bild der Ergebnisse setzt sich für die Variablen *Wahrscheinlichkeit Union zu wählen* und *Wahrscheinlichkeit SPD zu wählen* teilweise fort. Dies liegt zum einen an den „wahren“ Werten, die auf Vorstellungen von Extremtypen basieren. Das heißt, auch hier werden die Grenzen des Stresstests bei Unit Nonresponse deutlich. Daneben liegt es aber erstens an der

Höhe des Ausfalls und zweitens wohl am relativ geringen Beitrag, den die Prädiktoren zur Korrektur der Unit Nonresponse leisten können.

Die Gestalt der MI Schätzwerte für die beiden Variablen *Wahrscheinlichkeit Union zu wählen* und *Wahrscheinlichkeit SPD zu wählen* variiert aufgrund der höheren Anzahl an Ausprägungen mit steigendem Ausfall teilweise stärker. Für beide Parameter wurden zwei „wahre“ Werte ausgewählt: der des Extremtyps 1 und ein Wert unter MAR. Ausgehend von diesen beiden „wahren“ Werten ergeben sich folgende MSE:

*Wahrscheinlichkeit CDU zu wählen*

**Extremtyp 1**

Gew. 1: 0,0041 < MI 12,5 %: 0,0056 < MI 100 %: 0,0153 < Gew. 2: 0,0832

**wahrer Wert unter MAR**

Gew. 1: 0,00005 < MI 12,5 %: 0,0003 <Gew. 2: 0,0010 < MI 100 %: 0,0033

*Wahrscheinlichkeit SPD zu wählen*

**Extremtyp 1**

Gew. 2: 0,0019 < Gew. 1: 0,0064 < MI 12,5 %: 0,0080 < MI 100 %: 0,0085

**wahrer Wert unter MAR**

MI 12,5 %: 0,00012 < MI 100 %: 0,00014 < Gew. 2: 0,00030 < Gew. 1: 0,00036

Bei diesen beiden Variablen und den beiden Vergleichspunkten hat MI mit 100 % Ausfall kein einziges Mal den niedrigsten MSE; unter einem Ausfall von 12,5 % besitzt MI einmal den niedrigsten MSE. Zweimal trifft dies für Gewichtung 1 zu und einmal für Gewichtung 2. Innerhalb der Gewichtungen führt – wie schon mehrfach beobachtet – die komplexere Gewichtung 2 zu höheren MSE als Gewichtung 1. MI mit 100 % Ausfall ist aufgrund der höheren Unsicherheit mit einer größeren Varianz belastet und deshalb, wie auch bei den Parametern zuvor, immer schlechter als MI mit einem Ausfall von nur 12,5 %. Es sollte aber bedacht werden, dass gerade im Vergleich mit dem „wahren“ Wert unter MAR die jeweiligen MSE insgesamt sehr gering sind.
Die Zwischenstufen der Ausfälle mit den Standardabweichungen sind in der folgenden grafischen Darstellung aufbereitet (Abbildung 71 und 72); auch die Standardabweichungen der beiden Gewichtungen sind vermerkt.

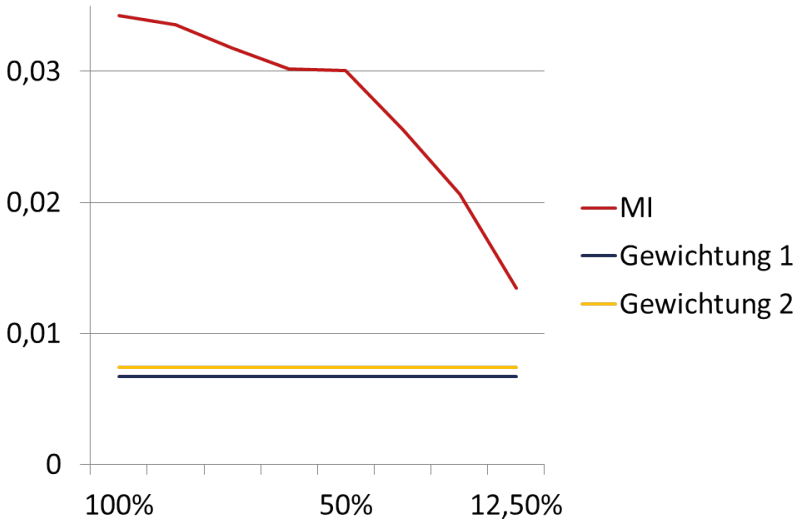



Abbildung 71: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen* unter MI mit sinkendem Ausfall

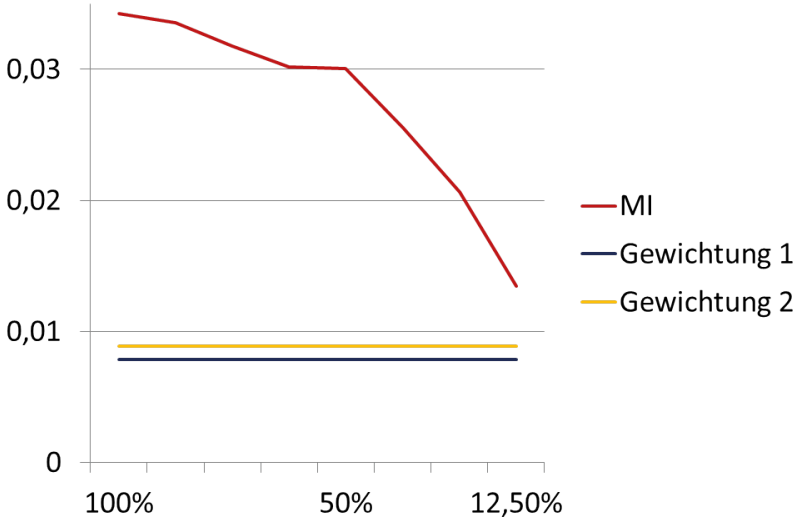



Abbildung 72: Veränderung der Standardabweichung der geschätzten Werte des Parameters der Variablen *Wahrscheinlichkeit SPD zu wählen* unter MI mit sinkendem Ausfall

#### 6.3.4.4 Ergebnis 4: Logitmodell

Die zweite multivariate Schätzung besteht in einem Logitmodell. Die Variable *Wahl der Union* soll mit drei Determinanten erklärt werden. Für den Achsenabschnitt wurden als „wahre“ Werte die Parameterwerte der beiden Extremtypen (rote bzw. blaue Linie) und des Wertes unter MAR (grüne Linie) ausgewählt. Betrachtet man die Schätzwerte, die mit MI berechnet wurden, zeigt sich gerade beim Achsenabschnitt, wie breit sich die Unsicherheit bei 100 % Ausfall in einer hohen Variation der Schätzwerte ausdrückt (Abbildung 73):

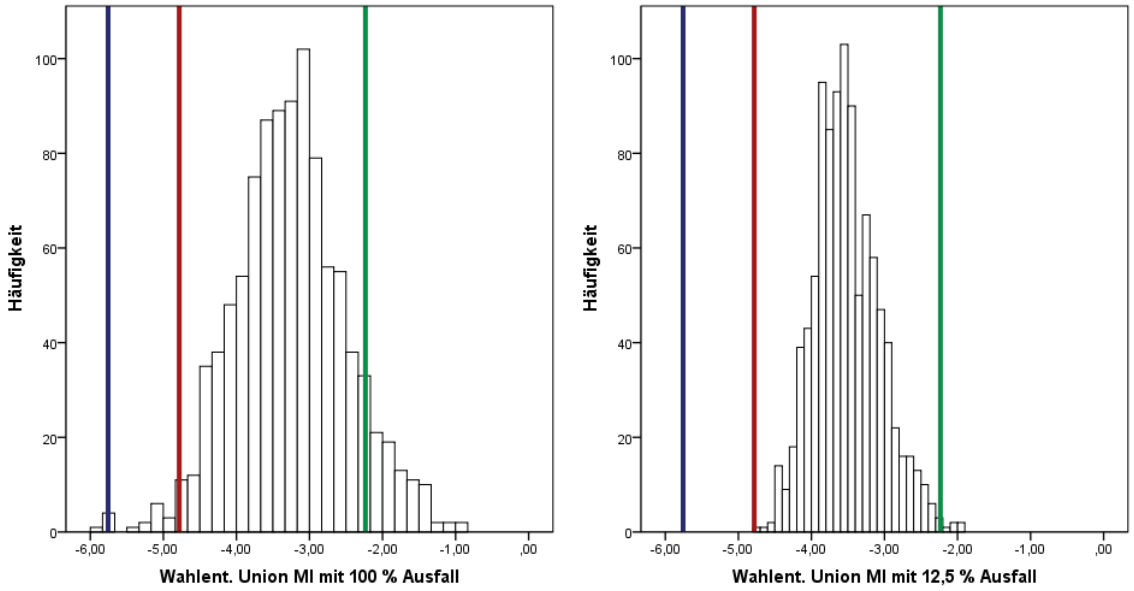



Abbildung 73: MI Histogramm des Achsenabschnitts des Logitmodells

Die Verteilung der Schätzwerte mit einem Ausfall von 100 % gruppiert sich um einen durchschnittlichen Wert von -3,2623 mit einer Standardabweichung von 0,76333. Der durchschnittlich geschätzte Wert bei einem Ausfall von 12,5 % beträgt -3,5132 und streut mit einer Standardabweichung von nur 0,45254. Das bedeutet, dass sich die Streuung mit sinkendem Ausfall um gut 40 % verringert.

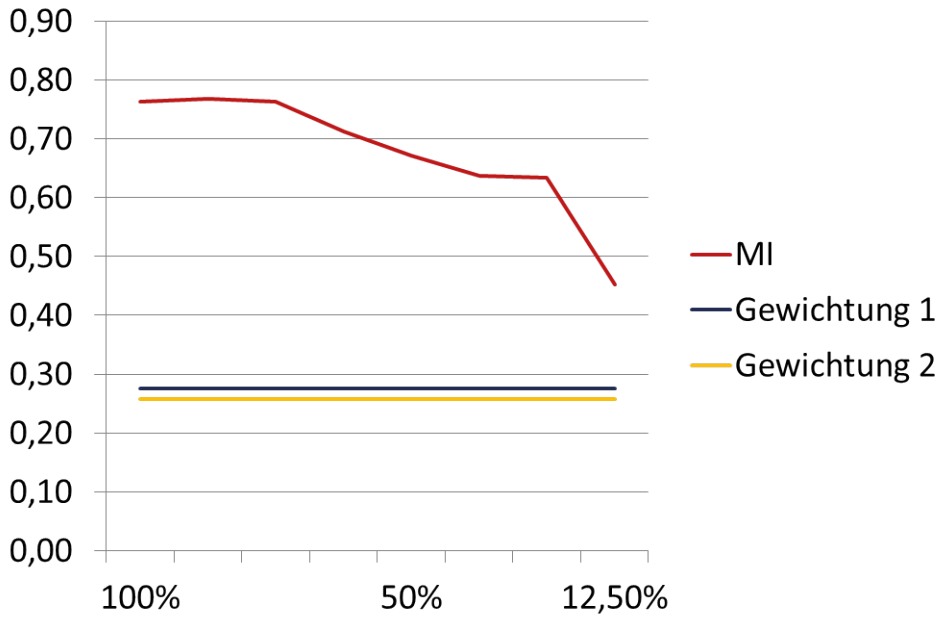



Abbildung 74: Veränderung der Standardabweichung der geschätzten Werte des Achsenabschnitts unter MI mit sinkendem Ausfall

In Abbildung 74 sind auch schon die Standardabweichungen der beiden Gewichtungen eingezeichnet. Die Lage der Verteilungen variiert stark: so ist der durchschnittlich geschätzte Achsenabschnitt bei Gewichtung 1 -3,5474, bei Gewichtung 2 beträgt er -2,3757 (Abbildung 75).

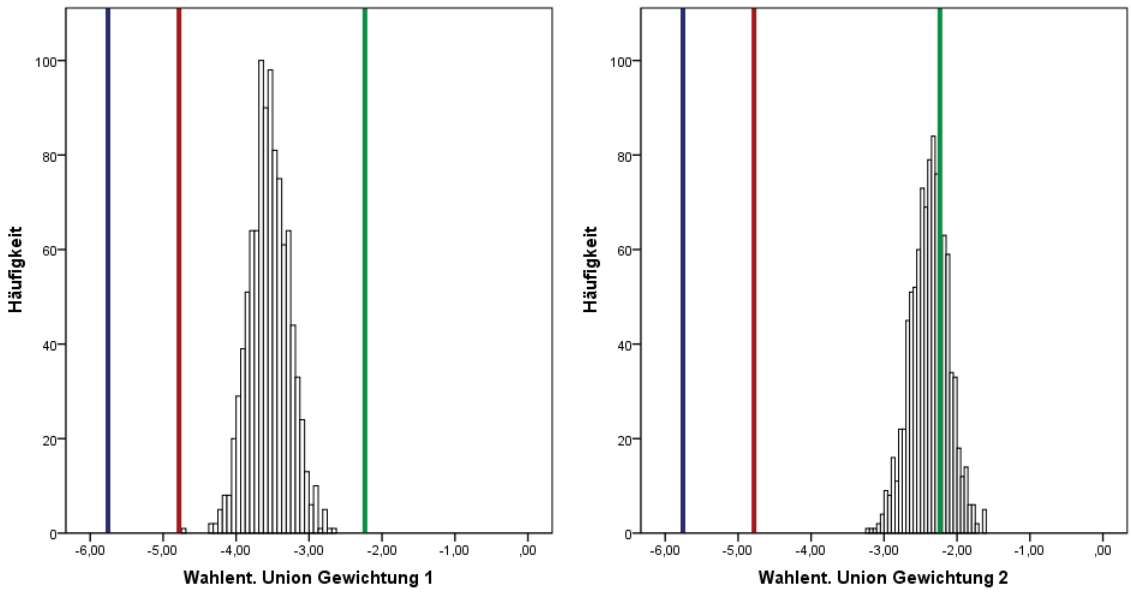



Abbildung 75: Gewichtung 1 und 2: Histogramm des Achsenabschnitts des Logitmodells

Mit diesen Informationen werden auf Grundlage der drei „wahren“ Werte die MSE berechnet.

**Extremtyp 1**

Gew. 1: 1,6032 < MI 12,5 %: 1,8177 < MI 100 %: 2,8958 < Gew. 2: 5,8626

**Extremtyp 2**

Gew. 1: 4,9566 < MI 12,5 %: 1,8177 < MI 100 %: 2,8958 < Gew. 2: 11,4971

**wahrer Wert unter MAR**

Gew. 1: 1,7334 < MI 12,5 %: 1,8437 < MI 100 %: 1,6421 < Gew. 2: 0,0200

Gewichtung 1 schneidet knapp vor MI mit sehr niedrigem Ausfall ab. Von der Höhe des MSE befindet sich MI mit 100 % Ausfall in der Mitte zwischen Gewichtung 1 und Gewichtung 2, die bei den beiden Extremtypen und dem „wahren“ Wert unter MAR mit Abstand den höchsten MSE aufweist. Diese Reihenfolge ist bei allen verwendeten Typen gleich.

Für den nächsten Parameter, der den Einfluss der Variable *Wahrscheinlichkeit CDU zu wählen* auf die abhängige Variable abbilden soll, war nur der Wert unter MAR als „wahrer“ Wert visuell darstellbar und dann für die Analyse sinnvoll. Bereits bei der Auswahl der „wahren“ Werte des Achsenabschnitts war schnell ersichtlich, dass zu extreme „wahre“ Werte die MSE sprengen.

Die Verteilungen der Schätzwerte für diesen Parameter sehen in der Zusammenschau über MI und die Gewichtungen folgendermaßen aus (Abbildung 76 und 77):

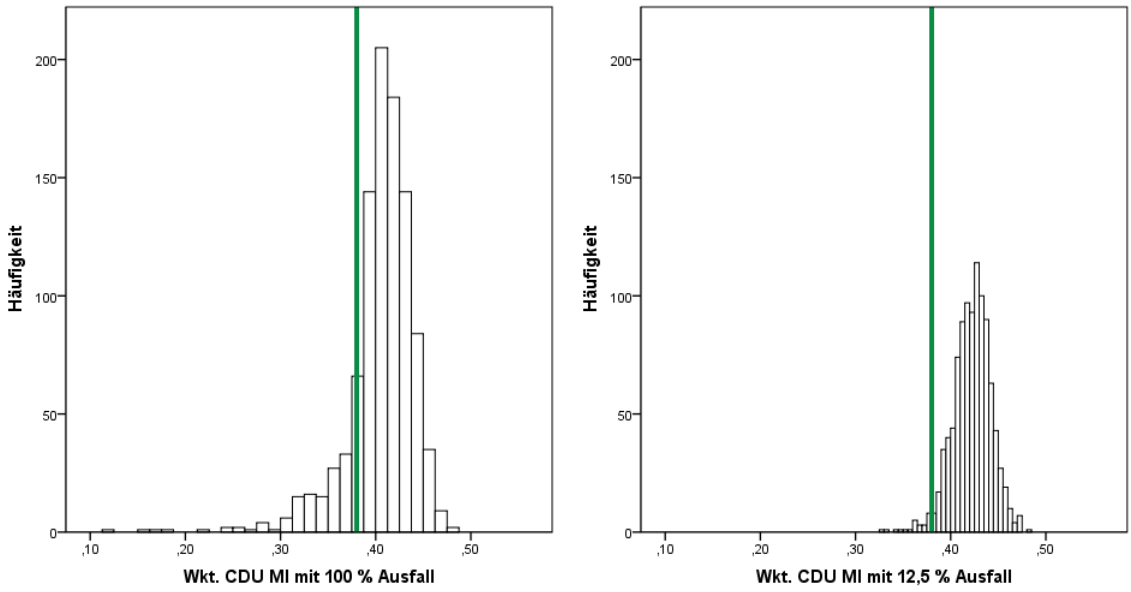



Abbildung 76: MI Histogramm des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen*

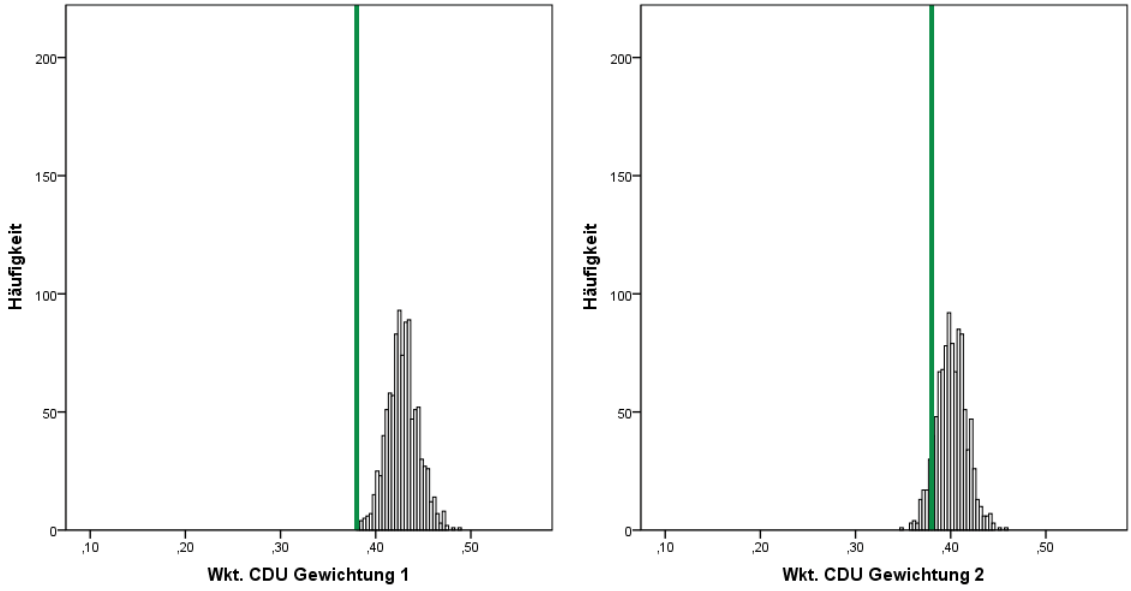



Abbildung 77: Gewichtung 1 und 2: Histogramm des Parameters der Variablen *Wahrscheinlichkeit CDU zu wählen*

Deutlich kann man aus den Grafiken die leichte Überschätzung der Parameterhöhe aller Korrekturmethoden herauslesen. Die Unterschiede in der Breite der Verteilungen wiederholem dagegen schon Gesehenes (der Verlauf befindet sich in Anhang 14). Dies führt insgesamt zu den einzelnen MSE:

**wahrer Wert unter MAR**

Gew. 1: 0,0028 < MI 12,5 %: 0,0022 < MI 100 %: 0,0020 < Gew. 2: 0,0007

Bezogen auf einen „wahren“ Wert unter MAR schneidet MI mit 12,5 % Ausfall und Gewichtung 1 wieder am besten ab. Dies ist nicht verwunderlich, wenn man bedenkt, dass diese beiden Verfahren dann am wenigsten an dem Ist-Zustand mit CC ändern. MI mit 100 % und Gewichtung 2 bringen dagegen Schätzwerte hervor, die eine stärkere Verzerrung bezogen auf MAR nach sich ziehen.

