

Masterarbeit
von
Claudia Baldermann

**METHODENVERGLEICH ZUR SCHÄTZUNG DES
ZEITEFFEKTS IM BUNDESGESUNDHEITSSURVEY**

Eingereicht bei
Prof. Dr. Ulrich Rendtel

am 10. April 2012

Freie Universität Berlin
Fachbereich Wirtschaftswissenschaft
Institut für Statistik und Ökonometrie

Masterstudiengang Statistik

Masterarbeit von Claudia Baldermann

Erstgutachter: Prof. Dr. Ulrich Rendtel
Zweitgutachterin: Angelika Schaffrath-Rosario, Dipl. Stat.

Claudia Baldermann

Matrikelnummer: 522601

Erich-Weinert-Str. 129

10409 Berlin

E-mail: claudia.baldermann@fu-berlin.de

Zusammenfassung

In dieser Arbeit werden statistische Methoden für die Analyse des Zeiteffekts im Bundesgesundheitsurvey (BGS) unter Berücksichtigung des komplexen Stichprobendesigns untersucht. Es werden modellbasierte und designbasierte Verfahren zur Analyse komplexer Stichproben vorgestellt. Das Generalisierte Lineare Modell, die Generalisierten Schätzgleichungen und das Generalisierte Lineare Gemischte Modell werden zur Schätzung des Zeiteffekts vorgeschlagen. Die Umsetzbarkeit der vorgestellten Methoden wird für die Statistik-Software SAS und STATA untersucht. In einer Simulation werden die drei Methoden hinsichtlich Verzerrung und Effizienz für stetige und binäre abhängige Merkmale verglichen. Zusätzlich werden die Methoden für Beispieldaten aus dem BGS untersucht. Basierend auf den Erkenntnissen der Simulation und der Beispieluntersuchung wird die Schätzung des Zeiteffekts durch designbasierte Generalisierte Lineare Modelle empfohlen.

Inhaltsverzeichnis

1	Einleitung	1
2	Das Stichprobendesign im BGS	3
2.1	Die erste Erhebung	3
2.2	Die zweite Erhebung	6
3	Theoretischer Rahmen	8
3.1	Modellbasierte Analyse von Surveydaten	8
3.2	Designbasierte Analyse von Surveydaten	13
3.2.1	Schätzung des Totalwertes	14
3.2.2	Pseudo-Maximum-Likelihood-Ansatz	19
4	Methoden zur Schätzung des Zeiteffekts	23
4.1	Generalisiertes lineares Modell	24
4.1.1	Arbeitsmodell und Schätzgleichung	24
4.1.2	Implementierung in SAS und STATA	25
4.2	Generalisierte Schätzgleichung	27
4.2.1	Arbeitsmodell und Schätzgleichung	27
4.2.2	Implementierung in SAS und STATA	30
4.3	Generalisierte lineare gemischte Modelle	31
4.3.1	Arbeitsmodell und Schätzgleichung	31
4.3.2	Implementierung in SAS und STATA	34
4.4	Zusammenfassung	35
5	Simulationsstudie	36
5.1	Vorüberlegungen	36
5.1.1	Modell der Grundgesamtheit	36
5.1.2	Abhängigkeitsstruktur in der Grundgesamtheit	38
5.1.3	Generierung der Grundgesamtheit	39

5.1.4	Stichprobenziehung	42
5.2	Vergleichskriterien für die Simulationsstudie	43
5.3	Resultate	44
5.4	Beispieldaten aus dem BGS	54
6	Abschließende Bemerkungen	58
	Literaturverzeichnis	59

Tabellenverzeichnis

5.1	In der Simulation verwendete Kombinationen von ICC_c und ICC_z und dazugehörige Standardfehler der zufälligen Effekte für metrische und dichotome Merkmale	39
5.2	Populationsparameter für binäre abhängige Variablen	41
5.3	Simulationsergebnisse für normalverteilte Merkmale für gewichtete Schätzungen.	45
5.4	Simulationsergebnisse für normalverteilte Merkmale für ungewichtete Schätzungen.	46
5.5	Standardfehler in Relation zu $WGLM_1$ für normalverteiltes Merkmal	48
5.6	Simulationsergebnisse für Merkmale mit log-normalverteiltem Fehlerterm für gewichteten Schätzungen	49
5.7	Simulationsergebnisse die Merkmale mit log-normalverteiltem Fehlerterm für ungewichteten Schätzungen.	50
5.8	Standardfehler in Relation zu $WGLM_1$ für die Merkmal mit log-normalverteiltem Fehlerterm	51
5.9	Simulationsergebnisse für binäre Merkmale für gewichtete Schätzungen.	52
5.10	Simulationsergebnisse für binäre Merkmale für ungewichtete Schätzungen.	53
5.11	Standardfehler in Relation zu $WGLM_1$ für binäre Merkmale	53
5.12	Geschätzte Zeiteffekte und Standardfehler für stetige Merkmale aus dem BGS	55
5.13	Geschätzte Zeiteffekte und Standardfehler für binäre Merkmale aus dem BGS	56

Abbildungsverzeichnis

2.1	Alte und neue Sample Points im BGS	4
3.1	Superpopulation und endliche Grundgesamtheiten	9
5.1	Eigenschaften der endlichen Grundgesamtheit und des Stichprobendesigns . .	43

1 Einleitung

Das Bundesgesundheitsurvey (BGS) ist eine repräsentative Erhebung zur Untersuchung des Gesundheitszustandes aller in der Bundesrepublik Deutschland lebenden 18- bis 79-Jährigen. Die Erhebung wurde vom Robert-Koch-Institut (RKI) im Auftrag des Bundesministerium für Gesundheit durchgeführt. Das Ziel der bundesweiten Befragung ist Daten für die Bearbeitung epidemiologischer Fragestellungen bereitzustellen. Die Befragungen und ärztlichen Untersuchungen begannen im Oktober 1997 und endeten im März 1999. Das BGS heißt heute „Studie zur Gesundheit Erwachsener in Deutschland“ (DEGS) und befindet sich seit November 2008 in seiner ersten Folgerhebung. In der Folgerhebung wurden auch Probanden vom ersten Erhebungszeitraum wieder befragt, sodass längsschnittliche Analysen möglich werden.