Die beiden noch ausstehenden Parameterschätzungen, die den Einfluss der Variable *Frau mit Kindern berufstätig* und *Nationalstolz* messen sollen, werden im Anhang grafisch umgesetzt. Zur Berechnung des MSE dienen zum einem Extremwert 2 und Wert unter MAR als „wahre“ Werte.

*Frau mit Kindern berufstätig*

**Extremtyp 1**

MI 12,5 %: 0,0017 < Gew. 1: 0,0039 < MI 100 %: 0,0099 < Gew. 2: 0,0131

**wahrer Wert unter MAR**

Gew. 1: 0,0176 < MI 12,5 %: 0,0318 < MI 100 %: 0,0491 < Gew. 2: 0,0760

*Stolz Deutscher zu sein*

**Extremtyp 1**

MI 12,5 %: 0,0734 < Gew. 1: 0,0748 < MI 100 %: 0,1150 < Gew. 2: 0,1654

**Extremtyp 2**

Gew. 1: 0,0597 < MI 100 %: 0,0604 < MI 12,5 %: 0,0989 < Gew. 2: 0,1433

**wahrer Wert unter MAR**

Gew. 1: 0,0034 < MI 12,5 %: 0,0097 < Gew. 2: 0,0132 < MI 100 %: 0,0269

Für diese beiden Determinanten weisen entweder Gewichtung 1 oder die MI-Schätzungen mit 12,5 % Datenausfall den geringsten MSE auf. Den höchsten MSE besitzt Gewichtung 2 oder MI mit 100 % Datenausfall.
Die anderen grafischen Darstellungen sowie die Verläufe, wie sich die Streuung unter verschieden hohen Ausfallhöhen entwickelt, finden sich im Anhang 14.

#### 6.3.4.5 Zusammenfassung

Die Ergebnisse für die Korrektur von Unit Nonresponse fallen wesentlich ernüchternder aus als für die Korrektur von Item Nonresponse. Während bei Item Nonresponse mit einzelnen Einschränkungen eine ziemlich klare Empfehlung zur Verwendung von MI gemacht werden konnte, ist dies für Unit Nonresponse nicht ganz eindeutig, auch wenn sich zum Teil das Muster fortsetzt, dass im univariaten Fall MI leicht im Nachteil ist. Zunächst kann für die Gewichtungen konstatiert werden: Es bestätigte sich die Kritik, dass der Effekt von Gewichtungen mitunter nicht vorhersehbar ist. So kann eine komplexere Gewichtung mit zusätzlichen Gewichtungsmerkmalen zu schlechteren Ergebnissen führen als eine weniger komplexe Gewichtung.
Leider stellt für diese Datenlage MI noch keine wirklich überzeugende Alternative dar. Zwar schneidet MI zusammen mit Gewichtung 1 unter Berücksichtigung eines nicht allzu hohen Ausfalls häufig gut ab. Die Ergebnisse sind jedoch nicht durchgehend positiv. Vollends ausgeschlossen erscheint die Berücksichtigung vollen Ausfalls (100 %), da die Variation der geschätzten Werte in keinem Verhältnis zum Nutzen steht. Freilich ist dieser Effekt bei univariaten Parametern (Anteilswert, Mittelwert) deutlich größer als bei multivariaten Modellen (OLS-, Logitmodell). In wenigen Fällen sank die Streuung der mit MI geschätzten Parameterwerte unter die der Gewichtungen. Die verhältnismäßig große Varianz wirkte sich oft negativ auf den jeweiligen MSE aus. Sie führte häufig zu einem höheren MSE als bei den Gewichtungen (Gewichtung 1), obwohl der Bias genauso hoch oder sogar niedriger als bei den beiden Gewichtungen lag. Die Ergebnisse der Unit Nonresponse-Beispiele stehen teilweise im Schatten des neuartigen Verfahrens, mit Extremtypen zu arbeiten. Bei Item Nonresponse konnte ja die Empfehlung gegeben werden, für den individuellen Fall das Imputationsmodell zu erweitern. So könnten zusätzliche Informationen für MI aus der Fülle der umfangreichen Erhebung herangezogen und damit die Ergebnisse für das Imputationsmodell verbessert werden. Für Unit Nonresponse ist dies nicht ohne Weiteres möglich, da die weitgehend monotone Ausfallstruktur nur die Erweiterung durch zusätzliche Variablen, etwa zum Erhebungsvorgang oder der Interviewersituation, zulässt. Die momentan – und das nur mit Mühe – erhältlichen Informationen sind mit Blick auf die Ergebnisse noch nicht ausreichend, um MI zur empfohlenen Korrekturmethode auch bei Unit Nonresponse zu erklären. Für eine vertiefte Analyse sollte erwogen werden, in irgendeiner Form a priori Informationen in die Imputation einzubeziehen.

## 6.3 Zwischenfazit Unit Nonresponse

Die Ursachen von Unit Nonresponse können von den Anfängen des Erhebungsprozesses, der Fragestellung, bis hin zur Editierung reichen (Abschnitt 5.1). Das Problem Unit Nonresponse wird dadurch verstärkt, dass – anders als bei Item Nonresponse – hinter den Ausfallgründen vollkommen andere Determinanten stehen und damit die Einteilung in Nichterreichbare, Befragungsunfähige und Verweigerer als sinnvoll erscheint. Eine umfangreiche Literatur zu theoretischen Aspekten ist die Folge. Aus dieser breiten Palette konnte für die Erstellung eines Erklärungsmodells geschöpft werden. Es erwies sich dabei wiederum als sinnvoll, nach den Hauptausfallgründen getrennte Modelle zu schätzen (Abschnitt 5.3.2). Zuvor war es unerlässlich, auf die Diskussion um die Entwicklung der Ausfallquote einzugehen und in Bezug zur ALLBUS-Erhebung zu setzen (Abschnitt 5.2.2.1). Der Trend für den ALLBUS geht dabei eindeutig zu einer niedrigeren Ausschöpfungsquote, auf deren Definition auch eingegangen wurde, im Zeitraum 1980-2008. Verschiedene Determinanten nach den einzelnen Ausfallgründen zeichneten dabei große Unterschiede in den Entwicklungen der Ausfälle: die Vermutung steigender Ausfälle lässt sich so zur Aussage steigender Verweigerungsraten präzisieren. Auch bestätigte sich, dass Faktoren der Feldphase sowie das erhebende Institut die Auschöpfungsquote beeinflussen (Abschnitt 5.2.2.2).
Der dann herangezogene Datensatz, der ALLBUS 2008 Pre_release, wies mit mehr als 60 % Ausfall die bis dato niedrigste Ausschöpfung auf. Für diesen Datensatz wurden getrennt nach Ausfallgründen Erklärungsmodelle geschätzt. Gerade für die Verweigerung, die im ALLBUS 2008 den bei weitem größten Ausfall verursachte, ergab sich eine vergleichsweise niedrige Modellgüte, während sich für Nichterreichbarkeit und Befragungsunfähigkeit viele Hypothesen bestätigten (Abschnitt 5.3.2).
Dennoch war der ALLBUS 2008 die Grundlage für den Korrekturmethodenvergleich. Obwohl sich die Datensituation ungünstiger gestaltete und beispielsweise auch die Konstruktion eines Stresstests zur Generierung „wahrer“ Werte schwieriger war, wurde anhand zweier unterschiedlich komplexer Gewichtungen und der Multiplen Imputation der Vergleich an ausgewählten Parametern durchgeführt. Dabei stellte sich schnell heraus, dass die Extremtypen bei Unit Nonresponse oft handhabbare Werte überstiegen und die Korrekturmethoden überforderten. Dies führte dazu, dass ein weiterer „wahrer“ Wert unter MAR als Vergleichskriterium herangezogen wurde. MI wurde zudem mit unterschiedlichen Ausfallhöhen geschätzt. Die Ergebnisse fasst das letzten Kapitel zusammen und es deutet zwar nicht auf ein übertrieben negatives Abschneiden gerade von MI hin, jedoch treten Herausforderungen und Probleme der Korrekturmethoden an sich als auch ihrer Evaluation, wie sie hier neuartig angewendet wurde, zu Tage.

# 7 Keine Angst vor Problemen mit Zähnen

Auf den letzten 182 Seiten wurde das Problem fehlender Werte in sozialwissenschaftlichen Umfragen eingehend erörtert, ihre Ursachen theoretisch diskutiert und empirisch analysiert; schließlich wurden Korrekturmöglichkeiten verglichen. Das Problem der fehlenden Werte hat tatsächlich Zähne, und diese Zähne wurzeln tief im Survey Lifecycle. Deshalb war es ein fruchtbarer Schritt, die Ursachen von Item und Unit Nonresponse im gesamten Kontext des Suvey Lifecycle zu betrachten. Dies ermöglichte die Einbindung z.T. fragmentarischer Theorie in die gesamte Fragestellung und deren Systematisierung. Darüber hinaus waren die theoretischen Überlegungen der Schlüssel zur Einschätzung des Ausfallmechanismus, der wiederum ausschlaggebend für die Auswahl der geeigneten Korrekturmethoden war. Für Item Nonresponse hat sich selbst unter extremen Werten insgesamt die Multiple Imputation als beste Korrekturmethode herausgestellt – zumindest gilt dies für die Schätzung multivariater Parameter. Bei der Analyse der Unit Nonresponse wurde die Notwendigkeit, die Entstehung der Daten genauer zu hinterfragen, noch deutlicher. Dies legt schon das Kapitel zur Diskussion um die Ausschöpfungsquote nahe. Mit Blick auf die Korrekturmethoden erweist sich Unit Nonresponse als das eigentliche Problem mit Zähnen, dessen Lösung für den Datennutzer nicht ohne Weiteres möglich ist. Dennoch hilft die Analyse und Einbeziehung der Daten und deren Entstehungsprozess an dieser Stelle, um Einschätzungen über die tatsächliche Größe des Problems zu erhalten.

Die anfangs festgestellte Kluft zwischen Datenerzeugern, Dateneditierern sowie Datennutzern kann von letzteren zumindest für das Problem Item Nonresponse in realistischem Umfang überbrückt werden. Das bereits erwähnte Ideal des Datenbereitstellers und des Datennutzers ist angesichts der weitgehenden Arbeitsteilung unrealistisch. Jedoch können Korrekturmethoden mittlerweile sinnvoll angewendet werden, die sich durch ausgereiftere Software zur Verwendung anbieten und, wie die Ergebnisse auch gezeigt haben, zu empfehlen sind. Die konstatierte Kluft scheint nach jetzigem Stand für Unit Nonresponse nicht ganz so einfach überbrückbar. Hier muss an die Adresse der Datenerzeuger gesagt werden, dass dringend zusätzliche und nicht unbedingt wesentlich konstenintensivere Informationen – z.B. für den Einsatz der Multiplen Imputation – notwendig sind. Diese Kluft kann vom Datennutzer allein nicht überbrückt werden. Wie wertvoll diese Informationen sein können, hat sich in dieser Arbeit einerseits bei der Analyse von Item und Unit Nonresponse gezeigt, anderseits wurden sie als tragende Teile des Imputationsmodells verwendet. Auch andere Arten der a priori Information sind denkbar und stellen einen realistischen Ansatzpunkt dar, die Herausforderung Unit Nonresponse anzugehen.

Die Systematisierung des Problems fehlender Werte mit Hilfe des Survey Lifecycles zeigt auch deutlich, dass Item und Unit Nonresponse zum einen eine starke Definitionskomponente besitzen (Abschnitt 2.2, Abschnitt 5.1), zum zweiten im Kontext mit anderen möglichen Fehlern einer

Erhebung gesehen werden müssen (Abschnitt 2.3 und 5.1). Dies soll die Zähne des Problems nicht verharmlosen, allerdings hilft es, die Diskussion auf eine praktische Ebene zu befördern und so Auswüchse wie die Verherrlichung der Ausschöpfungsquote nüchtern einzuschätzen. Zudem mahnt es diejenigen, die die Daten herausgeben, zu einer möglichst genauen und detaillierten Dokumentation des Erhebungsprozesses.
Auch wenn bei dieser Arbeit sozialwissenschaftliche Erhebungen im Fokus standen, stehen alle wissenschaftlichen und mehr noch kommerziellen Erhebungen unter ökonomischem Druck. Ein Teil der hier verglichenen Methoden hatte stark fallreduzierende Wirkung. Angesichts des Kostendrucks sollte im Zusammenhang mit dem Einsatz von Postsurvey Adjustments bei der Datenerhebung größere Sorgfalt walten. Zusätzliche Investitionen an dieser Stelle des Survey Lifecycle könnten sich daher im Hinblick auf den Einsatz von MI in der Korrekturphase lohnen.
Mit dieser Arbeit ist das Problem nicht gelöst, seine Zähne nicht gezogen; man muss aber auch keine Angst davor haben. Die Ergebnisse der Arbeit sollten Datennutzern im sozial- und wirtschaftswissenschaftlichen Bereich helfen, fehlende Werte einzuschätzen, zu analysieren und geeignet zu korrigieren.

# Anhang

**Anhang 1:** Liste der für den Item Nonresponse-Vektor verwendeten Variablen des ALLBUS 2006

| Variable | Label |
|---|---|
| v6 | BRAUCHT MAN FAMILIE ZUM GLUECK? |
| v7 | HEIRAT BEI DAUERNDEM ZUSAMMENLEBEN |
| v8 | LAGEVERSCHLECHTERUNG FUER EINFACHE LEUTE |
| v9 | BEI DIESER ZUKUNFT KEINE KINDER MEHR |
| v10 | POLITIKER UNINTERESSIERT AN EINF.LEUTEN |
| v11 | MEHRHEIT UNINTERESSIERT AN MITMENSCHEN |
| v12 | ABTREIB.- WENN WAHRSCH.BABY NICHT GESUND |
| v13 | ABTREIB.- VERH.FRAU,KEINE KINDER MEHR |
| v14 | ABTREIB.- BEI GESUNDHEITSGEFAEHRD.D.FRAU |
| v15 | ABTREIB.- BEI FINANZ.NOTLAGE DER FAMILIE |
| v16 | ABTREIBUNG- NACH VERGEWALTIGUNG |
| v17 | ABTREIB.- LEDIGE MUTTER, OHNE EHEWUNSCH |
| v18 | ABTREIBUNG - WENN DIE FRAU ES WILL |
| v19 | SUBJEKTIVE SCHICHTEINSTUFUNG, BEFR. |
| v20 | GERECHTER ANTEIL A.LEBENSSTANDARD,BEFR.? |
| v21 | ZUZUG VON: AUSSIEDLERN AUS OSTEUROPA |
| v22 | ZUZUG VON: ASYLSUCHENDEN |
| v23 | ZUZUG VON: EU-ARBEITNEHMERN |
| v24 | ZUZUG VON: NICHT-EU-ARBEITNEHMERN |
| v25 | GEBURTSMONAT: BEFRAGTE |
| v26 | GEBURTSJAHR: BEFRAGTE |
| v27 | ALTER: BEFRAGTE |
| v28 | ALTER: BEFRAGTE, KATEGORISIERT |
| v36 | BEFR.: HERKUNFTSLAND |
| v37 | BUNDESLAND, WO BEFRAGTER IN JUGEND LEBTE |
| v38 | LAND, WO BEFRAGTER IN DER JUGEND LEBTE |
| v39 | IMMIGRANT: SEIT WANN IN DEUTSCHLAND,JAHR |
| v40 | IMMIGRANT: SEIT WANN IN DEUTSCHLAND,KAT. |
| v41 | IMMIGRANT: WIEVIEL JAHRE IN DEUTSCHLAND? |
| v43 | AUSLAENDER: MEHR LEBENSSTILANPASSUNG |
| v44 | AUSLAEND.:WIEDER HEIM BEI KNAPPER ARBEIT |
| v45 | AUSLAENDER: POLIT.BETAETIGUNG UNTERSAGEN |
| v46 | AUSLAENDER: SOLLTEN UNTER SICH HEIRATEN |
| v48 | AUSLAENDER: KONTAKT I.D.EIGENEN FAMILIE? |
| v49 | AUSLAENDER: KONTAKT BEI DER ARBEIT? |
| v50 | AUSLAENDER: KONTAKT IN D. NACHBARSCHAFT? |
| v51 | AUSLAENDER: KONTAKT IM FREUNDESKREIS? |
| v52 | GENERELLER STOLZ, DEUTSCHER ZU SEIN |
| v53 | SCHLUSSSTRICH UNTER NAZIZEIT? |

| | |
|---|---|
| v54 | BEFR.: INFORMIERE MICH VOR WAHL |
| v55 | BEFR.: BIN MANCHMAL BELEIDIGT |
| v56 | BEFR.: BIN EIN GUTER ZUHOERER |
| v57 | BEFR.: HABE SCHON KRANK GEFEIERT |
| v58 | BEFR.: HABE SCHON PERSON AUSGENUTZT |
| v59 | BEFR.: KANN EIGENE FEHLER ZUGEBEN |
| v60 | BEFR.: TUE SELBST DAS, WAS ICH FORDERE |
| v61 | BEFR.: HOEFLICH ZU UNANGENEHMEN LEUTEN |
| v62 | BEFR.: AERGER UEBER BITTE UM GEFALLEN |
| v63 | BEFR.: NIE ABSICHTLICH GEFUEHLE VERLETZT |
| v64 | DANKBAR SEIN FUER FUEHRENDE KOEPFE |
| v65 | ANPASSUNG ALS KIND SPAETER NUETZLICH |
| v66 | FREMDER IM EIGENEN LAND DURCH AUSLAENDER |
| v67 | AUSLAENDER TUN DIE UNSCHOENEN ARBEITEN |
| v68 | AUSLAENDER BELASTEN UNSER SOZIALES NETZ |
| v69 | AUSLAENDER BEREICHERN UNSERE KULTUR |
| v70 | AUSLAENDER VERKNAPPEN WOHNUNGEN |
| v71 | AUSLAENDER STUETZEN DIE RENTENSICHERUNG |
| v72 | AUSLAENDER NEHMEN ARBEITSPLAETZE WEG |
| v73 | AUSLAENDER BEGEHEN HAEUFIGER STRAFTATEN |
| v74 | AUSLAENDER SCHAFFEN ARBEITSPLAETZE |
| v75 | VORKOMMEN: WIRT DISKRIMINIERT AUSLAENDER |
| v76 | VORKOMMEN: ELTERN DISKRIMINIEREN TUERKEN |
| v77 | VORKOMMEN: UNTERNEHMER DISKRIMINIEREN |
| v78 | MEINUNG: WIRT DISKRIMINIERT AUSLAENDER |
| v79 | MEINUNG: ELTERN DISKRIMINIEREN TUERKEN |
| v80 | MEINUNG: UNTERNEHMER DISKRIMINIEREN |
| v81 | ALLGEM. MEINUNG: DISKRIMINIERENDER WIRT |
| v82 | ALLGEM. MEINUNG: DISKRIMINIERENDE ELTERN |
| v83 | ALLG.MEINUNG: DISKRIMINIERENDER UNTERN. |
| v84 | ALLG.BEWERTUNG: AUSL.LEBENSSTIL ANPASSEN |
| v85 | ALLG.BEWERTUNG: AUSL.LEBENSSTIL BEHALTEN |
| v86 | ALLG.BEWERTUNG: AUSL.HEIM KNAPPER ARBEIT |
| v87 | ALLG.BEWERTUNG: BLEIBERECHT OHNE ARBEIT |
| v88 | ALLG.BEWERTUNG: AUSL.KEINE POLIT. AKTION |
| v89 | ALLG.BEWERTUNG: AUSL.POL.AKTION ERLAUBEN |
| v90 | ALLG.BEWERTUNG: AUSL.UNTER SICH HEIRATEN |
| v91 | ALLG.BEWERTUNG: AUSL. DEUTSCHE HEIRATEN |
| v92 | AUSLAENDERBEHANDLUNG DURCH BEHOERDEN |
| v93 | EINBUERGERUNG: SOLLTE HIER GEBOREN SEIN |
| v94 | EINBUERGERUNG: DEUTSCHE ABSTAMMUNG HABEN |
| v95 | EINBUERGERUNG: DEUTSCH SPRECHEN |
| v96 | EINBUERGERUNG: LANGE BEI UNS GELEBT |
| v97 | EINBUERGERUNG: LEBENSSTILANPASSUNG |
| v98 | EINBUERGERUNG: IN CHRISTLICH.KIRCHE SEIN |

| | |
|---|---|
| v99 | EINBUERGERUNG: KEINE STRAFTATEN |
| v100 | EINBUERGERUNG: EIGENER LEBENSUNTERHALT |
| v101 | EINBUERGERUNG: ZU GRUNDGESETZ BEKENNEN |
| v102 | DOPPELTE STAATSBUERGERSCHAFT ERLAUBEN |
| v103 | GLEICHE SOZIALLEISTUNGEN FUER AUSLAENDER |
| v104 | KOMMUNALES WAHLRECHT FUER AUSLAENDER |
| v105 | AN SCHULEN AUCH ISLAMUNTERRICHT ERLAUBEN |
| v106 | MEINUNG:ETHNISCH GEMISCHTE NACHBARSCHAFT |
| v108 | AUSLAENDERANTEILSCHAETZUNG WESTEN, KAT. |
| v110 | AUSLAENDERANTEILSCHAETZUNG OSTEN, KAT. |
| v111 | MOECHTE IN WOHNGEBIET 1 LEBEN |
| v137 | AUSLAENDERANTEIL IN EIGENER WOHNUMGEBUNG |
| v138 | LOKALES VERHAELTNIS ZW. AUSL.+DEUTSCHEN |
| v139 | POLITISCHES INTERESSE, BEFR. |
| v140 | WICHTIGKEIT VON RUHE UND ORDNUNG |
| v141 | WICHTIGKEIT VON BUERGEREINFLUSS |
| v142 | WICHTIGKEIT DER INFLATIONSBEKAEMPFUNG |
| v143 | WICHTIGKEIT V. FREIER MEINUNGSAEUSSERUNG |
| v144 | INGLEHART-INDEX |
| v145 | LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR. |
| v146 | WIRTSCHAFTSLAGE IN DER BRD HEUTE |
| v147 | WIRTSCHAFTSLAGE, BEFR. HEUTE |
| v148 | WIRTSCHAFTSLAGE DER BRD IN 1 JAHR |
| v149 | WIRTSCHAFTSLAGE, BEFR. IN 1 JAHR |
| v150 | LEBENSSTILUNTERSCHIED: ITALIENER IN BRD |
| v151 | LEBENSSTILUNTERSCHIED: AUSSIEDLER IN BRD |
| v152 | LEBENSSTILUNTERSCHIED: ASYLBEWERB. I.BRD |
| v153 | LEBENSSTILUNTERSCHIED: TUERKEN IN BRD |
| v154 | LEBENSSTILUNTERSCHIED: JUDEN IN BRD |
| v155 | WIE ANGENEHM ALS NACHBAR: ITALIENER |
| v156 | WIE ANGENEHM ALS NACHBAR: AUSSIEDLER |
| v157 | WIE ANGENEHM ALS NACHBAR: ASYLBEWERBER |
| v158 | WIE ANGENEHM ALS NACHBAR: TUERKE |
| v159 | WIE ANGENEHM ALS NACHBAR: JUDE |
| v160 | EINHEIRAT IN EIGENE FAMILIE: ITALIENER |
| v161 | EINHEIRAT IN EIGENE FAMILIE: AUSSIEDLER |
| v162 | EINHEIRAT IN EIGENE FAMILIE:ASYLBEWERBER |
| v163 | EINHEIRAT IN EIGENE FAMILIE: TUERKE |
| v164 | EINHEIRAT IN EIGENE FAMILIE: JUDE |
| v165 | RECHTSGLEICHSTELLUNG FUER: ITALIENER |
| v166 | RECHTSGLEICHSTELLUNG FUER: AUSSIEDLER |
| v167 | RECHTSGLEICHSTELLUNG FUER: ASYLBEWERBER |
| v168 | RECHTSGLEICHSTELLUNG FUER: TUERKEN |
| v169 | RECHTSGLEICHSTELLUNG FUER: JUDEN |
| v170 | JUDEN HABEN AUF DER WELT ZUVIEL EINFLUSS |

| | |
|---|---|
| v171 | SCHAM UEBER DEUTSCHE UNTATEN AN JUDEN |
| v172 | JUDEN NUTZEN DEUTSCHE VERGANGENHEIT AUS |
| v173 | JUDEN AN VERFOLGUNGEN NICHT UNSCHULDIG |
| v174 | GESCHLECHT |
| v175 | ALLGEMEINER SCHULABSCHLUSS |
| v187 | BEFR.: KEIN BERUFL. AUSBILDUNGSABSCHLUSS |
| v203 | BEFR.: JETZIGE BERUFSUNTERGRUPPE |
| v204 | BEFR.: JETZIGE BERUFSHAUPTGRUPPE |
| v206 | IM OEFFENTLICHEN DIENST TAETIG? |
| v208 | BEFR.: ARBEITSSTUNDEN PRO WOCHE, KATEG. |
| v209 | BERUFLICHE AUFSICHTSFUNKTION, BEFR.? |
| v211 | FURCHT: BETRIEBSVERLUST, SELBSTAENDIGE |
| v212 | BERUFST.: ARBEITSLOS I.D. LETZTEN 10 J.? |
| v214 | DAUER DER ARBEITSLOSIGKEIT, KATEGORIS. |
| v237 | NICHTBERUFST.:EHEDEM ARBEITSLOS GEWESEN? |
| v238 | ARBEITSLOS:EHEDEM ARBEITSLOS GEWESEN? |
| v239 | DAUER <EHEMALIGER> ARBEITSLOSIGKEIT |
| v240 | DAUER <EHEMALIGER> ARBEITSLOSIGKEIT,KAT. |
| v241 | GESUNDHEITSZUSTAND BEFR. |
| v242 | FAMILIENSTAND, BEFRAGTE |
| v243 | GEGENWAERTIGER EHEPARTNER: GEBURTSMONAT |
| v244 | GEGENWAERTIGER EHEPARTNER: GEBURTSJAHR |
| v246 | GEGENWAERTIGER EHEPARTNER: ALTER, KAT. |
| v250 | EHEP.: ZAHL DER STAATSBUERGERSCHAFTEN |
| v251 | EHEP.: VON GEBURT AN DEUTSCH? |
| v252 | EHEP.: URSPRUENGL. STAATSBUERGERSCHAFT |
| v253 | GEGENW.EHEP.: ALLGEMEIN.SCHULABSCHLUSS |
| v265 | GEGENW.EHEP.: KEIN BERUFL.ABSCHLUSS |
| v266 | GEGENWAERTIGER EHEP. BERUFSTAETIG? |
| v267 | GEGENW.EHEP.: JETZIGE BERUFL. STELLUNG |
| v270 | GEGENW.EHEP.: JETZIGER BERUF; ISCO 1988 |
| v284 | EHEP.: IM OEFFENTLICHEN DIENST TAETIG? |
| v288 | LEBENSPARTNER: GEBURTSJAHR |
| v290 | LEBENSPARTNER: ALTER, KAT. |
| v294 | LEBENSP.: ZAHL DER STAATSBUERGERSCHAFTEN |
| v295 | LEBENSPARTNER: VON GEBURT AN DEUTSCH? |
| v296 | LEBENSP.:URSPRUENGL.STAATSBUERGERSCHAFT |
| v297 | LEBENSPARTNER: ALLG.SCHULABSCHLUSS |
| v310 | LEBENSPARTNER: BERUFSTAETIG? |
| v311 | LEBENSPARTNER: JETZIGE BERUFL.STELLUNG |
| v328 | LEBENSP: IM OEFFENTLICHEN DIENST TAETIG? |
| v329 | LEBENSP: STATUS D.NICHTERWERBSTAETIGKEIT |
| v330 | HERKUNFTSLAND: VATER |
| v331 | HERKUNFTSLAND: GROSSVATER, VATERSEITS |
| v332 | HERKUNFTSLAND: GROSSMUTTER, VATERSEITS |

| | |
|---|---|
| v333 | HERKUNFTSLAND: MUTTER |
| v334 | HERKUNFTSLAND: GROSSVATER, MUTTERSEITS |
| v335 | HERKUNFTSLAND: GROSSMUTTER, MUTTERSEITS |
| v336 | ELTERN: DAMALS MIT BEFR. ZUSAMMENGELEBT |
| v337 | VATER: BERUFLICHE STELLUNG |
| v354 | MUTTER: BERUFLICHE STELLUNG |
| v371 | VATER: ALLGEMEINER SCHULABSCHLUSS |
| v372 | MUTTER: ALLGEMEINER SCHULABSCHLUSS |
| v373 | VATER: BERUFSAUSBILDUNG |
| v374 | MUTTER: BERUFSAUSBILDUNG |
| v375 | FAMILIENMEINUNG: ZU VIELE AUSLAENDER |
| v376 | FAMILIE EINIG ZUM THEMA AUSLAENDER |
| v377 | FREUNDE MEINUNG: ZU VIELE AUSLAENDER |
| v378 | FREUNDE EINIG ZUM THEMA AUSLAENDER |
| v379 | BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE |
| v380 | BEFR.: NETTOEINKOMMEN, LISTENABFRAGE |
| v381 | BFR.:NETTOEINKOMMEN <OFFENE+LISTENANGABE> |
| v382 | NETTOEINKOMMEN < OFFENE+LISTENANGABE>,KAT. |
| v383 | MEHRPERSONENHAUSHALT? |
| v384 | MEHR ALS 8 HAUSHALTSPERSONEN? |
| v385 | ANZAHL WEITERER HAUSHALTSPERSONEN |
| v386 | ANZAHL DER HAUSHALTSPERSONEN |
| v485 | MIT KIND: WUNSCH NACH WEITEREN KINDERN? |
| v486 | MIT KIND: WIEVIELE KINDER GEWUENSCHT? |
| v487 | KINDERLOS: WUNSCH NACH KINDERN? |
| v488 | KINDERLOS: WIEVIELE KINDER GEWUENSCHT? |
| v489 | BEFR.: TYP DER WOHNUNG |
| v490 | SELBSTBESCHREIBUNG DES WOHNORTS |
| v491 | IM WESTEN MEHR OPFERBEREITSCHAFT ZEIGEN |
| v492 | IM OSTEN MEHR GEDULD ZEIGEN |
| v493 | WIEDERVEREIN. MEHR VORTEILE FUER WESTEN |
| v494 | WIEDERVEREIN. MEHR VORTEILE FUER OSTEN |
| v495 | ZUKUNFT IM OSTEN HAENGT VON LEISTUNG AB |
| v496 | BUERGER IM ANDEREN TEIL DER BRD FREMD? |
| v497 | NEUE LAENDER: LEISTUNGSDRUCK ZU GROSS? |
| v498 | NICHT NACH STASI-VERGANGENHEIT FRAGEN |
| v499 | SOZIALISMUS: GUTE IDEE, SCHLECHT AUSGEF. |
| v500 | KONFESSION, BEFRAGTE |
| v501 | WELCHE NICHTCHRISTLICHE RELIGION? |
| v502 | KIRCHGANGSHAEUFIGKEIT |
| v503 | MITGLIED IN EINER GEWERKSCHAFT? |
| v504 | FRUEHER GEWERKSCHAFTSMITGLIED? |
| v505 | MITGLIED: POLITISCHE PARTEI |
| v506 | WAHLABSICHT, BUNDESTAGSWAHL; BEFR. |
| v507 | WAHLBETEILIGUNG, LETZTE BUNDESTAGSWAHL? |

**Anhang 2:** Items für Beispiel 1 des Methodenvergleichs bei Item Nonresponse[1]

| Variablen | Labels | Nonresponse-Kategorien |
|---|---|---|
| v8 | LAGEVERSCHLECHTERUNG FUER EINFACHE LEUTE | 8 =„w. n.“ 9 =„k. A.“ |
| v9 | BEI DIESER ZUKUNFT KEINE KINDER MEHR | 8 =„w. n.“ 9 =„k. A.“ |
| v11 | MEHRHEIT UNINTERESSIERT AN MITMENSCHEN | 8 =„w. n.“ 9 =„k. A.“ |
| v53 | SCHLUSSSTRICH UNTER NAZIZEIT? | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v66 | FREMDER IM EIGENEN LAND DURCH AUSLAENDER | 0 =„Trifft nicht zu“ 99 =„k. A.“ |
| v142 | WICHTIGKEIT DER INFLATIONSBEKAEMPFUNG | 8 =„w. n.“ 9 =„k. A.“ |
| v143 | WICHTIGKEIT V. FREIER MEINUNGSAEUSSERUNG | 8 =„w. n.“ 9 =„k. A.“ |
| v170 | JUDEN HABEN AUF DER WELT ZUVIEL EINFLUSS | 99 =„k. A.“ |
| v171 | SCHAM UEBER DEUTSCHE UNTATEN AN JUDEN | 99 =„k. A.“ |
| v172 | JUDEN NUTZEN DEUTSCHE VERGANGENHEIT AUS | 99 =„k. A.“ |
| v173 | JUDEN AN VERFOLGUNGEN NICHT UNSCHULDIG | 99 =„k. A.“ |



[1] „weiß nicht“=„w. n.“; „keine Angabe“=„k. A.“

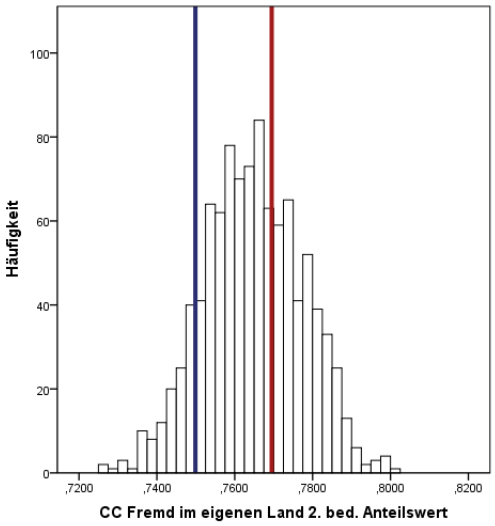
**Anhang 3:** Items für Beispiel 2 des Methodenvergleichs bei Item Nonresponse

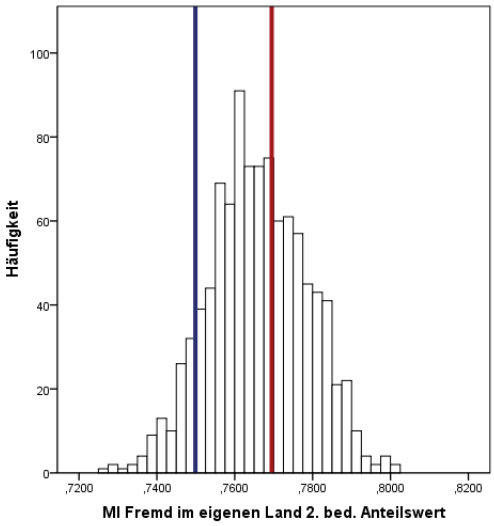
| Variablen | Labels | Nonresponse-Kategorien |
|---|---|---|
| v3 | ERHEBUNGSGEBIET: WEST - OST | - |
| v106 | LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR. | 99= „k. A.“ |
| v182 | GESCHLECHT, BEFRAGTE | - |
| v185 | ALTER: BEFRAGTE | 999 =„k. A.“ |
| v187 | ALLGEMEINER SCHULABSCHLUSS | 9 =„keine Angabe“ |
| v196 | BEFR.: FACHHOCHSCHULABSCHLUSS | 6 =„Trifft nicht zu“ 9 =„k. A.“ |
| v197 | BEFR.: HOCHSCHULABSCHLUSS | 6 =„Trifft nicht zu“ 9 =„k. A.“ |
| v201 | BEFR.: JETZIGE BERUFLICHE STELLUNG | 0 =„Trifft nicht zu“ 99 =„k. A.“ |
| v213 | IM OEFFENTLICHEN DIENST TAETIG? | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v217 | BERUFST.: ARBEITSLOS I.D. LETZTEN 10 J.? | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v220 | BEFR.: STATUS DER NICHTERWERBSTAETIGKEIT | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v235 | ARBEITSLOS:EHEDEM ARBEITSLOS GEWESEN? | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v236 | NICHTBERUFST.:EHEDEM ARBEITSLOS GEWESEN? | 0 =„Trifft nicht zu“ 9 =„k. A.“ |
| v306 | VATER: BERUFLICHE STELLUNG | 96 =„Vater unbekannt“ 98 =„w. n.“ 99 =„k. A.“ |
| v318 | VATER: ALLGEMEINER SCHULABSCHLUSS | 0 =„Trifft nicht zu“ 8 =„w. n.“ 9 =„k. A.“ |
| v320 | MUTTER: ALLGEMEINER SCHULABSCHLUSS | 8 =„w. n.“ 9 =„k. A.“ |
| v517 | MITGLIED IN EINER GEWERKSCHAFT? | 9 =„k. A.“ |
| v519 | FRUEHER GEWERKSCHAFTSMITGLIED? | 0 =„Trifft nicht zu“ 7 =„verweigert“ 9 =„k. A.“ |

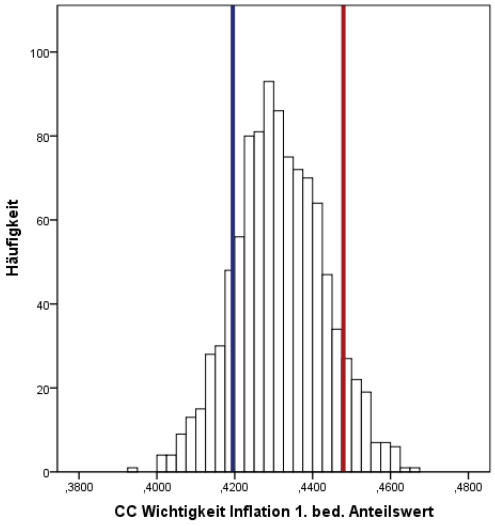
**Anhang 4:** Items für Beispiel 3 des Methodenvergleichs bei Item Nonresponse