Im Fokus der Datenauswertung steht die Analyse des Zeiteffekts. Hierbei wird untersucht ob sich der Gesundheitszustand zwischen den beiden Erhebungen verändert hat. Diese Untersuchungen soll für stetige (z.B. Blutdruck), sowie für binäre (z.B. Atembeschwerden ja/nein) Gesundheitsmerkmale durchgeführt werden. Neben der Variable Zeit, die hier von primärem Interesse ist, sollen weitere Kontrollvariablen, wie Alter und Geschlecht in das Modell integriert werden.

Gegenstand dieser Arbeit ist die Untersuchung und der Vergleich geeigneter statistischer Methoden für die Analyse des Zeiteffekts im BGS unter Berücksichtigung des komplexen Stichprobendesigns. Insbesondere sollen drei Aspekte des Stichprobendesigns berücksichtigt werden: die Clusterstruktur; die Design- und Anpassungsgewichtung unter Berücksichtigung der Wiederteilnahmewahrscheinlichkeit von Längsschnittprobanden und die longitudinale Korrelation innerhalb von Längsschnittprobanden. Die Herausforderung besteht darin geeignete statistische Methoden zu finden, bei denen unter Berücksichtigung der Datenstruktur unverzerrte und effiziente Schätzer gewährleistet sind. Die Untersuchung zielt darauf ab praktikable Methoden für die Schätzung des Zeiteffekts zu finden. Eine Empfehlungen sollte insbesondere die Umsetzbarkeit in bekannten Statistischen Softwarepaketen beinhalten. Die Umsetzbarkeit der untersuchten Methoden wird für die Statistik-Software SAS und STATA untersucht.

Im zweiten Kapitel wird das Stichprobendesign im BGS beschrieben. Zuerst wird auf die zweistufig geschichtete Ziehung der Stichprobe der ersten Erhebung eingegangen, gefolgt vom

ebenfalls zweistufigem Design der Folgestudie.

Im dritten Kapitel folgen theoretische Grundlagen zur Analyse von Surveydaten. Es wird auf modell- und designbasierte Analysemethoden eingegangen. Zusätzlich wird Der Pseudo-Maximum-Likelihood Ansatz für die designbasierte Analyse von Surveydaten eingeführt.

Im vierten Kapitel werden drei Methoden vorgestellt mit deren Hilfe der Zeiteffekt geschätzt werden kann. Diese bauen für designbasierte Analysen auf dem Prinzip des Pseudo-Maximum-Likelihood Ansatzes auf. Es werden zunächst die lineare Regression für metrische und die logistische Regression für binäre abhängige Variablen vorgestellt. Anschließend wird die Methode der verallgemeinerten Schätzgleichung als Lösungsmöglichkeit diskutiert. Als eine dritte Methode wird das generalisierte gemischte lineare Modell mit Designgewichten für die verschiedenen Ziehungsstufen vorgestellt. Zu jedem Verfahren wird auf Implementierungen in den Programmen SAS und STATA verwiesen.

Anschließend werden in Kapitel 5 die vorgestellten Methoden anhand einer Simulationsstudie untersucht. Hierfür wird eine vereinfachte Grundgesamtheit generiert. Die wesentlichen Eigenschaften der Grundgesamtheit im BGS werden jedoch aufgegriffen, um eventuelle Rückschlüsse auf die Parameterschätzung im BGS ziehen zu können. Im Anschluss daran wird der Zeiteffekt für einige Testvariablen aus dem BGS-Datensatz geschätzt. Es wird untersucht, ob Erkenntnisse aus der Simulationsstudie sich auch in den BGS-Daten wiederfinden.

Im sechsten Kapitel folgen einige abschließende Bemerkungen.

2 Das Stichprobendesign im BGS

Die Befragung im BGS besteht aus der Beantwortung eines Fragenkatalogs und einer medizinischen Untersuchung der gezogenen Probanden. In der medizinischen Untersuchung werden unter anderem Körpergewicht, Körpergröße, Blutdruck und Aspekte körperlicher Fitness erfasst. Die Zielpopulation im BGS sind die in der Bundesrepublik lebenden 18- bis 79-Jährigen. Die erhobenen Daten sollen Aufschluss über Volkskrankheiten, gesundheitsbezogene Risikofaktoren oder auch eventuelle Versorgungsengpässe im Gesundheitsbereich geben. Durch wiederholte Befragungen können zusätzlich längsschnittliche Analysen durchgeführt werden. Hierbei sind kausale Zusammenhänge zwischen Risikofaktoren und gesundheitlichen Problemen und die Analyse von Krankheitsverläufen von besonderem Interesse (RKI, 2009). Nachfolgend wird die Abkürzung BGS98 für die Basiserhebung aus den Jahren 1997 bis 1999 verwendet und DEGS1 für die erste Folgerhebung, die im Jahr 2008 begann.