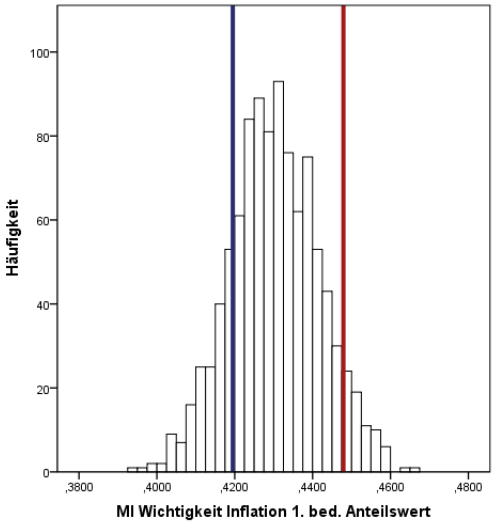
| Variablen | Labels | Nonresponse-Kategorien |
|---|---|---|
| v3 | ERHEBUNGSGEBIET: WEST - OST | - |
| v55 | GESCHLECHT, BEFRAGTE¡R¿ | - |
| v58 | ALTER: BEFRAGTE¡R¿ | 999 =„k. A.“ |
| v60 | ALLGEMEINER SCHULABSCHLUSS | 9 =„k. A.“ |
| v182 | POLITISCHES INTERESSE, BEFR. | 9 =„k. A.“ |
| v811 | POL.AKT.: UNTERSCHRIFTENSAMMLUNG | 6 =„KEIN ISSP“ 8=„w. n.“ 9 =„k. A.“ |
| v812 | POL.AKT.: KRITISCHER KONSUM | 6 =„KEIN ISSP“ 8 =„w. n.“ 9 =„k. A.“ |
| v813 | POL.AKT.: DEMONSTRATION | 6 =„KEIN ISSP“ 8 =„w. n.“ 9 =„k. A.“ |
| v814 | POL.AKT.: POLITISCHE VERSAMMLUNG | 6 =„KEIN ISSP“ 8 =„w. n.“ 9 =„k. A.“ |
| v815 | POL.AKT.: KONTAKT MIT POLITIKER, BEAMTEM | 6 =„KEIN ISSP“ 8 =„w. n.“ 9 =„k. A.“ |
| v816 | POL.AKT.: GELD GESPENDET ODER GESAMMELT | 6 =„KEIN ISSP“ 8 =„w. n.“ 9 =„k. A.“ |
| v482 | REDUZIERTE HAUSHALTSGROESSE | 99 =„k. A.“ |
| v553 | MEHRPERS.HAUSH.:EINKOMMEN <OFFENE ABFR.> | 99997 =„verweigert“ 99999 =„k. A.“ |
| v854 | WIE GUT FUNKTIONIERT DEMOKRATIE IN BRD? | 96 =„KEIN ISSP“ 98 =„w. n.“ 99 =„k. A.“ |
| v874 | WAHLBETEILIGUNG, LETZTE BTW? | 0= „NICHT WAHLBE.“ 6 =„KEIN ISSP“ 9 =„k. A.“ |

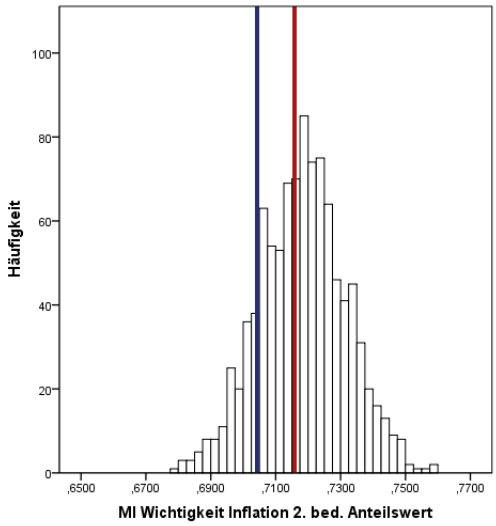
**Anhang 5:** Sonstige visuelle Aufbereitung für Beispiel 1 des Methodenvergleichs bei Item Nonresponse

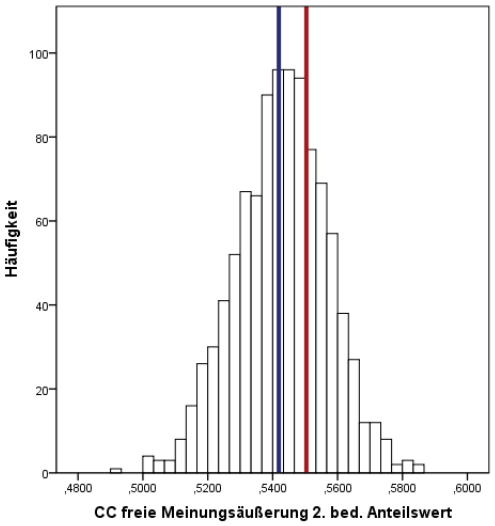


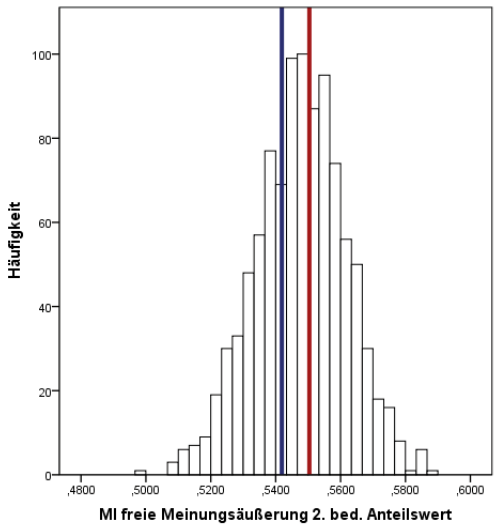



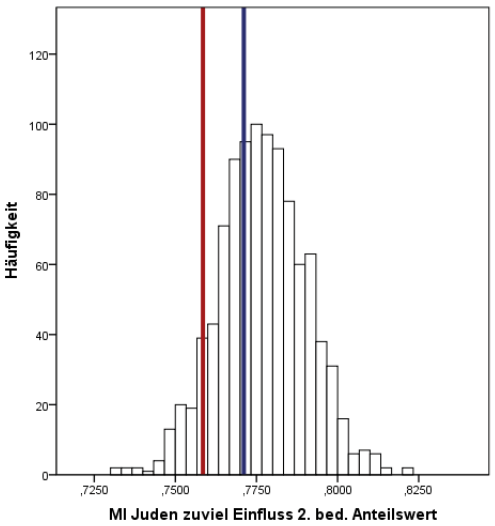



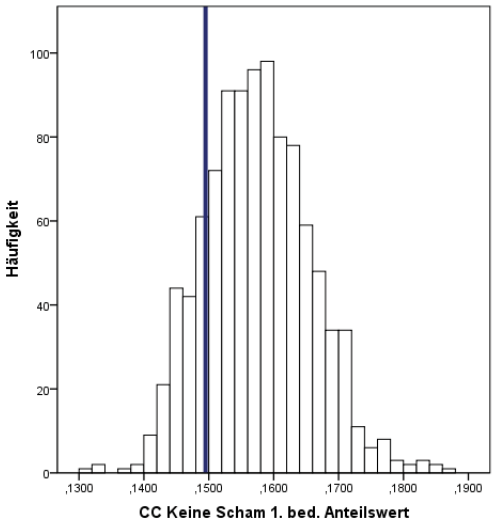



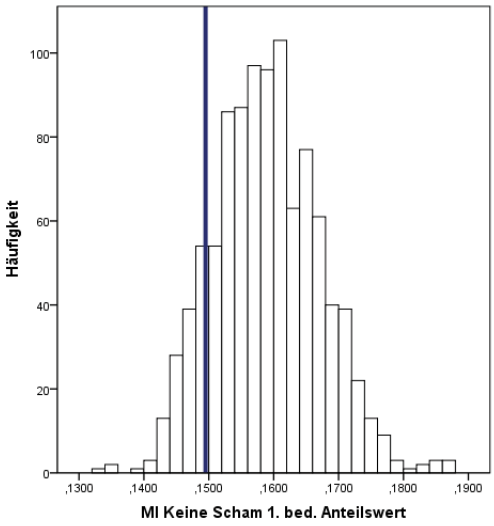

Häufigkeit
100
80
60
40
20
0
,6500
,6700
,6900
,7100
,7300
,7500
,7700
MI Wichtigkeit Inflation 2. bed. Anteilswert

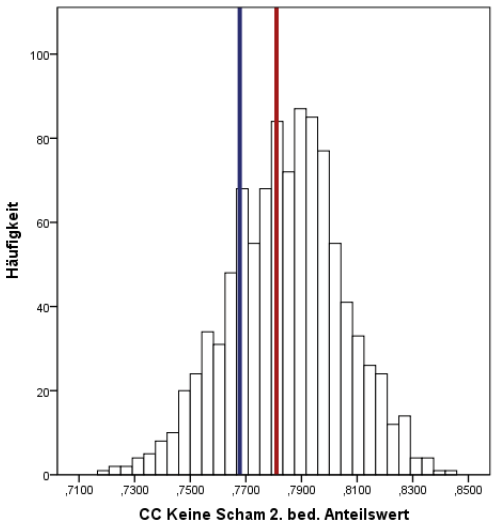

Häufigkeit
100
80
60
40
20
0
,4800
,5000
,5200
,5400
,5600
,5800
,6000
CC freie Meinungsäußerung 2. bed. Anteilswert

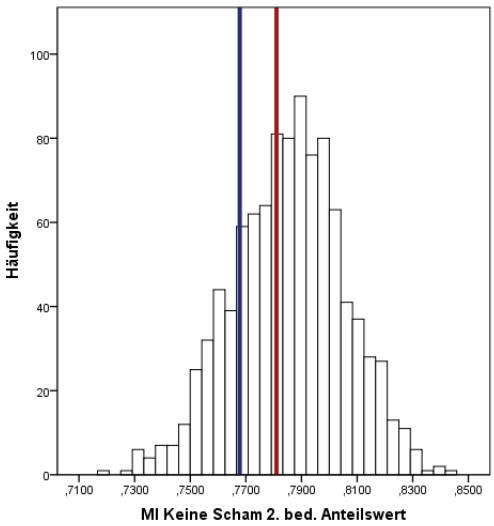

Häufigkeit
100
80
60
40
20
0
,4800
,5000
,5200
,5400
,5600
,5800
,6000
MI freie Meinungsäußerung 2. bed. Anteilswert

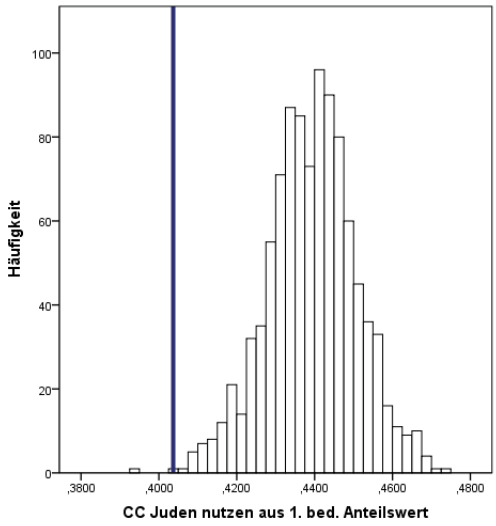

Häufigkeit
120
100
80
60
40
20
0
,7250
,7500
,7750
,8000
,8250
MI Juden zuviel Einfluss 2. bed. Anteilswert

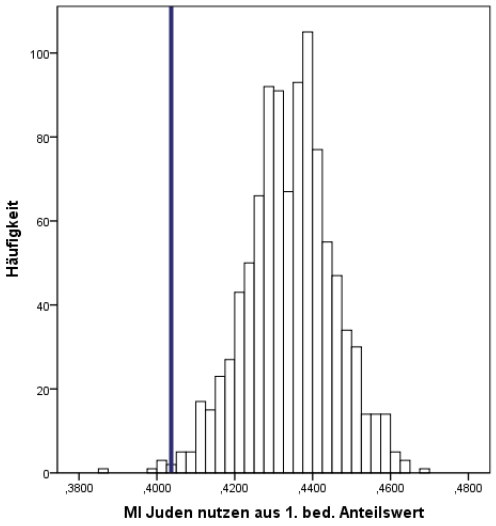

Häufigkeit
100
80
60
40
20
0
,1300
,1400
,1500
,1600
,1700
,1800
,1900
CC Keine Scham 1. bed. Anteilswert

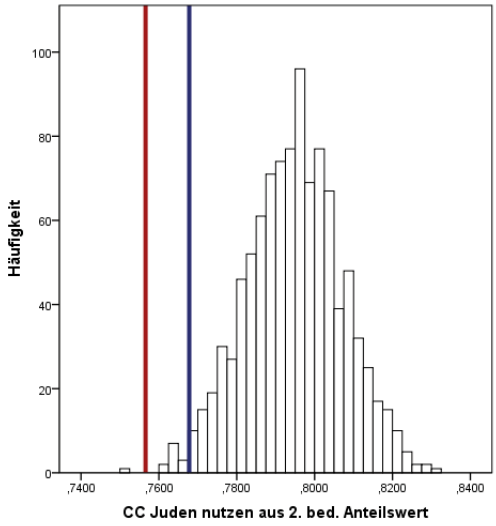

Häufigkeit
100
80
60
40
20
0
,1300
,1400
,1500
,1600
,1700
,1800
,1900
MI Keine Scham 1. bed. Anteilswert

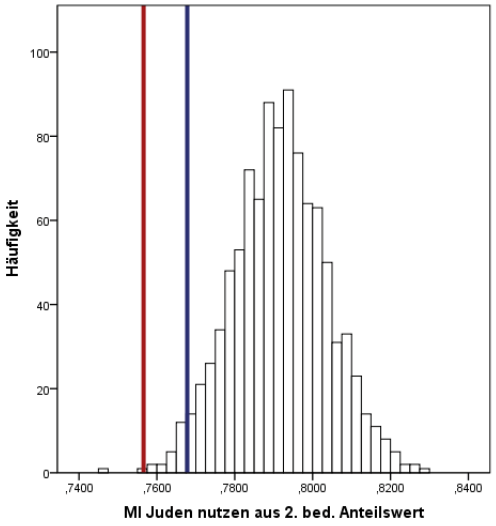

Häufigkeit
100
80
60
40
20
0
,7100
,7300
,7500
,7700
,7900
,8100
,8300
,8500
CC Keine Scham 2. bed. Anteilswert

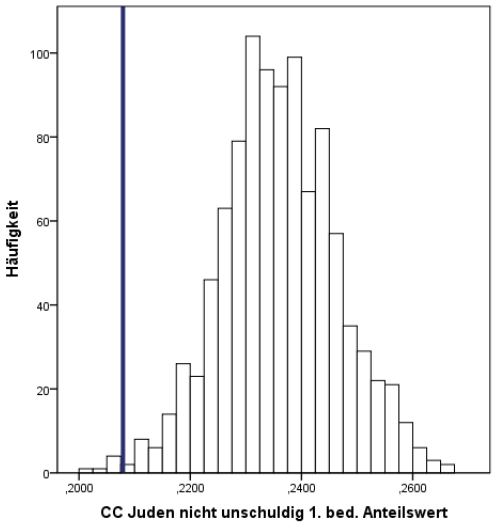

Häufigkeit
100
80
60
40
20
0
,7100
,7300
,7500
,7700
,7900
,8100
,8300
,8500
MI Keine Scham 2. bed. Anteilswert

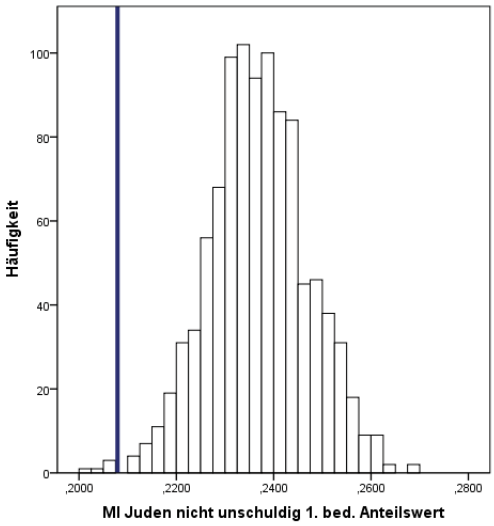

Häufigkeit
0
20
40
60
80
100
,3800
,4000
,4200
,4400
,4600
,4800
CC Juden nutzen aus 1. bed. Anteilswert

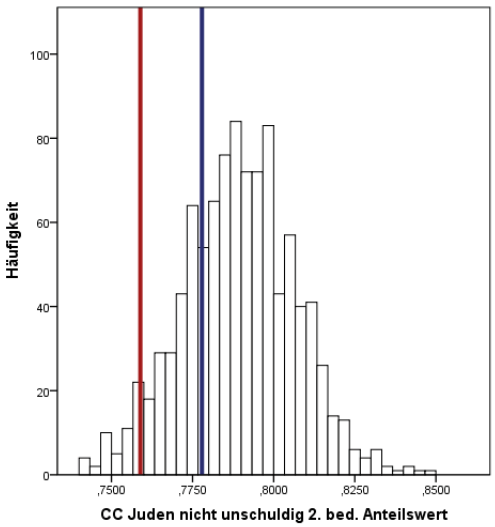

Häufigkeit
0
20
40
60
80
100
,3800
,4000
,4200
,4400
,4600
,4800
MI Juden nutzen aus 1. bed. Anteilswert

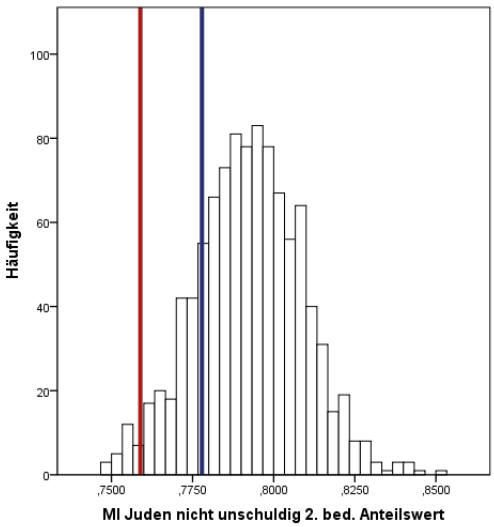

Häufigkeit
0
20
40
60
80
100
,7400
,7600
,7800
,8000
,8200
,8400
CC Juden nutzen aus 2. bed. Anteilswert

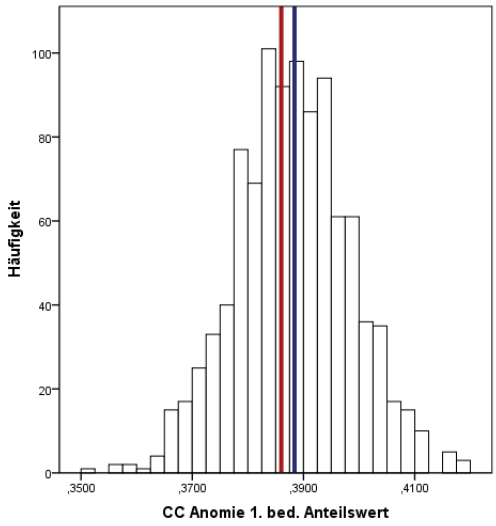

Häufigkeit
0
20
40
60
80
100
,7400
,7600
,7800
,8000
,8200
,8400
MI Juden nutzen aus 2. bed. Anteilswert