2.1 Die erste Erhebung

Für das BGS98 wurde ein zweistufig geschichtetes Stichprobendesign entwickelt. In der ersten Stufe wurden die Gemeinden Deutschlands nach den 16 Bundesländern und 10 Gemeindetypen geschichtet, aus denen insgesamt 120 Untersuchungsorte (nachfolgend Sample Points genannt) gezogen wurden. Um anschließend aussagekräftige Ost/West-Vergleiche machen zu können, wurde ein Oversampling für die neuen Bundesländer angestrebt. Aus den neuen Bundesländern sollten dabei insgesamt 40 und aus den alten Bundesländern insgesamt 80 Sample Points gezogen werden (RKI, 2009).

Die 10 Gemeindetypen werden durch die 10-stufige BIK-Klassifikation (nachfolgend BIK10-Klassifikation genannt) definiert. Die BIK10-Klassifikation schichtet die Gemeinden in Größenklassen. Die Größe wird hierbei mit der Zahl der erwerbstätigen Einwohner gemessen.¹ Jeder der 160 Schichtungszellen (16 Bundesländer \times 10 BIK-Klassen) wurde ein Bedeutungsgewicht, proportional zur Bevölkerungszahl in der Zielpopulation in der jeweiligen Zelle zu-

¹Eine detaillierte Beschreibung der BIK-Klassifikation liefern die Produktbeschreibungen der GmbH BIK ASCHPURWIS + BEHRENS: <http://www.bik-gmbh.de>

geordnet. Diese Gewichte wurden für die alten Bundesländer auf die Summe 80 und für die neuen Bundesländern auf die Summe 40 normiert. Mit Hilfe des Cox-Algorithmus² wurden die Bedeutungsgewichte auf ganze Zahlen gerundet, sodass für jede Schichtungszelle die Anzahl der auszuwählenden Gemeinden feststand.

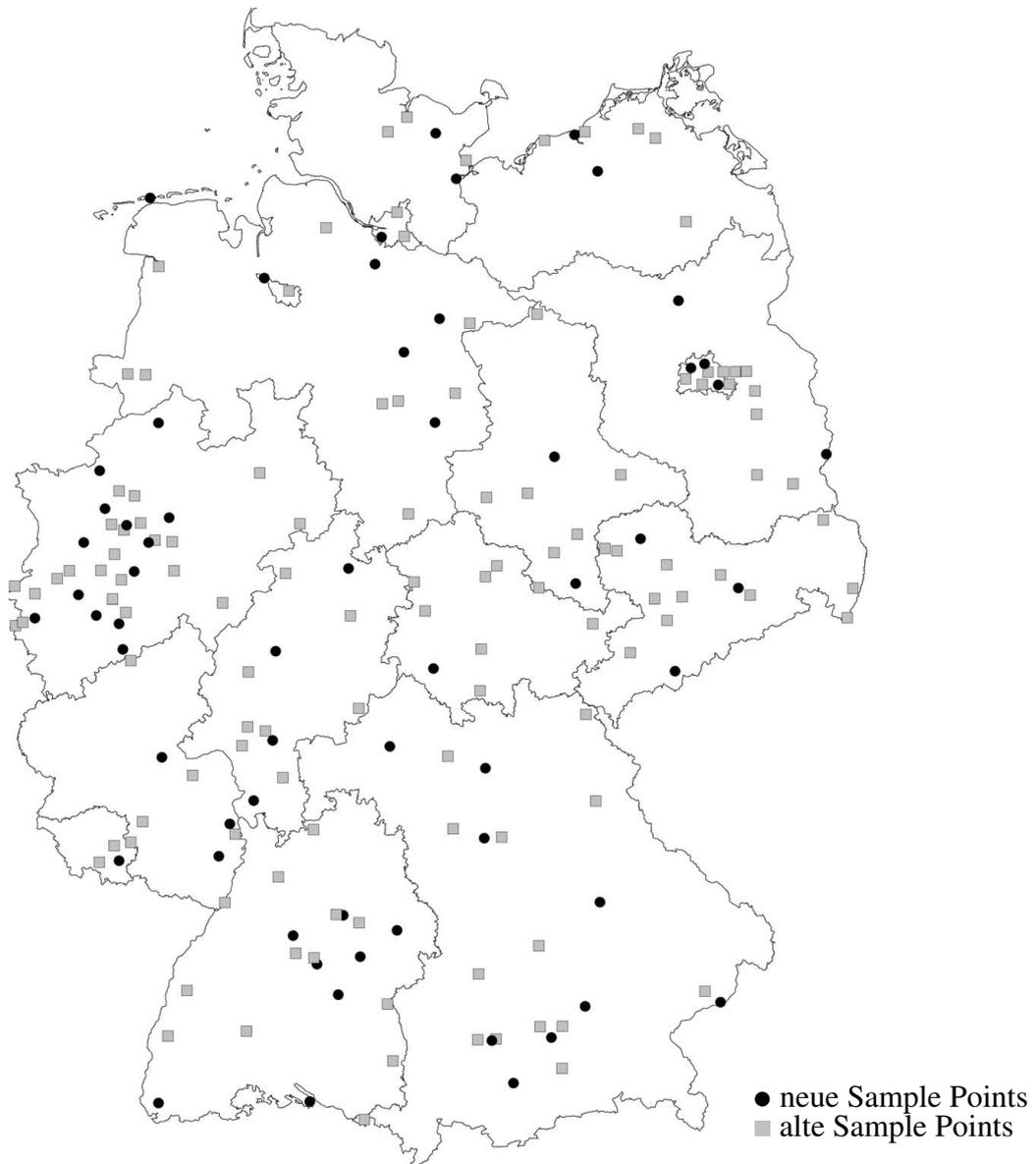


Abbildung 2.1: Alte und neue Sample Points im BGS
Quelle: RKI (2012)

²Eine detaillierte Beschreibung dieser Rundungsmethode ist in Cox (1987) nachzulesen.