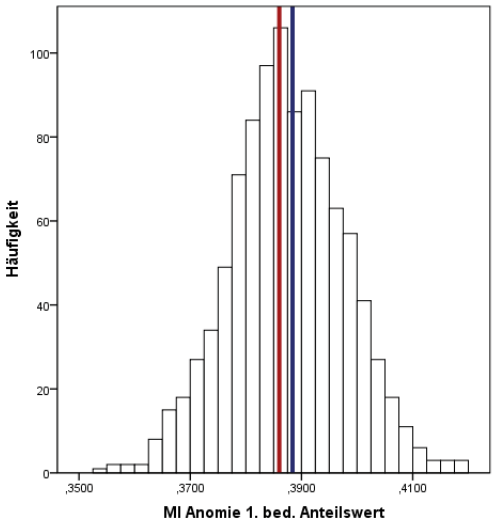

Häufigkeit
100
80
60
40
20
0
,2000
,2200
,2400
,2600
CC Juden nicht unschuldig 1. bed. Anteilswert

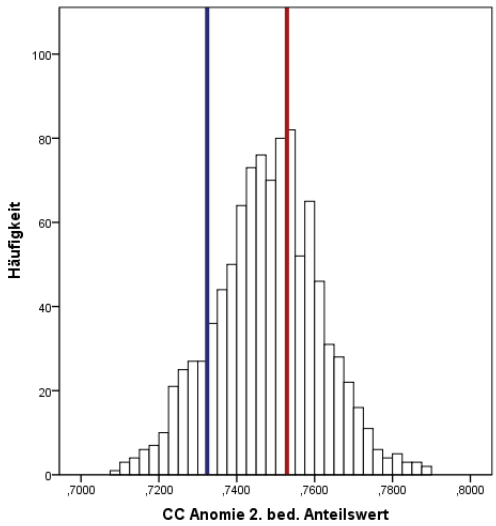

Häufigkeit
100
80
60
40
20
0
,2000
,2200
,2400
,2600
,2800
MI Juden nicht unschuldig 1. bed. Anteilswert

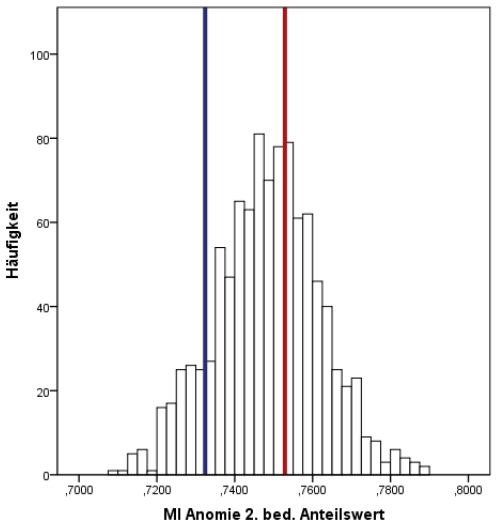

Häufigkeit
100
80
60
40
20
0
,7500
,7750
,8000
,8250
,8500
CC Juden nicht unschuldig 2. bed. Anteilswert


Häufigkeit
100
80
60
40
20
0
,7500
,7750
,8000
,8250
,8500
MI Juden nicht unschuldig 2. bed. Anteilswert

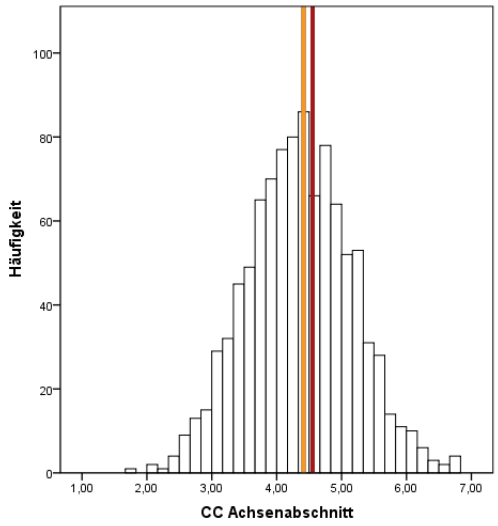

Häufigkeit
0
20
40
60
80
100
,3500
,3700
,3900
,4100
CC Anomie 1. bed. Anteilswert

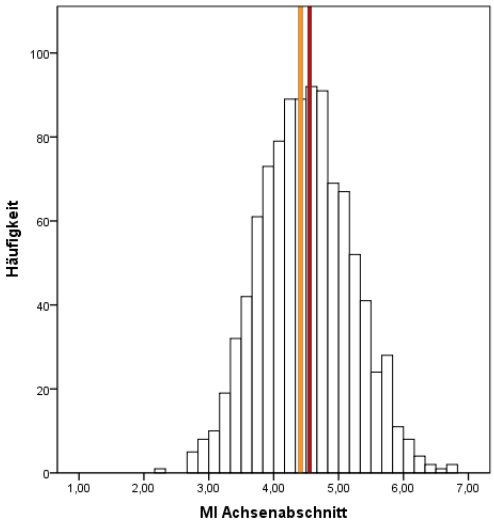

Häufigkeit
0
20
40
60
80
100
,3500
,3700
,3900
,4100
MI Anomie 1. bed. Anteilswert

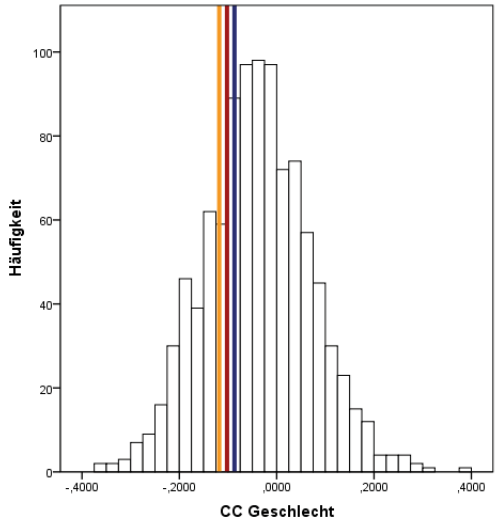

Häufigkeit
0
20
40
60
80
100
,7000
,7200
,7400
,7600
,7800
,8000
CC Anomie 2. bed. Anteilswert

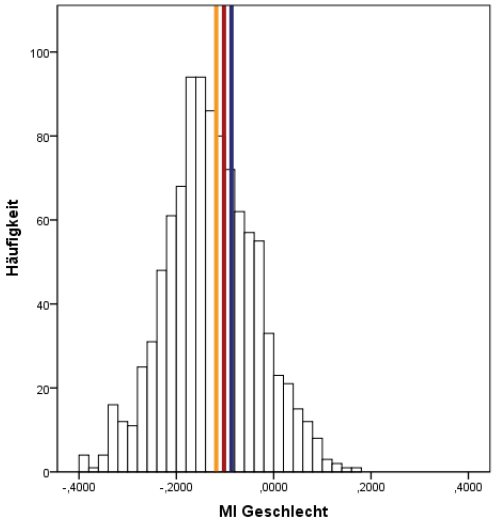

Häufigkeit
0
20
40
60
80
100
,7000
,7200
,7400
,7600
,7800
,8000
MI Anomie 2. bed. Anteilswert

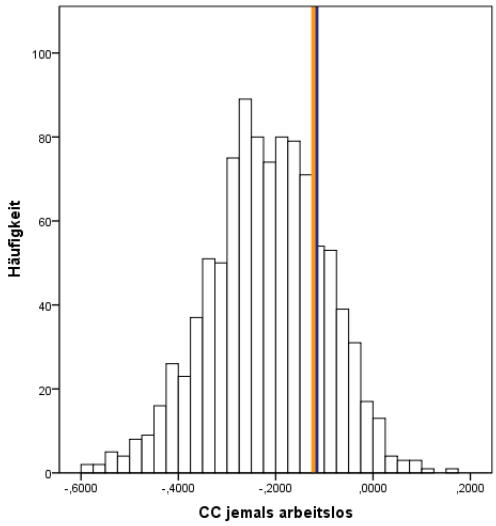
**Anhang 6:** Sonstige visuelle Aufbereitung für Beispiel 2 des Methodenvergleichs bei Item Nonresponse

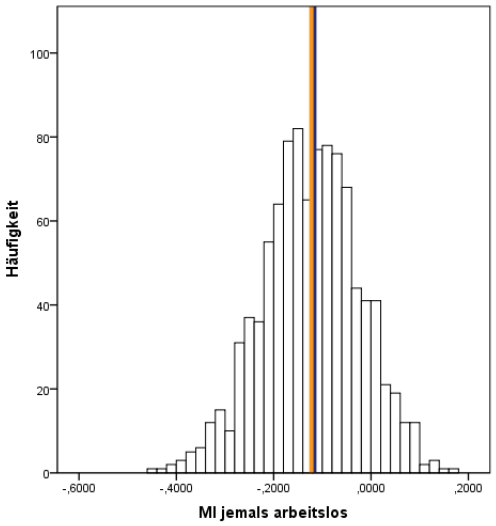


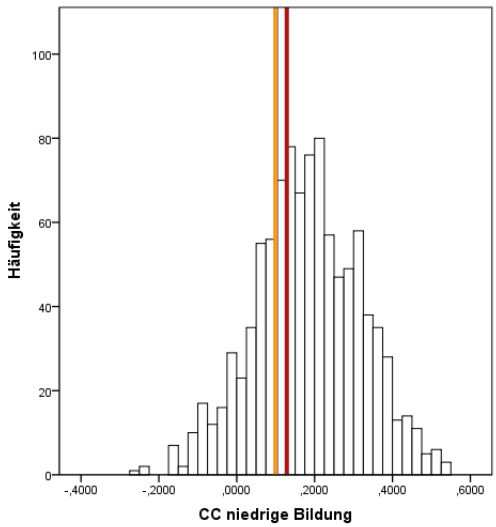



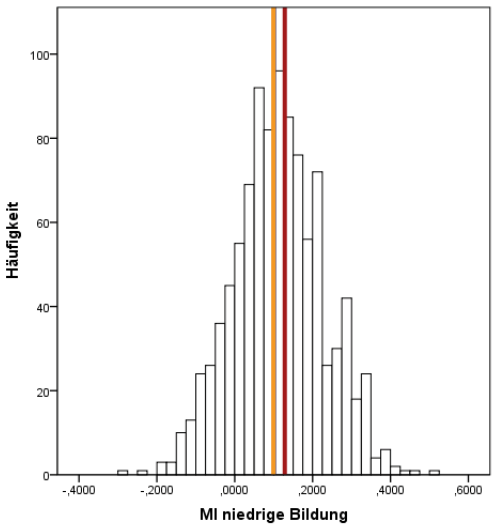



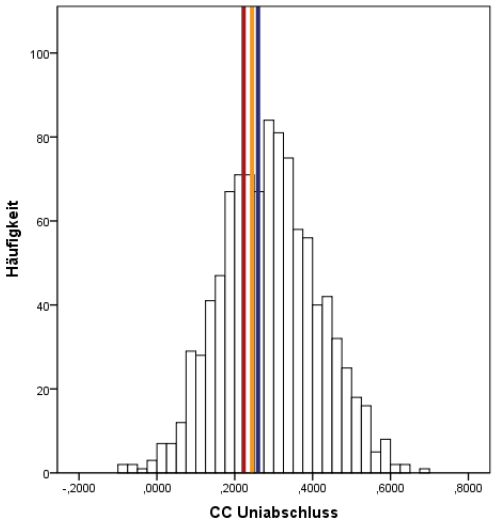



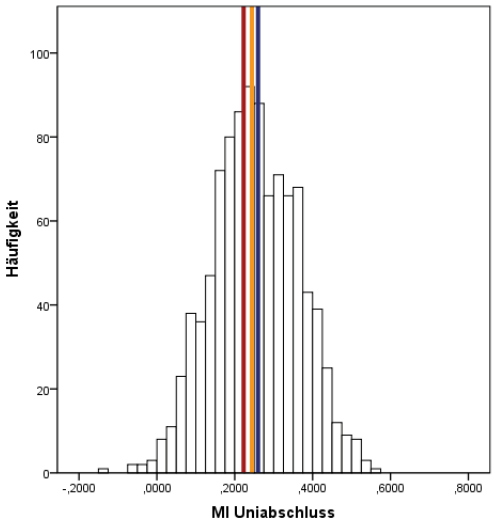

Häufigkeit
100
80
60
40
20
0
-,6000
-,4000
-,2000
,0000
,2000
CC jemals arbeitslos

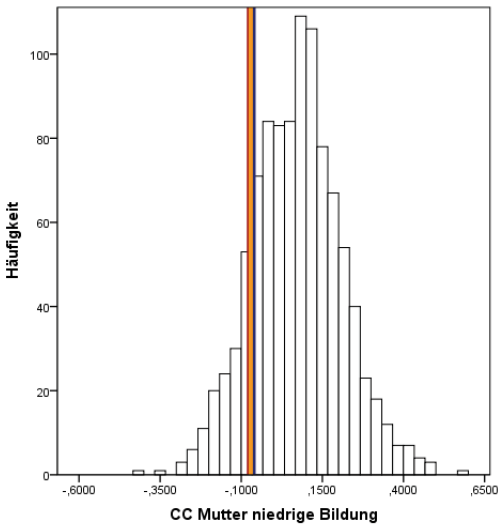

Häufigkeit
100
80
60
40
20
0
-,6000
-,4000
-,2000
,0000
,2000
MI jemals arbeitslos

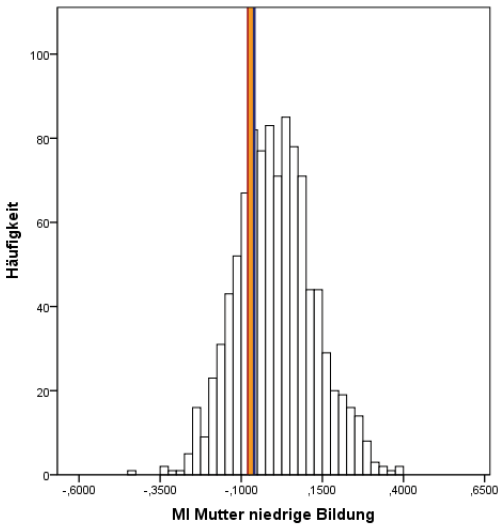

Häufigkeit
100
80
60
40
20
0
-,4000
-,2000
,0000
,2000
,4000
,6000
CC niedrige Bildung

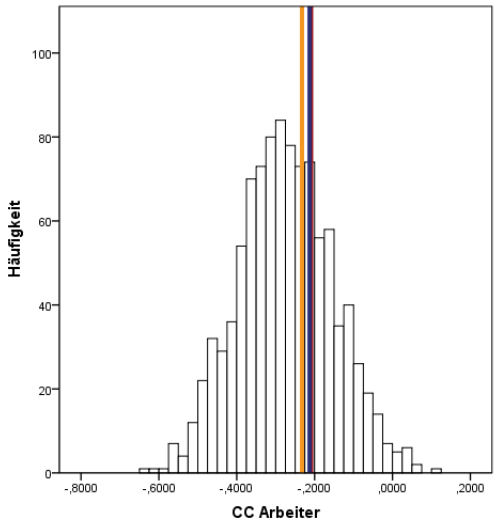

Häufigkeit
100
80
60
40
20
0
-,4000
-,2000
,0000
,2000
,4000
,6000
MI niedrige Bildung

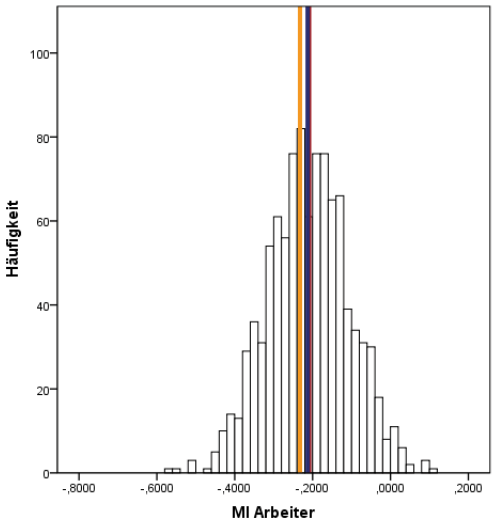

Häufigkeit
CC Uniabschluss

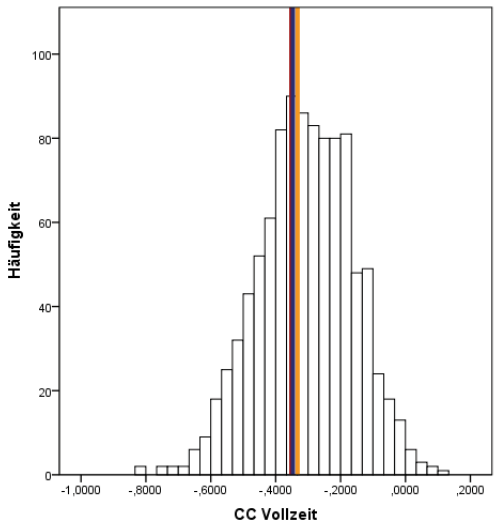

Häufigkeit
MI Uniabschluss

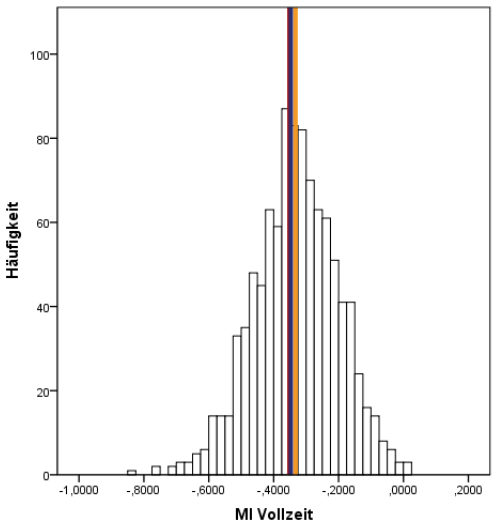

Häufigkeit
CC Mutter niedrige Bildung

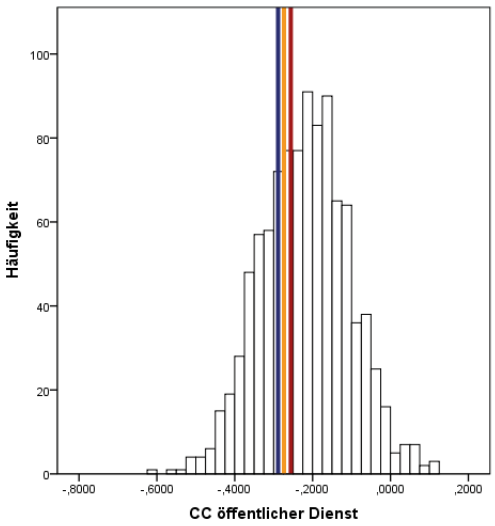

Häufigkeit
MI Mutter niedrige Bildung

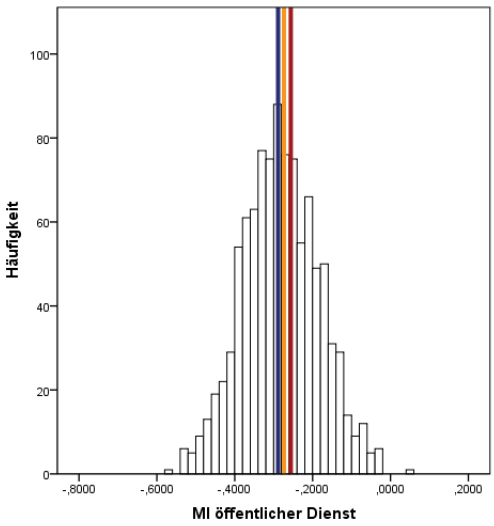

Häufigkeit
100
80
60
40
20
0
-,8000
-,6000
-,4000
-,2000
,0000
,2000
CC Arbeiter


Häufigkeit
100
80
60
40
20
0
-,8000
-,6000
-,4000
-,2000
,0000
,2000
MI Arbeiter

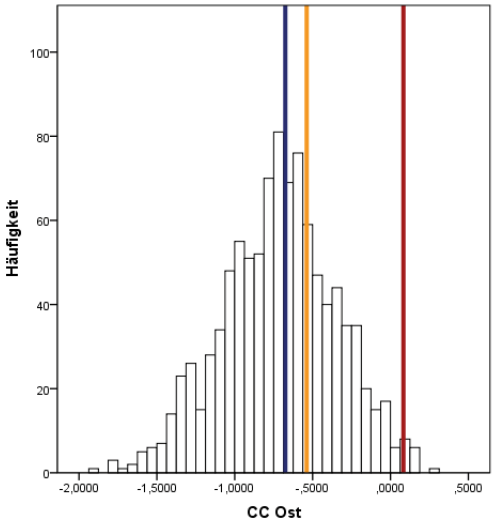

Häufigkeit
100
80
60
40
20
0
-1,0000
-,8000
-,6000
-,4000
-,2000
,0000
,2000
CC Vollzeit

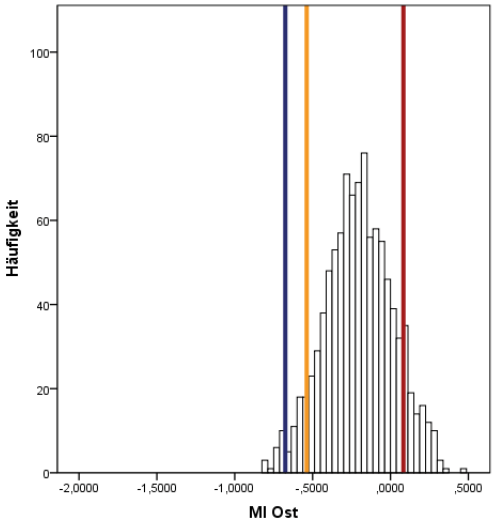

Häufigkeit
100
80
60
40
20
0
-1,0000
-,8000
-,6000
-,4000
-,2000
,0000
,2000
MI Vollzeit

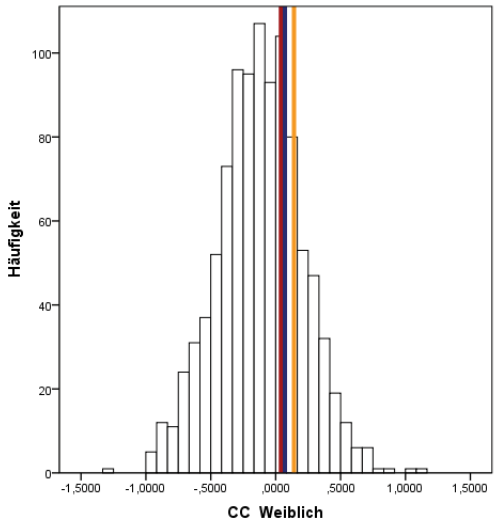


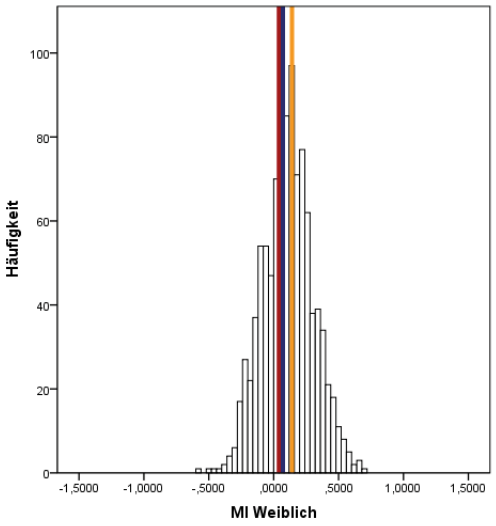



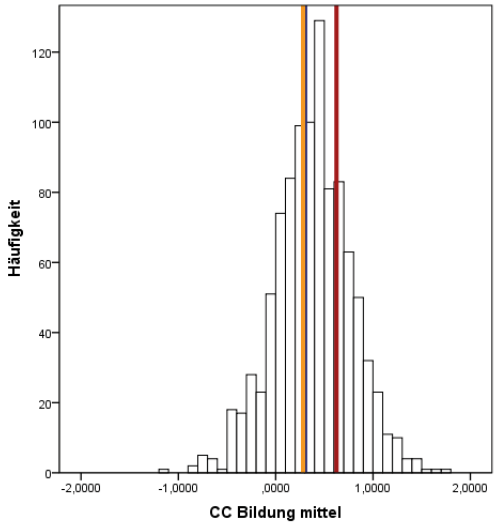

**Anhang 7:** Sonstige visuelle Aufbereitung für Beispiel 3 des Methodenvergleichs bei Item Nonresponse

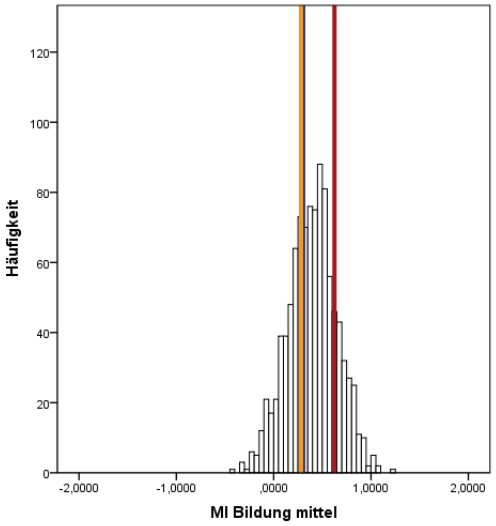


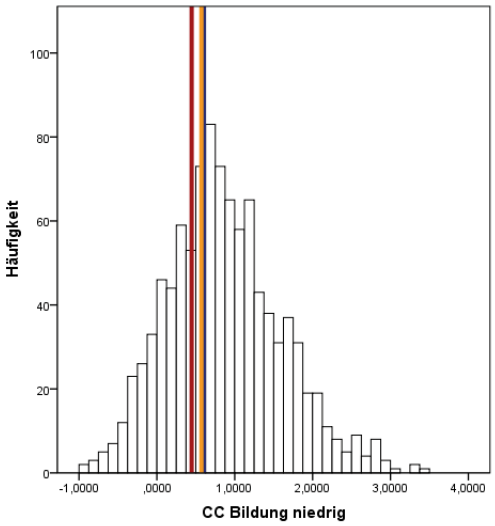



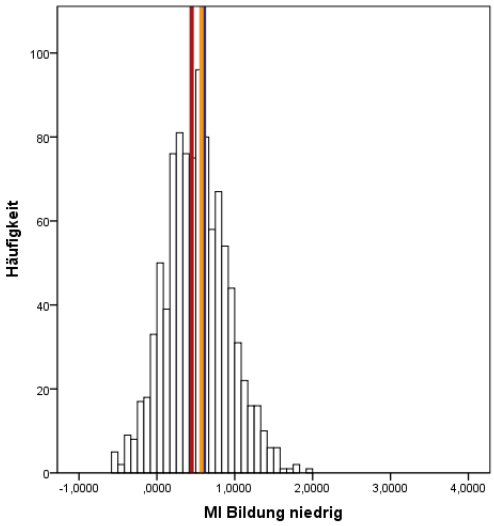

Häufigkeit
CC Weiblich

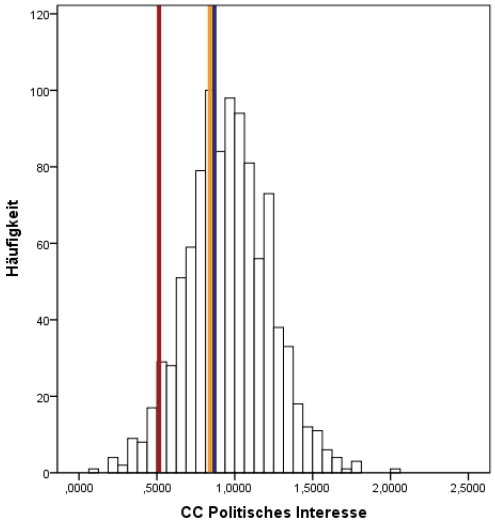

Häufigkeit
MI Weiblich

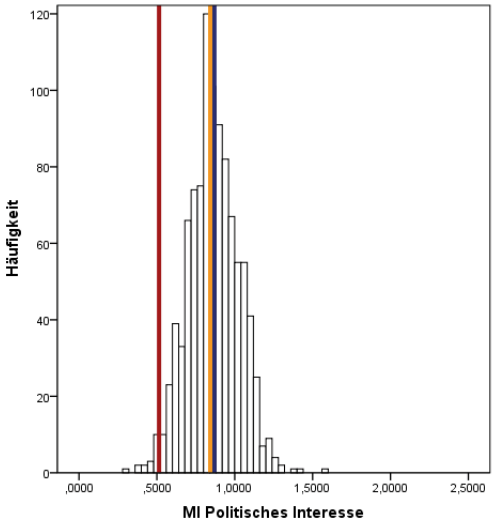

Häufigkeit
CC Bildung mittel

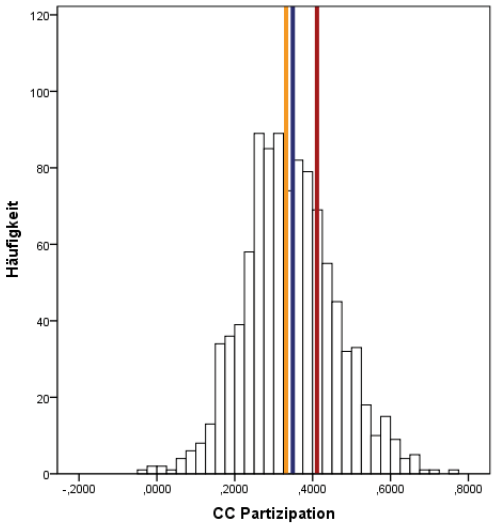

Häufigkeit
MI Bildung mittel

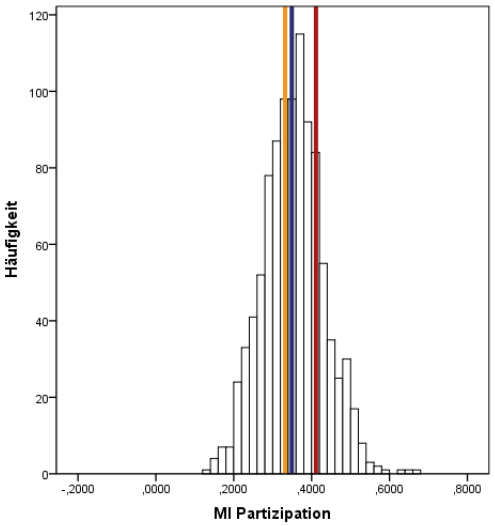

Häufigkeit
CC Bildung niedrig


Häufigkeit
MI Bildung niedrig


Häufigkeit
CC Politisches Interesse


Häufigkeit
MI Politisches Interesse


Häufigkeit
120
100
80
60
40
20
0
-,2000
,0000
,2000
,4000
,6000
,8000
CC Partizipation


Häufigkeit
120
100
80
60
40
20
0
-,2000
,0000
,2000
,4000
,6000
,8000
MI Partizipation

**Anhang 8:** Änderung der erfassten Ausfallkategorien des ALLBUS von 1980-2008

| | ‘80 | ‘82 | ‘84 | ‘86 | ‘88 | ‘90 | ‘91 | ‘92 | ‘94 | ‘96 | ‘98 | ‘00 | ‘02 | ‘04 | ‘06 | ‘08 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Im HH niemanden angetroffen | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| ZP nicht angetroffen | - | - | - | X | X | X | X | X | X | X | X | X | X | X | X | X |
| ZP trotz Terminvereinbarung/ trotz mehrfacher Besuche nicht angetroffen/ ZP über den Befragungszeitraum abwesend (Auslandsauf., Montage u.a.) | X | X | X | - | - | - | - | - | - | - | - | - | - | - | - | - |
| ZP ist krank/ nicht befragungsfähig | X | X | X | - | - | - | - | - | - | - | X | - | - | - | - | - |
| ZP nicht befragungsfähig | - | - | - | X | X | X | X | X | X | X | - | X | X | X | X | X |
| ZP verweigert telefonisch bei Infratest Geschäftsleitung | - | - | - | - | - | - | - | - | - | - | - | - | - | - | X | - |
| ZP verweigert das Interview/ lässt sich verleugnen/ lässt Interview durch andere Person verweigern | X | X | X | - | - | - | - | - | - | - | - | - | - | - | - | - |
| ZP aus Zeitgründen nicht zum Interview bereits | - | - | - | - | - | - | - | - | X | X | - | X | X | X | X | X |
| HH oder Zielperson nicht kooperativ | - | - | - | X | X | X | X | X | - | - | - | - | - | - | - | - |
| Angetroffene Person verweigert jede Auskunft, auch HH-Auflistung | - | - | - | - | - | - | - | - | - | - | X | - | - | - | - | - |
| Angetroffene Person ist nicht ZP, verhindert aber nach HH-Auflistung die Durchführung des Interviews | - | - | - | - | - | - | - | - | - | - | X | - | - | - | - | - |
| ZP verweigert das Interview | - | - | - | - | - | - | - | - | - | - | X | - | - | - | - | - |
| ZP generell nicht zum Interview bereit | - | - | - | - | - | - | - | - | X | X | - | X | X | X | X | X |
| ZP spricht nicht hinreichend gut deutsch | - | - | - | - | - | - | - | - | - | - | - | - | X | X | X | X |
| Mit Haushalt/ZP ist keine Verständigung auf Deutsch möglich | - | - | - | - | - | - | - | - | X | - | X | - | - | - | - | - |
| Andere Ausfallgründe | - | - | - | - | - | - | - | - | X | - | X | - | - | - | - | - |
| Anderer Grund | - | - | - | - | - | X | - | - | - | - | - | - | - | - | - | - |
| Anderer Grund bzw. keine Angabe des Ausfallgrunds | X | X | X | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Adresse nicht abschließend bearbeitet | - | - | - | - | - | - | - | - | - | - | - | - | - | X | X | X |
| Adresse bearbeitet, aber kein Ausfallprotokoll | - | - | - | - | X | - | - | - | - | - | - | - | - | - | - | - |
| Interview als (Teil-)Fälschung identifiziert | - | - | - | - | - | - | - | - | - | - | - | - | - | X | X | X |
| Interview nicht korrekt durchgeführt | - | - | - | - | - | - | - | - | - | X | - | X | - | - | - | - |
| Zweifel an korrekter Durchführung | - | - | - | - | - | - | - | - | - | - | - | - | X | - | - | - |
| Interview nicht vollständig durchgeführt | - | - | - | - | - | - | - | - | - | - | - | - | X | - | - | - |
| Interview nicht auswertbar wg. Zweifel an korrekter Durchführung | - | - | - | - | - | - | - | - | X | - | X | - | - | - | - | - |

**Anhang 9:** Item für den Anteilswert des Methodenvergleichs bei Unit Nonresponse

| Variablen | Labels |
|---|---|
| v534 | WAHLABSICHT, BUNDESTAGSWAHL |

**Anhang 10:**Item für den Mittelwert des Methodenvergleichs bei Unit Nonresponse

| Variablen | Labels |
|---|---|
| v106 | LINKS-RECHTS-SELBSTEINSTUFUNG |

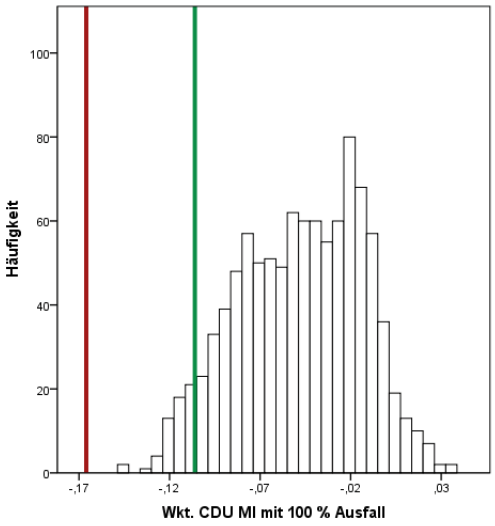
**Anhang 11:** Items für das OLS-Modell des Methodenvergleichs bei Unit Nonresponse

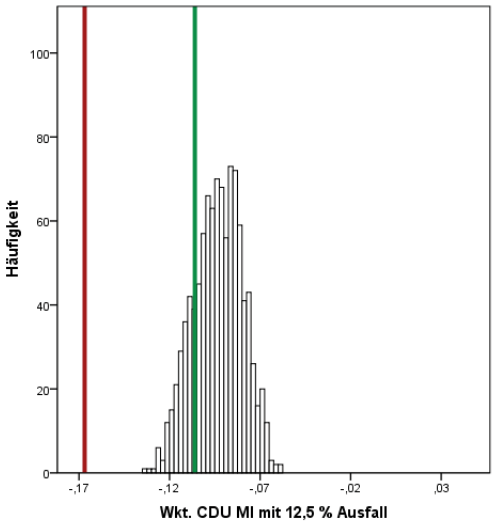
| Variablen | Labels |
|---|---|
| v95 | ZUFRIEDEN MIT LEISTUNG D.BUNDESREGIERUNG |
| v105 | LINKS-RECHTS-SELBSTEINSTUFUNG |
| v115 | FRAU, LIEBER MANN BEI D.KARRIERE HELFEN? |
| v131 | WAHRSCHEINLICHKEIT: CDU-CSU WAEHLEN |
| v132 | WAHRSCHEINLICHKEIT: SPD WAEHLEN |

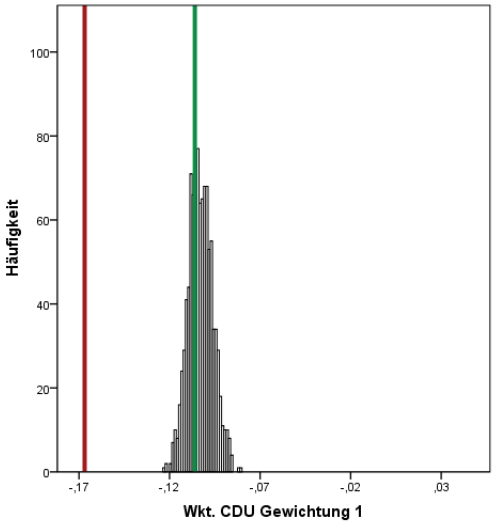
**Anhang 12:** Items für das Logitmodell des Methodenvergleichs bei Unit Nonresponse

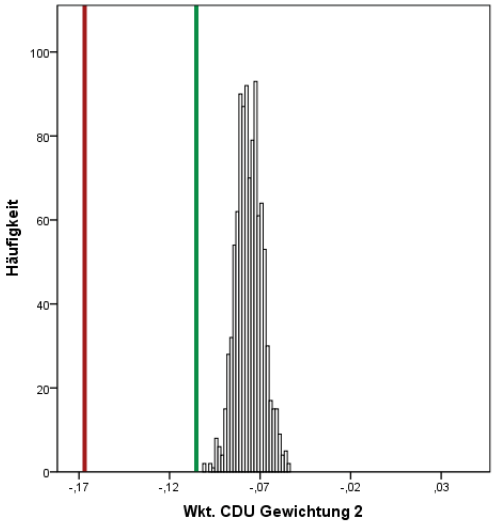
| Variablen | Labels |
|---|---|
| v534 | WAHLABSICHT, BUNDESTAGSWAHL |
| v117 | FRAU, ZU HAUSE BLEIBEN+KINDER VERSORGEN? |
| v131 | WAHRSCHEINLICHKEIT: CDU-CSU WAEHLEN |
| v141 | ZUSTIMMUNG: STOLZ, DEUTSCHER ZU SEIN |