Die 120 Sample Points wurden aus dem aktuellen Gemeindeverzeichnis mit einer Wahrscheinlichkeit proportional zu ihrer Größe gezogen. Die Größe misst sich hierbei an der Anzahl der in der Gemeinde gemeldeten 18- bis 79-jährigen Personen (Stand 31.12. 1995). Innerhalb der Gemeinden wurde ein Stadtteil oder Wahlbezirk zufällig ausgewählt, der den Sample Point repräsentiert. Bei Gemeinden mit bis zu 50.000 Einwohnern repräsentieren die Sample Points die gesamte Gemeinde (Stolzenberg, 2000). Die Abbildung 2.1 zeigt die Verteilung der gezogenen Sample Points über das Gebiet der Bundesrepublik Deutschland. Die Sample Points aus dem BGS98 sind mit einem grauen Kästchen gekennzeichnet.

In der zweiten Stufe wurden mit Hilfe der Melderegister der ausgewählten Gemeinden jeweils die gleiche Anzahl an Personenadressen, geschichtet nach Altersgruppen³ zufällig gezogen und angeschrieben. Insgesamt wurden 13.222 Probanden angeschrieben, von denen 7.124 teilnahmen. Im Mittel wurden 59 Probanden pro Sample Point untersucht. Die Feldarbeit erstreckte sich über drei Jahre, wobei jeder Sample Point dreimal im Abstand von einer Woche aufgesucht wurde. Dieses Vorgehen erleichterte den potentiellen Probanden die Terminfindung und erhöhte somit ihre Teilnahmeneigung. Aufgrund der langwierigen Feldarbeit, wurden die Personenadressen erst kurz vor dem vorgesehenen Untersuchungszeitraum des jeweiligen Sample Points gezogen, um die Aktualität der Adresdaten zu gewährleisten (Potthoff, Schöder, Reis und Klamert, 1999).

Grundsätzlich ist dieses zweistufige Verfahren mit Ziehungswahrscheinlichkeiten der ersten Stufe proportional zur Größe und einer konstanten Anzahl zu ziehender Probanden in jedem Sample Point selbstgewichtigend. Die Teilnahmebereitschaft der gezogenen Probanden fiel jedoch für die Sample Points und für verschiedene Altersklassen sehr unterschiedlich aus. Es wurde unter anderem festgestellt, dass die Teilnehmerate in Großstädten mit mehr als 500.000 Einwohnern deutlich geringer ausfiel als in kleineren Gemeinden (Thefeld, Stolzenberg und Bellach, 1999). Bezüglich der Altersgruppen zeigten beispielsweise die 20- bis 29-Jährigen eine geringere Teilnahmebereitschaft als die übrigen Altersgruppen (Stolzenberg, 2000).

Auf Grund der variierenden Responseraten wurden Gewichte für die beobachteten Probanden eingeführt. Diese werden über das Produkt der inversen Auswahlwahrscheinlichkeiten der ersten und der zweiten Stufe berechnet. Die Auswahlwahrscheinlichkeit der zweiten Stufe ist die Wahrscheinlichkeit eines Probanden gezogen zu werden, bedingt aus die vorherige Ziehung des Sample Points.

³Die Altersgruppen sind durch eine 10-Jahres-Gruppierung definiert.

2.2 Die zweite Erhebung

Das Stichprobendesign der Folgerhebung DEGS1 entspricht ebenfalls einer zweistufig geschichteten Zufallsauswahl. Um eine wiederholte Befragung der Probanden des BGS98 zu ermöglichen, wurden alle 120 Sample Points wieder in die Stichprobe aufgenommen. Zusätzlich wurden nach dem oben beschriebenen Ziehungsprinzip der ersten Stufe 60 neue Sample Points gezogen. Durch die Neuziehungen sollte sichergestellt werden, dass die insgesamt 180 Sample Points den aktuellen Bevölkerungsstand widerspiegeln. Ein Oversampling der neuen Bundesländer sollte im DEGS1 nicht fortgeführt werden (RKI, 2009). Die neu gezogenen Sample Points sind in Abbildung 2.1 mit einem schwarzen Punkt markiert.

In der zweiten Auswahlstufe wurden aus den neu gezogenen Sample Points wiederum Personenadressen, geschichtet nach Altersgruppen aus den Einwohnermelderegistern zufällig gezogen. Im DEGS1 wurde eine durchschnittliche Anzahl von 40 bis 45 Probanden pro Sample Point angestrebt (RKI, 2009). Aus den alten Sample Points wurde die Stichprobe ebenfalls mit neuen Probanden aufgestockt. Hierbei wurden vor allem Probanden nachgezogen, die im BGS98 noch nicht zur Zielpopulation gehörten. Das betrifft die damals unter 18-Jährigen, die zum Zeitpunkt der Folgerhebung ein Mindestalter von 18 erreicht haben. Etwa die Hälfte der ursprünglich Befragten aus dem BGS98 konnten für die Folgestudie DEGS1 wiedergewonnen werden.