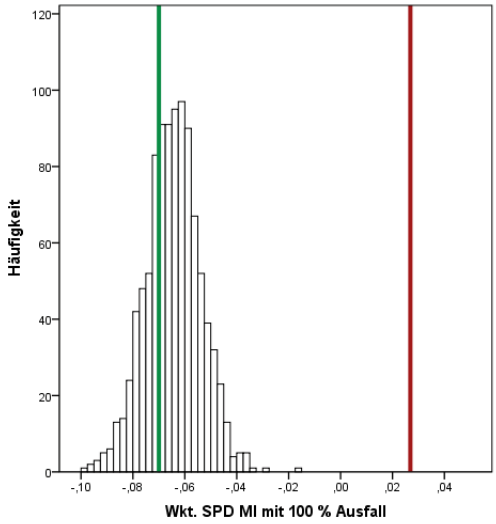
**Anhang 13:** Sonstige visuelle Aufbereitung für das OLS-Modell des Methodenvergleichs bei Unit Nonresponse

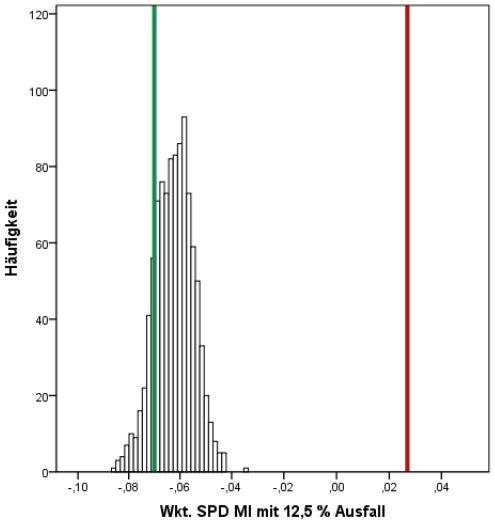


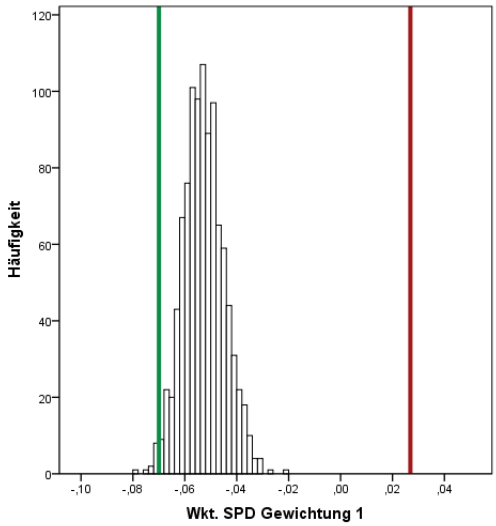



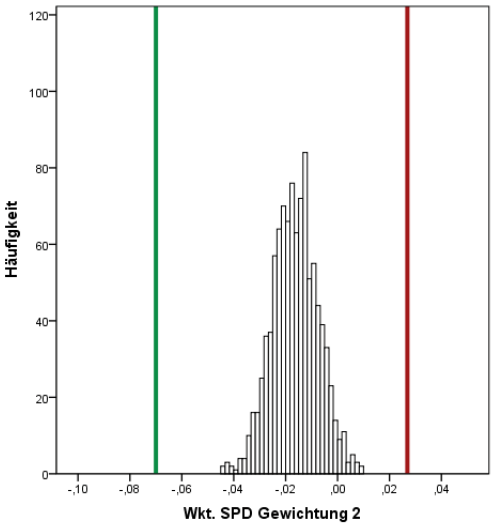



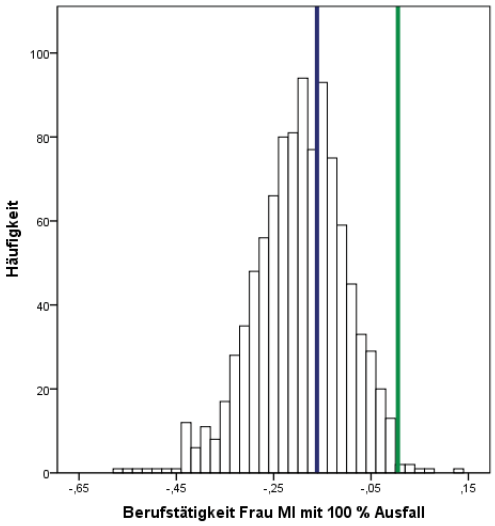

Häufigkeit
120
100
80
60
40
20
0
-,10
-,08
-,06
-,04
-,02
,00
,02
,04
Wkt. SPD MI mit 100 % Ausfall

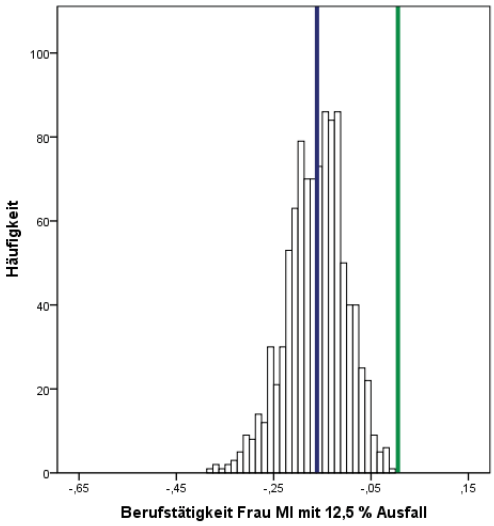

Häufigkeit
120
100
80
60
40
20
0
-,10
-,08
-,06
-,04
-,02
,00
,02
,04
Wkt. SPD MI mit 12,5 % Ausfall

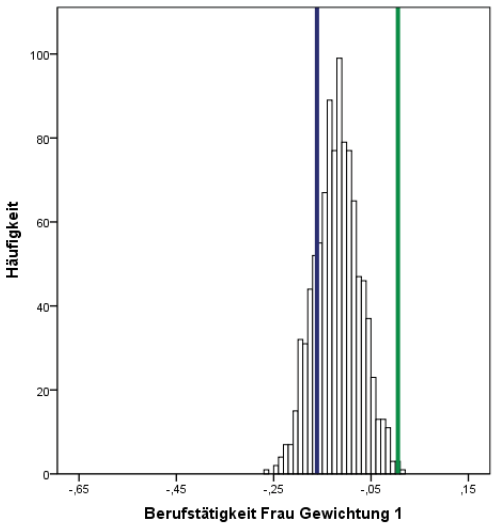

Häufigkeit
120
100
80
60
40
20
0
-,10
-,08
-,06
-,04
-,02
,00
,02
,04
Wkt. SPD Gewichtung 1

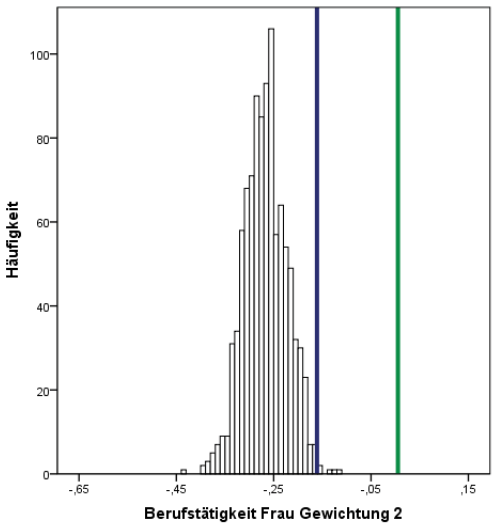

Häufigkeit
120
100
80
60
40
20
0
-,10
-,08
-,06
-,04
-,02
,00
,02
,04
Wkt. SPD Gewichtung 2

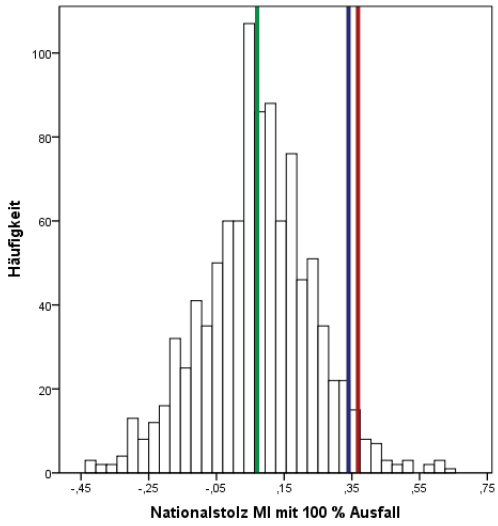
**Anhang 14:** Sonstige visuelle Aufbereitung für das Logitmodell des Methodenvergleichs bei Unit Nonresponse

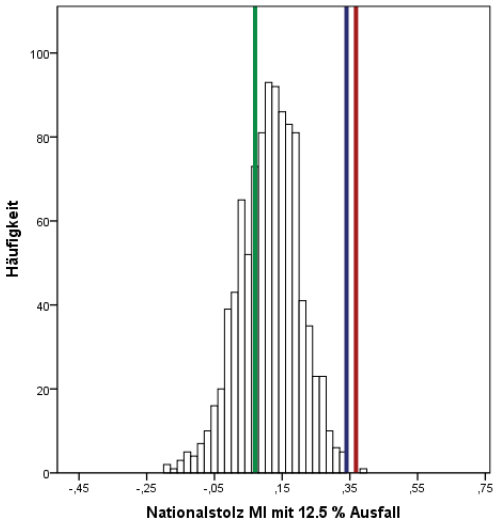


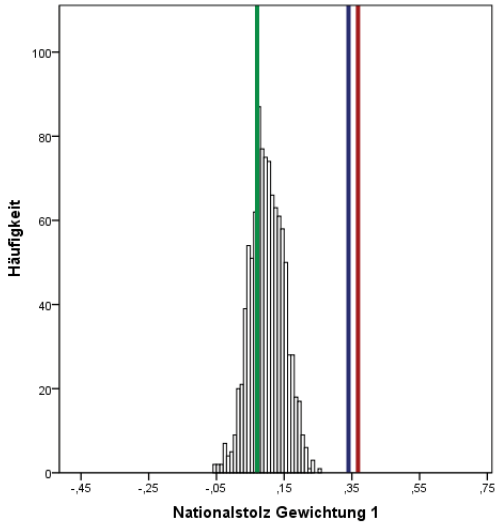



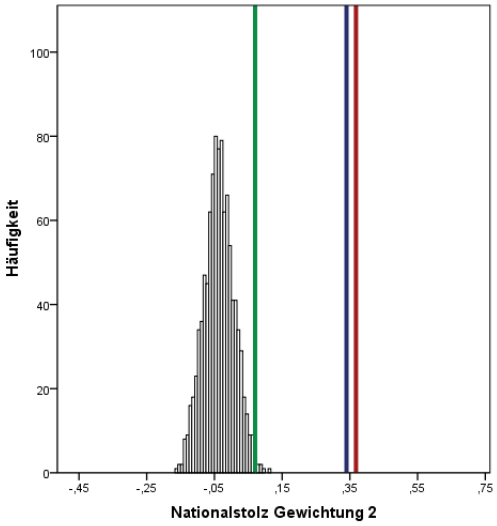




Häufigkeit
100
80
60
40
20
0
-,45
-,25
-,05
,15
,35
,55
,75
Nationalstolz MI mit 100 % Ausfall


Häufigkeit
100
80
60
40
20
0
-,45
-,25
-,05
,15
,35
,55
,75
Nationalstolz MI mit 12.5 % Ausfall


Häufigkeit
100
80
60
40
20
0
-,45
-,25
-,05
,15
,35
,55
,75
Nationalstolz Gewichtung 1


Häufigkeit
100
80
60
40
20
0
-,45
-,25
-,05
,15
,35
,55
,75
Nationalstolz Gewichtung 2

# Literaturverzeichnis

ABELSON, R.P. (1981) Psychological Status of the Script Concept, American Psychologist, 7, S.715-729.

ALBERS, I. (1997) Einwohnermelderegister-Stichproben in der Praxis. Ein Erfahrungsbericht, Stichproben in der Umfragepraxis (Hrsg. Gabler, S. und J. Hoffmeyer-Zlotnik), Westdeutscher Verlag, Opladen.

ALLISON, P.D. (2002) Missing Data, Series: Quantitative Applications in the Social Sciences, Sage University Press, Thousand Oaks.

ALWIN, D.F. und J.A. KROSNICK (1991) Aging, Cohorts, and the Stability of Socialpolitical Orientations over the Life Span, The American Journal of Sociology, 97/1, S.169-195.

ATTESLANDER, P. (2010) Methoden der empirischen Sozialforschung, Schmidt, Berlin.

ATTESLÄNDER, P. und H.-U. KNEUBÜHLER (1975) Verzerrungen in Interviews. Zu einer Fehlertheorie der Befragung, VS für Sozialwissenschaften, Wiesbaden.

BANDILLA, W. et al. (1992) Methodenbericht zum DFG-Projekt ALLBUS Baseline-Studie, ZUMA Arbeitsbericht 4.

BANKHOFER, U. (1995) Unvollständige Daten- und Distanzmatrizen in der multivariaten Datenanalyse, Eul, Bergisch Gladbach.

BARRON, D.N. (1992) The Analysis of Count Data: Overdispersion and Autocorrelation, Sociological Methodology, 22, S.179-220.

BEATTY, P. und D. HERRMANN (2002) To Answer or Not to Anwere: Decision Process Related to Survey Item Nonresponse, Survey Nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.71-86.

BECKER, I. und R. HAUSER (2003) Anatomie der Einkommensverteilung. Ergebnisse der Einkommens und Verbrauchsstichproben 1969-1998, edition sigma, Berlin.

BIEMER, P.P. und S.L. CHRIST (2008) Constructing the survey weights. Sampling of Populations: Methods and Applications (Hrsg. Levy, P.S. und S. Lemeshow), Wiley, New York.

BIRKELBACH, K. (1998) Befragungsthema und Panelmortalität. Ausfälle in einer Lebenslauferhebung, ZA-Informationen, 42, S.128-147.

BLOHM, M. et al. (2003) Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“(ALLBUS) 2002. ZUMA Methodenbericht,
2003/12.

BLOHM, M. (2005) Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS). Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund (Hrsg. Grözinger, G. und W. Matiaske), Hampp Verlag, München, S.43-55.

BLOHM, M. et al. (2011) ALLBUS-Bibliographie 25. Fassung, Stand: März 2011, Technical Reports, 2011/6.

BÖHNING, D. et al. (1999) The Zero-Inflated Poisson Model and the Decay, Missing and Filled Teeth index in dental epidemiology, Royal Statistics Society, 162, S.195-209.

BORG, I. und T. STAUFENBIEL (2007) Lehrbuch Theorien und Methoden der Skalierung, Huber, Bern.

BOTMAN, S.L. und O.T. THORNBERRY (1992) Survey Design Features Correlates of Nonresponse, ASA Proceedings of the Section on Survey Research Methods, S.309-314.

BREDENKAMP, J. und B. VATERRODT (1992) Direkte und indirekte Gedächtnisprüfung skriptbezogener Informationen, Sprache und Kognition, 11, S.14-26.

BREHM, J. (1993) The Phantom Respondents. Opinion Surveys and Political Representation, Michigan Studies in Politicial Analysis, The University of Michigan Press, Ann Arbor.

BÜSCHGES, G. et al. (1995) Grundzüge der Soziologie, Oldenbourg, München.

CANTOR, D. et al. (2004) Testing an Automated Refusal Avoidance Training Methodology. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

CATANIA, J.A. et al. (1990) Methodological Problems in Aids Behavioral Research. Influences on Measurement Error and Participations in Studies of Sexual Behavior, Psychological Bulletin, 108/3, S.339-362.

CIALDINI, R.B. (1989) Social motivations to comply: Norms, values, and principles, Taxpayer Compliance (Hrsg. Roth, J. und J. Scholz), University of Pennsylvania Press, Philadelphia.

CALDINI, R.B. (1994) Principles and techniques of social influence, Advanced social psychology (Hrsg. Tesser, A.), McGraw-Hill, New York.

CIALDINI, R.B. und B.J. SAGARIN (2005) Principles of interpersonal influence, Persuasion: Psychological insights and perspectives (Hrsg. Brock, T. und M. Green), Sage University Press, Newbury Park, S.143-169.

COCHRAN, W.G. (1977) Sampling Techniques, 2nd Ed., Wiley, New York.

CONSUL, P.C. (1989) Generalized Poisson Distributions. Proberties and Applications, Marcel Dekker, New York.

CONVERSE, P.E. (1970) Attitudes and Non-Attitudes: Continuation of a Dialogue, Quantitative Analysis of Social Problems (Hrsg. Tufte, E.), MA (Addison-Wesley), Reading, S.168-189.

COUNCIL OF AMERICAN SURVEY RESEARCH ORGANIZATIONS (1982) „On the Definition of Response Rates," L. Frankel, chairman, Completion Rates Task Force, Fort Jefferson, NY, CASRO.

CZADO, C. et al. (2007) Zero-inflated generalized Poisson models with Regression Effects on the Mean, Dispersion and Zero-inflation Level Applied to Patent Outsourcing Rates, Statistical Modelling, 7, S.125-153.

DE HEER, W. (1999) International Response Trends: Results of an International Survey, Journal of Official Statistics, 15/2, S.129–142.

DE LEEUW, E.D. et al. (2008) The Cornerstones of Survey Researchm, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.1-17.

DE LEEUW, E.D. (2008) Choosing the Method of Data Collection, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.113-135.

DE NÈVE, D. (2009) NichtwählerInnen – eine Gefahr für die Demokratie? Budrich, Opladen.

DEMPSTER, A.P. et al. (1977) Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 39, S.1-38.

DIEKMANN, A. (2006) Aktuelle Probleme der empirischen Sozialforschung, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44: Methoden der Sozialforschung, S.8-32.

DILLMAN, D.A. (1978) Mail und Telephone Surveys. The Total Design Method, Wiley, New York.

DILLMAN, D.A. et al. (2002) Survey Nonresponse in Design, Data Collection, and Analysis, Survey Nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.3-26.

DILLMAN, D.A. (2008) The Logic and Psychology of Constructing Questionaires, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.163.

ENDERS, C. (2010) Applied Missing Data Analysis, Guilford Press, New York.

ENGEL, U. et al. (2004) Nonresponse und Stichprobenqualität. Ausschöpfung in Umfragen der Markt- und Sozialforschung, Universtität Bremen, Frankfurt am Main.

ESSER, H. (1984) Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Interviewereffekten, Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Beiträge zu methodischen Problemen des ALLBUS 1980 (Hrsg. Mayer, K. U. et al.), Campus Verlag, Frankfurt, S.26-71.

FITZGERALD, R. und L. FULLER (1982) „I Hear You Knocking but You Can't Come".The Effects of Reluctant Respondents and Refusers on Sample Survey Estimates, Sociological Methods and Research, 11, S.3–32.

FOWLER, F.J. und C. CONSENZA (2008) Writing effective questions, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.136-160.

GÖTHLICH, S.E. (2007) Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen, Methoden der empirischen Sozialforschung (Hrsg. Albers, S. et al.), Gabler, Wiesbaden, S.119-134.

GOYDER, J. (1987) The Silent Minority: Nonrespondents on Sample Surveys, Polity Press/Westview Press, Cambridge.

GREEN, D.P. und I. SHAPIRO (1999) Rational Choice. Eine Kritik am Beispiel von Anwendungen in der politischen Wissenschaft, Oldenbourg, München.

GROVES, R.M. et al. (2000), Leverage-Salience Theory of survey participation. Description and an illustration, The Public Opinion Quarterly, 64/3, S.299-308.

GROVES, R.M. et. al (2004) Survey Methodology, Wiley, New York.

GROVES, R.M. und M. COUPER (1998) Nonresponse in Household Interview Surveys, Wiley, New York.

HAARMANN, A. et al. (2006) Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften"(ALLBUS) 2004, ZUMA Methodenbericht, 2006/6.

HARKNESS, J. et al. (1998) Incentives in Two German Mail Surveys Methods, Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.173-187.