Für die Gewichtung der Längsschnittprobanden muss die Wahrscheinlichkeit der erneuten Teilnahme berücksichtigt werden. Die Wiederteilnahmewahrscheinlichkeit kann mit Hilfe der Daten aus dem BGS98 modelliert werden. Hierbei wird untersucht, ob nicht erneut teilnehmende Probanden sich bezüglich der im BGS98 erhobenen Merkmale von wieder teilnehmenden Probanden unterscheiden. Die in der Zwischenzeit verstorbenen Probanden werden aus dieser Untersuchung ausgeschlossen. Ebenso werden Probanden, die im BGS98 älter als 69 Jahre alt waren, aus der Modellierung der Wiederteilnahme ausgeschlossen. Die Wiederteilnahmewahrscheinlichkeit kann mit Hilfe einer logistischen Regression geschätzt werden. Vorläufig konnte festgestellt werden, dass Männer und Probanden mit höherem Sozialstatus häufiger wiederteilnahmebereit waren. Probanden mit adipöser Neigung, Hypertoniker und tägliche Raucher waren hingegen seltener bereit, wieder teilzunehmen.⁴ Das Designgewicht der Längsschnittprobanden ist das Produkt aus den inversen Ziehungswahrscheinlichkeiten der ersten und zweiten Stufe und der geschätzten inversen Wiederteilnahmewahrscheinlichkeit.

Abschließend wird eine gemeinsame Anpassungsgewichtung für die Längs- und Querschnitts-

⁴Diese Resultate beziehen sich auf vorläufige Ergebnisse interner Berechnungen am RKI. Die Zweitbetreuerin dieser Arbeit, Frau Schaffrath Rosario, hat diese Informationen bereitgestellt.

probanden der beider Erhebungen durchgeführt. Die demografischen Merkmale Alter und Geschlecht werden hierbei pro Bundesland an bekannte Verteilungen aus der amtlichen Statistik auf den Bevölkerungsstand vom 31.12.2008 angepasst.

Durch das beschriebene Stichprobendesign und die Zusammenführung der Daten aus dem BGS98 und dem DESG1 entsteht eine komplexe Datenstruktur. Diese besteht aus zwei Komponenten: dem zweistufigen Design mit Ziehung der Sample Points in der ersten Stufe und der Longitudinalstruktur mit zwei abhängigen Beobachtungen für einen Teil der Probanden. Bevor Methoden zur Schätzung des Zeiteffekts in Kapitel 4 diskutiert werden, erfolgt im nächsten Kapitel eine Beschreibung des theoretischen Rahmens für die Analyse von komplexen Stichproben.

3 Theoretischer Rahmen

Zunächst werden einige Begriffe und Notationen, die im Verlauf dieser Arbeit verwendet werden eingeführt. $U = \{1, \dots, i, \dots, N\}$ sei eine endliche Grundgesamtheit, die aus N Einheiten besteht. Jedem Element i der endlichen Grundgesamtheit kann ein Wert des interessierenden Merkmals y_i und ein p -dimensionaler Vektor erklärender Variablen \mathbf{x}_i zugeordnet werden. Die zufällig gezogene Stichprobe s mit einer Größe von n ist eine Teilmenge von U . Das komplexe Stichprobendesign ist durch den Inklusionsindikator $\mathbf{I} = (I_1, \dots, I_N)$ festgelegt, wobei I_i eine dichotome Variable ist, mit $I_i = 1$, wenn Element i in die Stichprobe s gezogen wurde und $I_i = 0$ wenn Element i nicht enthalten in s ist.

Die Suche nach einem geeigneten Verfahren für die Analyse von komplexen Surveydaten beinhaltet die Entscheidung darüber wie das Stichprobendesign in der Analyse berücksichtigt wird (Korn und Graubard, 1995; Binder und Roberts, 2004).

Stichproben aus einem komplexen Ziehungsmechanismus können auf Basis modellbasierter und designbasierter Verfahren analysiert werden. Traditionell sind in der designbasierten Analyse von Surveydaten deskriptive Parameter der endlichen Grundgesamtheit Ziel statistischer Inferenz. Im modellbasierten Ansatz besteht das Hauptinteresse darin, Schlüsse über Parameter eines übergeordneten stochastischen Prozesses, der die Merkmale der Grundgesamtheit generiert ziehen zu können (Skinner, 1989; Binder und Roberts, 2009). Die theoretischen Annahmen und die unterschiedlichen Sichtweisen hinter diesen beiden Ansätzen werden in den folgenden beiden Abschnitten beschreiben.

3.1 Modellbasierte Analyse von Surveydaten

Der modellbasierte Ansatz fußt auf der Annahme, dass die Merkmale y_i der endlichen Grundgesamtheit das Produkt eines hypothetischen stochastischen Prozesses sind. Dieser Prozess beschreibt einen angenommenen Zusammenhang zwischen dem Zielmerkmal y_i und den erklärenden Variablen \mathbf{x}_i . Als Beispiel für diesen Zusammenhang dient in diesem Kapitel ein

lineares Modell:

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad (3.1)$$

wobei ε_i eine unabhängige normalverteilte Zufallsvariable ist mit Erwartungswert 0 und Varianz σ^2 . Der p -dimensionale Vektor $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ beinhaltet die unbekanntenen Modellparameter, die den Zusammenhang zwischen y und \mathbf{x} beschreiben.

Die endliche Grundgesamtheit wird als eine zufällige Realisierung aus unendlich vielen möglichen Grundgesamtheiten angesehen. Der zugrunde liegende stochastische Prozess wird allgemein als Superpopulation bezeichnet. Das angenommene Modell der Superpopulation kann als Generierungsmechanismus der endlichen Grundgesamtheiten und insbesondere der Zielpopulation einer bestimmten Stichprobe gesehen werden. Die untersuchte Grundgesamtheit, aus der eine komplexe Stichprobe gezogen wurde, ist in diesem Ansatz eine mögliche Realisation aller endlichen Populationen (Binder und Roberts, 2004; Heeringa et al., 2010 S. 54 ff.). Abbildung 3.1 liefert eine Visualisierung dieses hypothetischen Prozesses, wobei die beiden dargestellten endlichen Populationen stellvertretend für eine unendlich große Anzahl endlicher Populationen abgebildet sind.

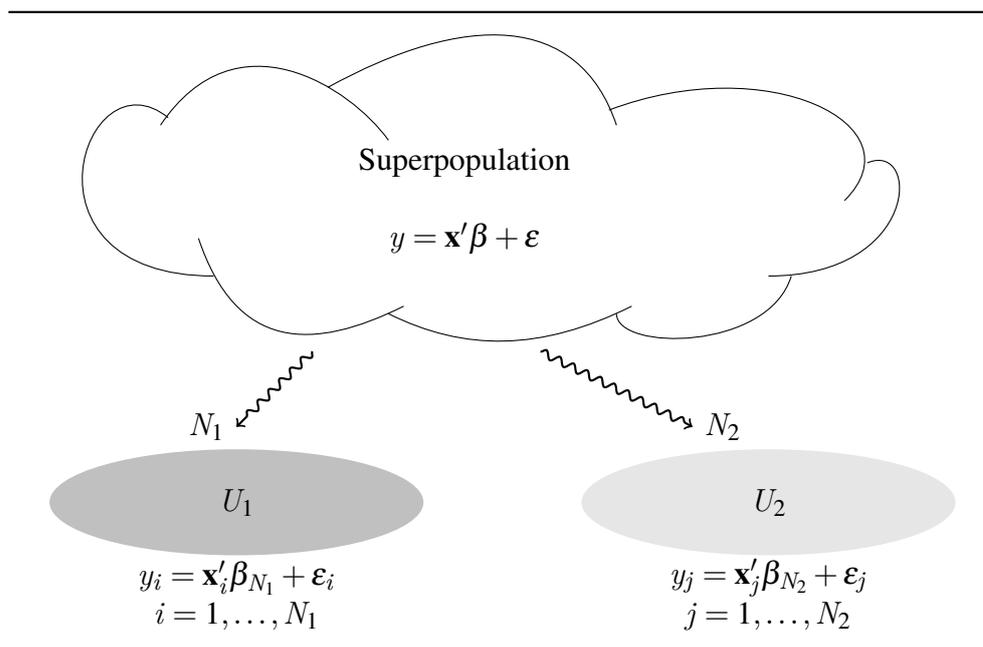


Abbildung 3.1: Superpopulation und endliche Grundgesamtheiten

Quelle: Eigene Darstellung in Anlehnung an Heeringa et al. (2010, S. 55)

Die endlichen Populationen können sich unterscheiden, beide wurden jedoch durch den

selben übergeordneten Prozess generiert. Die Realisierungen der Merkmale y_i werden demzufolge als Zufallsvariablen betrachtet, die aus der Verteilung der hypothetischen Superpopulation stammen. Aufgrund dieses Zufallsprozesses sind die jeweiligen Parameter β_N ebenfalls Zufallsvariablen und können zwischen den Populationen variieren. Im rein modellbasierten Ansatz ist die Variabilität von β_N auf die zufällige Realisierung der Grundgesamtheit zurückzuführen. Die gezogene Stichprobe hingegen wird als fest angenommen. Treffen die Modellannahmen zu, hat die zufällige Abweichung vom Parameter der Superpopulation ($\beta_N - \beta$) einen Erwartungswert von 0. Der Parameter β_N der Grundgesamtheit ist folglich ein unverzerrter Schätzer des Modellparameters β (Godambe und Thompson, 1986; Binder und Roberts, 2004).

Der Zielparameter statistischer Inferenz in der modellbasierten Schätzung ist β , der Modellparameter aus der Superpopulation.

Wäre es möglich die gesamte Zielpopulation zu beobachten, könnte der Parameter β durch β_N erwartungstreu geschätzt werden. Die Schätzung für β erfolgt über die Maximierung der Populationslikelihood. Die grundlegende Annahme für die Maximum-Likelihood-Schätzung ist, dass die Merkmale y_i der Grundgesamtheit unabhängige Realisationen aus der Superpopulation sind. Unter dieser Unabhängigkeitsannahme ergibt sich die Likelihoodfunktion der Grundgesamtheit aus dem Produkt der Einzeldichten (Fahrmeir et al. 2007, S. 93):

$$L(\beta_N) = \prod_{i \in U} f(y_i | \beta_N, \mathbf{x}_i), \quad (3.2)$$

wobei $f(y_i | \beta_N, \mathbf{x}_i)$ die Dichtefunktion des Merkmals y_i von Beobachtungseinheit i ist, bedingt auf die erklärenden Variablen \mathbf{x}_i . Unter der in Modellgleichung 3.1 angenommenen Normalverteilung hat $f(y_i | \beta_N, \mathbf{x}_i)$ die Form:

$$f(y_i | \beta_N, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (y_i - \mathbf{x}'_i \beta_N)^2 \right]. \quad (3.3)$$

Der logarithmierte Likelihoodbeitrag einer Beobachtung i hat die Form:

$$l(y_i | \beta_N, \mathbf{x}_i) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y_i - \mathbf{x}'_i \beta_N)^2. \quad (3.4)$$

Der Populationsparameter β_N ist der populationsbasierte Maximum-Likelihood-Schätzer (nachfolgend ML-Schätzer genannt) aus der endlichen Grundgesamtheit für β (Skinner, 1989). Der Parameter β_N wird so gewählt, dass die Populationslikelihood maximal wird.