HASHER, L. und M. GRIFFIN (1978) Reconstructive and reproductive processes in forgetting, Journal of Experimental Psychology: Human Learning and Memory, 4, S.318-330.

HAUNBERGER, S. (2006) Das standardisierte Interview als soziale Interaktion: Interviewereffekte in der Umfrageforschung, ZA Informationen, 58, S.23-46.

HAVASI, E. und A. MARTON (1998) Nonresponse in the 1996 Income Survey, (Supplement to the Microcensus), Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.65-74.

HEBERLEIN, T.A. und R. BAUMGARTNER (1978) Fractors Affecting Response Rates to Mail Questionnaires: A Quantitativ Analysis of the Published Literature, American Sociological Review, 43/4, S.447-462.

HECKMAN, J.J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models, Annals of Economic and Social Measurement, 5/4, S.475-492.

HENNIG, E. (2009) „Einen Schlussstrich unter die nationalsozialistische Vergangenheit ziehen" Zur politischen Soziologie eines historischen Deutungsmuster, Einsicht, 2, S.42-49.

HILBE, J.M. (2007) Negative Binomial Regression, Cambridge University Press, Cambridge.

HOROWITZ, J.L. und C.F. MANSKI (2000) Nonparametric Analysis of Randomized Experiments with Covariates and Outcome Data, Journal of the American Statistical Association, 95, S.77-84.

HOX, J. et al. (1998) Fighting Nonresponse in Telephone Interviews; Successful Interviewer Tactics, Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.173-185.

INGLEHART, R. (1977) The Silent Revolution: Changing Values and Political Styles among Western Publics, Princeton University Press, Princeton.

INGLEHART, R. und P.R. ABRAMSON (1999) „Measuring Postmaterialism". The American Political Science Review, 93/3, S.665-677.

JOBE, J.B. und D.J. HERRMANN (1996) Implications of Models of Survey Cognition for Memory Theory, Basic and Applied Memory Research (Hrsg. Herrmann, D.J. et al.)(Band 2), Erlbaum, Hillsdale, S.193-205.

JOHNSON, T.P. et al. (2002) Culture and Survey Nonresponse, Survey nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.55-69.

KALTON, G. (1983) Introduction to survey sampling, Sage University Press, Beverly Hills.

KENDALL, M.G. und W.R. BUCKLAND (1960) A dictionary of statistical terms. Prepared for the International Statistical Institute with the assistance of the UNESCO, Oliver and Boyd for the International Statistical Institute, Edinburgh.

KISH, L. (1965) Survey sampling, Wiley, New York.

KOCH, A. (1997) ADM-Design und Einwohnermelderegisterstichprobe. Stichprobenverfahren bei mündlichen Bevölkerungsumfragen, Stichproben in der Umfragepraxis (Hrsg. Gabler, S. und J. Hoffmeyer-Zlotnik), Westdeutscher Verlag, Opladen, S.99-116.

KOCH, A. (2002) 20 Jahre Feldarbeit im ALLBUS: Ein Einblick in die Blackbox, ZUMA Nachrichten, 51.

KOCH, A. et al. (1999) Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften"(ALLBUS) 1998. ZUMA Arbeitsbericht, 1999/2.

KOCH, A. et al. (2001) Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften"(ALLBUS) 2000, ZUMA Methodenbericht, 2001/5.

KOCH, A. und M. WASMER (2004) Der ALLBUS als Instrument zur Untersuchung sozialen Wandels: Eine Zwischenbilanz nach 20 Jahren, Sozialer und politischer Wandel in Deutschland. Analysen mit ALLBUS-Daten aus zwei Jahrzehnten (Hrsg. Schmitt-Beck, R. et al.), VS Verlag für Sozialwissenschaften, Wiesbaden, S.13-41.

KOLLER-MEINFELDER, F. (2010) Analysis of incomplete survey data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching, Bamberg.

KREBS, D. und K. SCHÜSSLER (1987) Soziale Empfindungen. Ein interkultureller Skalenvergleich bei Deutschen und Amerikanern, Campus Verlag, New York.

KREUTER, F. und K. OLSON (2011) Multiple auxiliary variables in nonresponse adjustment, Sociological Methods and Research, 40/2, Document Actions, S.311-332.

KROSNICK, J.A. (2002) The Causes fo No-Opinion Responses to Attitudes Measures in Surveys: The Are Rarely What They Appear to Be, Survey nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.87-100.

LAMBERT, D. (1992) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, Technometrics, 34, S.1-14.

LENSVELT-MULDERS, G. (2008) Surveying sensitive topics, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.461-478.

LESSLER, J.T. und W.D. KALSBECK (1992) Nonsampling Errors in Surveys, Wiley, New York.

LIEVESLEY, D. (1986) Unit Non Response in Interview Surveys, Joint Centre for Survey Methods, SCPR, London.

LITTLE, R.J.A. (1988) Missing-Data adjustments in large surveys, Journal of Business and Economic Statistics, 6, S.287-296.

LITTLE, R.J.A. und D.B. RUBIN (2002) Statistical analysis with missing data, Wiley, Hoboken.

LITTLE, R.J.A. und N. SCHENKER (1994) Missing data, Handbook for Statistical Modeling in the Social and Behavioral Sciences. (Hrsg. Arminger, G. et al.), Plenum, New York, S.39-75.

LOHR, S. (2008) Coverage and Sampling, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.97-112.

LONGFORD, N.T. (2000) Multiple imputation in an internation database of social science surveys, ZA-Information, 46, S.72-95.

LONGFORD, N.T. (2008) Studying Human Population. An Advanced Course in Statistics, Springer, New York.

LOOSVELDT, G. (1998) Interaction characteristics of the difficult-to-interview respondents, International Journal of Public Opinion Research, 9/4, S.385-394.

LOOSVELDT, G. (2008) Face-to-face interviews, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.201-220.

LOOSVELDT, G. et al. (1998) The effect of interviewer and respondent characteristics on refusals in a panel survey, Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.249-262.

LOOSVELDT, G. et al. (2002) Item nonresponse as a predictor of unit nonresponse in a panel survey. Journal of official statistics, 18/4, S.545-557.

LOOSVELDT, G. et al. (2007) Possibilities of evaluating interviewer effects through the use of interviewer variance, Measuring meaningful data in social research (Hrsg. Loosveldt, G. et al.), Acco, Leuven, S.177-194.

LÜESSE, T. (2007) Bürgerverantwortung und abnehmende Wahlbeteiligung, Lang, Frankfurt am Main.

LYNN, P. (1998) Data Collection Mode Effects on Responses to Attitudinal Questions, Journal of Official Statistics, 14/1, S.1-14.

LYNN, P. (2008) The Problem of Nonresponse, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.35-55.

MAAS, C.F. und W.F. DE HEER (1995) Response developments and the fieldwork strategy, Bulletin de Methodologie Sociologique 48, S.36-51.

MADOW, W.G. et al. (1983) Incomplete Data in Sample Surveys. Report and Case Studies, New York.

MARTIN, J. et al. (1993) The Use of CAPI for Attitude Surveys: An Experimental Comparison with Traditional Methods, Journal of Official Statistics, 9/3, S.641-661.

MCCULLAGH, P. und J. NELDER (1989) Generalized Linear Models, Chapman and Hall/CRC, Boca Raton.

MCKNIGHT, P.E. et al. (2007) Missing Data. A gentle introduction, Guilford Press, New York.

MENG, X.-L. (2003) A congenial overview and investigation of multiple imputation inferences under uncongeniality, Survey Nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.343-356.

MENOLD, N. und C. ZÜLL (2010) Codierung von Gründen der Verweigerung der Teilnahme an Interviews: ein Kategorienschema, GESIS-Technical Report, 2010/11.

MOHLER, P. ET AL. (2008) Survey documentation: Towards professional knowledge management in sample surveys, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.403-420.

NORMAN, D.A. (1982) Learning and memory, Freeman & Company, New York.

NEALON, J. (1983) The Effects of Male vs. Female Telephone Interviewers, Statistical Reporting Service, U.S. Department of Agriculture. Report AGES830617, Washington.

NORDHOLT, E.S. (1998) Imputation. Methods, Simulation Experiments, and Practical Examples, International Statistical Review, 66/2, S.157-180.

OH, H.L. und F.J. SCHEUREN (1983) Weighting adjustments for unit non-response, Incomplete Data in sample surveys: Theory and bibliographies (Hrsg. Madow, W.G. et al.) (Band 2), Academic Press, New York, S.143-184.

PICKERY, J. und G. LOOSVELDT (1998) The impact of respondent and interviewer characteristics on the number of ‚no opinion'answers. A multilevel model for count data, Quality and Quantity, 32, S.31-45.

PICKERY, J. und G. LOOSVELDT (2001) An exploration of question characteristics that mediate interviewer effects on item nonresponse, Journal of Official Statistics, 17/3, S.337-350.

PÖSCHL, H. (1993) Werbung und Beteiligung der Haushalte an der Einkommens- und Verbrauchsstichprobe 1993, Wirtschaft und Statistik, 6, S.385-390.

PÖTSCHKE, M. und C. MÜLLER (2006) Erreichbarkeit und Teilnahmebereitschaft in Telefoninterviews: Versuch einer mehrebenenanalytischen Erklärung, ZA-Informationen, 59, S.83-99.

PORST, R. (1996) Ausschöpfungen bei sozialwissenschaftlichen Umfragen. Die Sicht der Institute, ZUMA-Arbeitsbericht, 1996/7.

PORST, R. und C. VON BRIEL (1995) Wären Sie vielleicht bereit, sich gegebenenfalls wiederbefragen zu lassen? ZUMA-Arbeitsbericht 4.

PORST, R. und M. SCHNEID (1988) Ausfälle und Verweigerungen bei Panelbefragungen – ein Beispiel. ZUMA-Arbeitsbericht 12.

RAGHUNATHAN, T.E. (2000) Comment, Journal of the American Statistical Association, 95, S.85-87.

RÄSSLER, S. (2000) Ergänzung fehlender Daten in Umfragen, Jahrbücher für Nationalökonomie, 220, 1, S.64-94.

RÄSSLER, S. und R. SCHNELL (2004) A comparison of multiple imputation and other unit-nonresponse compensating techniques in fear of crime studies, Toronto, Joint Statistical Meeting.

REDER, L.M. (1988) Strategic control of retrieval strategies, The Psychology of Learning and Motivation (Hrsg. Bower, G.) (Band 22), Academic Press, New York, S.227-259.

REINECKE, J. (1991) Interviewer- und Befragtenverhalten. Theoretische Ansätze und methodische Konzepte, Westdeutscher Verlag, Opladen.

REUBAND, K.-H. (1984) Dritte Personen beim Interview – Zuhörer, Adressaten oder Katalysatoren der Kommunikation? Soziale Realität im Interview (Hrsg. Meulemann, H. und K.-H. Reuband), Campus Verlag, Frankfurt, S.117-156.

REUBAND, K.-H. (2006) Postalische Befragung alter Menschen. Kooperationsverhalten, Beantwortungsstrategien und Qualität der Antworten, ZA-Informationen, 59, S.100-127.

RÖSCH, G. (1994) Kriterien der Gewichtung einer nationalen Bevölkerungsstichprobe, Gewichtung in der Umfragepraxis (f. Gabler, S. et al.), Westdeutscher Verlag, Opladen, S.7-26.

ROGELBERG , S.G. und A. LUONG (1998) Nonresponse to mailed surveys: A review and guide, Current Directions in Psychological Science, 7, S.60-65.

ROTHE, G. (1994) Wie (un)wichtig sind Gewichtungen? Eine Untersuchung am ALLBUS 1986, Gewichtung in der Umfragepraxis (Hrsg. Gabler, S. et al.), Westdeutscher Verlag, Opladen, S.62-87.

ROTHE, G. und M. WIEDENBECK (1994) Stichprobengewichtung: Ist Repräsentativität machbar? Gewichtung in der Umfragepraxis (Hrsg. Gabler, S. et al.), Westdeutscher Verlag, Opladen, S.46-61.

ROYSTON, P. (2005) Multiple imputation of missing values: Update, The Stata Journal, 5, S.1-14.

RUBIN, D.B. (1976) Inference and missing data, Biometrika, 63/3, S.581-592.

RUBIN, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations, Journal of Business and Economic Statistics, 4, S.87-94.

RUBIN, D.B. (1987) Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

RUBIN, D.B. et al. (1995) Handling „Don't Know" Survey Responses: The Case of the Slovenian Plebiscite, Journal of the American Statistical Association, 90, S.822-828.

SACKSHAUG, J. und F. KREUTER (2011) Using Paradata and Other Auxiliary Data to Examine Mode Switch Nonresponse in a „Recruit-and-Switch" Telephone Survey, Journal of Official Statistics, 27/2, S.339-357.

SCHÄFER, A. (2009) Alles halb so schlimm? Warum eine sinkende Wahlbeteiligung der Demokratie schadet, MPIfG Jahrbuch, S.5-10.

SCHAFER, J.L. (1997a) Analysis of Incomplete Multivariate Data, Chapman & Hall, London.

SCHAFER, J.L. (1997b) Imputation of missing covariates under a general linear mixed model, Technical report, Deptartment of Statistics, Penn State University.

SCHANZ, V. und P. SCHMIDT (1984) Interviewsituation, Interviewmerkmale und Reaktionen von Befragten im Interview: eine multivariate Analyse, Allgemeine Bevölkerungsumfrage der Sozialwissenschaften – Beiträge zu methodischen Problemen des ALLBUS 1980 (Hrsg. Mayer, K.U. und P. Schmidt), Campus, Frankfurt, S.72-113.

SCHEUREN, F.J. (2004) What is a survey? http://www.whatisasurvey.info.

SCHNABEL, C. und J. WAGNER (2005) Who Are the Workers Who Never Joined a Union? Empirical Evidence from Germany, IZA Paper, 1698.

SCHNELL, R. (1991) Wer ist das Volk? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen“: Undercoverage, Schwererreichbare und Nichtbefragbare, Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43, S.106-137.

SCHNELL, R. (1993) Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität“ und Gewichtungsverfahren, Zeitschrift für Soziologie, 22, S.16-32.

SCHNELL, R. (1997) Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen, Leske+Budrich, Opladen.

SCHUMAN, H. und S. PRESSER (1981) Questions and answers in attitude surveys, Academic Press, New York.

SCHWARZ, N. et al. (2008) The Psychology of Asking Questions, Interational Handbook of Survey Methodology (Hrsg. De Leeuw, E.D. et al.), Taylor & Francis, New York, S.18-34.

SINGER, E. et al. (1998) Does the Payment of lncentives Create Expectation Effects? Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.229-237

SINGER, E. (2002) The Use of Incentives to Reduce Nonresponse in Household Surveys, Survey Nonresponse (Hrsg. Groves, R.M. et al.), Wiley, New York, S.163-177.

SKRONDAL, A. und S. RABE-HESKETH (2008) Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models, Stata Press, College Station Texas.

SOMMER, R. (1987) Der Mythos der Ausschöpfung, Planung und Analyse, 14, S.300-301.

SPIESS, M. (2008) Missing-Data Techniken. Analyse von Daten mit fehlenden Werten, Lit, Hamburg.

STEEH, C.G. (1981) Trends in Nonresponse Rates 1952-1979, Public Opinion Quarterly, 45, S.40-57.

STEINER, H. (1984) Das Interview als soziale Interaktion, Soziale Realität im Interview (Hrsg. Meulemann, H. und K.-H. Reuband), Campus Verlag, Frankfurt, S.17-59.

STENGER, H. (1994) Anforderungen an eine repräsentative Stichprobe aus der Sicht des Statistikers, Gewichtung in der Umfragepraxis (Hrsg. Gabler, S. et al.), Westdeutscher Verlag, Opladen, S.42-45.

STOCKÉ, V. und C. HUNKLER (2004) Die angemessene Erfassung der Stärke und Richtung von Anreizen durch soziale Erwünschtheit, ZA-Information, 54, S.53-88.

STOOP, I. (2004) Surveying Nonrespondents, Field Methods, 16/1, S.23-54.

STOOP, I. (2007) No time, too busy. Time strain and survey cooperation. Measuring Meaningful Data in Social Research (Hrsg. Loosveldt, G. et al.), Acco, Leuven, S.301-314.

STOOP, I. et al. (2010) Improving Survey Response. Lessons Learned from the European Social Survey, Wiley, Chichester.

TERWEY, M. (2000) ALLBUS: A German General Social Survey, Schmollers Jahrbuch, 120, S.151-158.

THIESSEN, V. und J. BLASIUS (1998) Using Multiple Correspondence Analysis to Distinguish between Substantive and Nonstubstantive Response, Visualization of Categorical Data (Hrsg. Blasius, J. und M. Greenacre), Academic Press, San Diego, S.239-252.

TOURANGEAU, R. et al. (2000) The Psychology of Survey Nonresponse, Cambridge University Press, Cambridge.

TRAUGOTT, M.W. et al. (1987) Using dual frame designs to reduce nonresponse in telephone surveys, Public Opinion Quarterly, 51, S.522-539.

VAN BUUREN, S. und K. GROOTHUIS-OUDSHOORN (2009) MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, forthcoming, http://www.stefvanbuuren.nl/publications/MICEinR-Draft.pdf.

WATZLAWICK, P. et al. (2011) Menschliche Kommunikation: Formen, Störungen, Paradoxien, Huber, Bern.

WASMER, M. und A. KOCH (2002) Konzeption und Durchführung der PAPI-Methodenstudie zur "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2000, ZUMA Methodenbericht, 2002/1.

WASMER, M. et al. (2007) Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2006, ZUMA Methodenbericht, 2007/9.

WILLIAMS, B. et al. (2007) When „No" Might Not Quite Mean „No„. The Importance of Informed and Meaningful Non-Consent: Results from a Survey of Individuals Refusing Participation in a Health-Related Research Project, BMC Health Services Research, 7.

WILLIS, G.B. et al. (1999) Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques, Cognition and Survey (Hrsg. Sirken M. et al.), Wiley, New York, S.133-153.

WHITE, A. et al. (1998) Improving Advance Letters for Major Goverment Surveys, Nonresponse in Survey Research. Proceedings of the Eighth International Workshop on Household Survey Nonresponse 24-26 September 1997 (Hrsg. Koch, A. und R. Porst), Mannheim, S.151-171.

WISEMAN, F. und M. BILLINGTON (1984) Commet on a Standard Definition of Response Rates, Journal of Marketing Research, 21, S.336-338.

WINKELMANN, R. (2008) Econometric Analysis of Count Data, Springer, Berlin.

ZEH, J. (1976) Der Verzerrungsfehler durch Ausfälle bei Meinungsbefragungen, Dissertation, Bonn.

UNIVERSITY OF BAMBERG PRESS

Fehlende Werte bilden seit jeher eine große methodische Herausforderung für quantitative Analysen in den Sozialwissenschaften. Die Methodenforschung deutet darauf hin, dass sich in einigen Bereichen das Problem fehlender Werte noch verschärfen wird. In dieser Arbeit werden zunächst die wichtigsten Theorien zu Item und Unit Nonresponse zusammengefasst. Anhand von Beispielen aus dem ALLBUS werden anschließend Analysen zu Item und Unit Nonresponse vorgenommen. Im Weiteren werden Korrekturmethoden für fehlende Werte durch ein auf realen Daten basierendes neues Verfahren verglichen. Dabei steht im Zentrum des Methodenvergleichs das Abschneiden der Multiplen Imputation als Korrekturmethode. Die Ergebnisse sprechen bei Item Nonreponse klar für die Anwendung der Multiplen Imputation vor allem bei multivariaten Analysen. Obwohl die Ergebnisse für Unit Nonresponse keine klaren Empfehlungen zulassen, erscheint der Weg, Multiple Imputation auch dort zu verwenden, bei günstiger Datenlage als Alternative zu Gewichtungen.

eISBN 978-3-86309-123-1

9 783863 091231

www.uni-bamberg.de/ubp/