Zu diesem Zweck wird die logarithmierte Populationslikelihood bezüglich β_N differen-

ziert und gleich Null gesetzt. Der Parametervektor β_N erfüllt das Gleichungssystem:

$$\mathbf{U}_N(\beta_N) = \sum_{i \in U} \mathbf{x}_i (y_i - \mathbf{x}_i' \beta_N) = \sum_{i \in U} \mathbf{u}_i(\beta_N) = \mathbf{0}, \quad (3.5)$$

wobei $\mathbf{U}_N(\beta_N)$ der p -dimensionale Vektor partieller Ableitungen der logarithmierten Populationslikelihood bezüglich β_N ist. $\mathbf{U}_N(\beta_N)$ wird als Scorefunktion bezeichnet, wobei $\mathbf{u}_i(\beta_N)$ den Scorebeitrag der Einheit i kennzeichnet.

Eine Maximierung des Terms $-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i' \beta_N)^2$ aus Gleichung 3.4 ist gleichbedeutend mit einer Minimierung von $(y_i - \mathbf{x}_i' \beta_N)^2$ (Fahrmeir et al. 2007 S. 93). Die ML-Schätzung β_N entspricht dem Kleinste-Quadrate-Schätzer (nachfolgend KQ-Schätzer genannt) für β aus der Modellgleichung 3.1:

$$\beta_N = \left(\sum_{i \in U} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' y_i. \quad (3.6)$$

Treffen die Modellannahmen aus Gleichung 3.1 zu, ist β_N eine unverzerrte und konsistente Schätzung für β (Skinner, 1989; Binder und Roberts, 2009).

Es wird angenommen, dass eine zufällige Stichprobe aus der endlichen Population gezogen wurde und die Zielvariable, bedingt auf die erklärenden Variablen, unabhängig vom Stichprobendesign ist. Der unbekannte Parametervektor β kann dann mit Hilfe der Beobachtungen aus der Stichprobe geschätzt werden:

$$\hat{\mathbf{U}}(\hat{\beta}) = \sum_{i \in U} I_i x_i (y_i - \mathbf{x}_i' \hat{\beta}) = \sum_{i=1}^n \mathbf{u}_i(\hat{\beta}) = \mathbf{0}. \quad (3.7)$$

Im Unterschied zu Gleichung 3.5 werden in diesem Ausdruck ausschließlich die gezogenen Beobachtungen aus der Stichprobe verwendet. Der geschätzte Parameter $\hat{\beta}$ ist der modellbasierte Stichproben-MLS für Modellparameter β und entspricht im linearen Fall einer stichprobenbasierten KQ-Schätzung (Binder und Roberts, 2009):

$$\hat{\beta} = \left(\sum_{i \in s} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in s} \mathbf{x}_i' y_i. \quad (3.8)$$

Die Eigenschaften der ML-Schätzung können durch eine Taylor-Linearisierung der Score-

funktion $\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}})$ an der Stelle $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ gezeigt werden:

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}) \approx \hat{\mathbf{U}}(\boldsymbol{\beta}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \frac{\delta \hat{\mathbf{U}}(\hat{\boldsymbol{\beta}})}{\delta \boldsymbol{\beta}}. \quad (3.9)$$

Nach Umstellen ergibt sich die Erwartungstreue der ML-Schätzung $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$:

$$E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} \sum_{i \in U} I_i E(\mathbf{u}_i(\boldsymbol{\beta})) = \mathbf{0}, \quad (3.10)$$

wobei $\mathbf{I}(\hat{\boldsymbol{\beta}}) = -\delta \hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}) / \delta \boldsymbol{\beta}$ die Informationmatrix der ML-Schätzung ist (Greene, 2008 S. 482). Die Varianz von $\hat{\boldsymbol{\beta}}$ ist durch die quadratische Form

$$V(\hat{\boldsymbol{\beta}}) = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} V(\hat{\mathbf{U}}(\boldsymbol{\beta})) \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} \quad (3.11)$$

definiert. Trifft die Verteilungsannahme der abhängigen Variable zu, entspricht die Inverse der Informationsmatrix $\mathbf{I}(\hat{\boldsymbol{\beta}})$ der Varianz der Scorefunktion. In dem Fall kann die Varianz von $\hat{\boldsymbol{\beta}}$ durch

$$V(\hat{\boldsymbol{\beta}}) = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} \quad (3.12)$$

ausgedrückt werden. Die Varianz aus Gleichung 3.11 ist robust gegenüber einer Verletzung der Verteilungsannahme.⁵

Modellbasierte Methoden liefern für komplexe Stichprobendesigns erwartungstreue und effiziente Schätzungen für $\boldsymbol{\beta}$, wenn angenommen werden kann, dass das Stichprobendesign vernachlässigbar ist. Dies gilt insbesondere, wenn die Verteilung der Zielvariable y_i in der Stichprobe der Verteilung in der endlichen Grundgesamtheit entspricht. Trifft dies nicht zu, können die Schätzungen der Parameter und ihrer Standardfehler verzerrt sein. Eine Gewichtung auf Grund unterschiedlicher Inklusionswahrscheinlichkeiten ist im modellbasierten Ansatz nicht vorgesehen (Binder und Roberts 2004; Kott, 1991).

Im modellbasierten Ansatz werden die Eigenschaften der Stichprobe weitestgehend ignoriert, es sei denn das Stichprobendesign wird direkt in das Modell integriert. Es können zum Beispiel Designvariablen, wie die Schicht- oder Clusterinformation als zusätzliche erklärende Variablen in das Modell aufgenommen werden. Bedingt auf diese Designvariablen wird das Stichprobendesign als vernachlässigbar angenommen (Pfeffermann und Sverchkov, 2009).

⁵Die Herleitung der asymptotischen Eigenschaften der ML-Schätzung geht über den Rahmen dieser Arbeit hinaus. Eine ausführliche Beschreibung der Maximum-Likelihood-Schätzung und die Herleitung der asymptotischen Eigenschaften bietet zum Beispiel Greene (2008, S 482 ff.)

Zusammenfassend werden die wichtigsten Annahmen der modellbasierten Analyse von Surveydaten wiederholt. Die Zielpopulation wurde durch eine übergeordnete hypothetische Superpopulation generiert. Die realisierten Merkmale dieser Population werden als Zufallsvariablen aufgefasst, wohingegen die gezogene Stichprobe als fest angenommen wird. Wenn die Modellannahmen für das Modell der Superpopulation zutreffen, wird angenommen dass das Stichprobendesign, bedingt auf die in der Modellgleichung integrierten Kovariablen, keinen Einfluss auf die Zielvariable hat. Ziel im modellbasierten Ansatz ist es den Modellparameter β unverzerrt und effizient zu schätzen. In diesem Ziel liegt ein wesentlicher Unterschied zur designbasierten Analyse von Surveydaten. Ein weiterer wichtiger Unterschied liegt in der Betrachtungsweise der Grundgesamtheit.

3.2 Designbasierte Analyse von Surveydaten

Im designbasierten Ansatz wird die gezogene Stichprobe als eine mögliche Zufallsstichprobe aus einer endlichen Grundgesamtheit betrachtet. Die Grundgesamtheit wird hierbei als fest angenommen. In dieser Betrachtungsweise liegt der wesentliche Unterschied zur modellbasierten Analyse von Surveydaten. Die Quelle der Zufälligkeit liegt hier nicht in der Realisierung der Merkmale y_i , sondern im zufälligen Ziehungsmechanismus der Stichprobe. Der Inklusionsindikator I_i einer Einheit i wird als Zufallsvariable betrachtet und nicht das Merkmal selbst. Dementsprechend wird in rein designbasierten Verfahren kein Bezug zu einer theoretischen, übergeordneten Superpopulation hergestellt (Binder und Roberts, 2004, 2009; Skinner, 1989).

Das Ziel rein designbasierter Verfahren ist es deskriptive Populationsgrößen der endlichen Grundgesamtheit unverzerrt zu schätzen. Eine deskriptive Populationsgröße ist der Totalwert eines Merkmals y_i in der endlichen Grundgesamtheit, auf dem im nächsten Unterabschnitt detailliert eingegangen wird.

Mit Hilfe von Designgewichten können Populationsgrößen unverzerrt aus der zufälligen Stichprobe geschätzt werden. Als Designgewichte dienen die inversen Inklusionswahrscheinlichkeiten der gezogenen Einheiten aus der Stichprobe. Die Wahrscheinlichkeit einer Einheit i in die Stichprobe zu gelangen ist der Erwartungswert des binären Inklusionsindikators I_i : $E(I_i) = P(I_i = 1) = \pi_i$. Die Kovarianz Δ_{ik} des Inklusionsindikators ist $\pi_{i,k} - \pi_i\pi_k$, wobei $\Delta_{ik} = \pi_i(1 - \pi_i)$ für $i = k$ ist. Der Term $\pi_{i,k} = P(I_i = 1 \text{ und } I_k = 1)$ bezeichnet die Ziehungswahrscheinlichkeit zweiter Ordnung. Das ist die Wahrscheinlichkeit der Einheiten i und k gemeinsam in die Stichprobe s zu gelangen. Für $i = k$ ist $\pi_{i,k} = \pi_i$.

Das Designgewicht einer gezogenen Einheit i wird mit $w_i = 1/\pi_i$ gekennzeichnet. Es wird im

Folgendes angenommen, dass die Inklusionswahrscheinlichkeit proportional zu einer Größenvariable z_i ist und die Einheiten ohne Zurücklegen gezogen werden: $\pi_i = n z_i / \sum_{i=1}^N z_i$. Dieses Stichprobendesign wird nachfolgend als *pps*-Stichprobe ohne Zurücklegen bezeichnet.⁶

3.2.1 Schätzung des Totalwertes

Die Schätzung des Totalwertes als Populationsgröße ist ein grundlegendes Prinzip im designbasierten Ansatz. Auf diesem Prinzip baut der Pseudo-Maximum-Likelihood-Ansatz auf. Dieser Ansatz erlaubt eine designbasierte Schätzung möglicher Zusammenhänge zwischen einem Zielmerkmal y_i und zusätzlichen erklärenden Variablen in einer endlichen Grundgesamtheit. Der Pseudo-Maximum-Likelihood-Ansatz wird im nächsten Unterabschnitt beschrieben. Ein designbasierter Schätzer für den Totalwert T einer Population ist der Horvitz-Thompson-Schätzer (nachfolgend HT-Schätzer genannt) (Horvitz und Thompson, 1952):

$$\hat{T} = \sum_{i=1}^N I_i w_i y_i = \sum_{i \in s} w_i y_i. \quad (3.13)$$

Dieser Schätzer liefert eine erwartungstreue Schätzung für den Totalwert in der Population:

$$E(\hat{T}) = \sum_{i=1}^N E(I_i) w_i y_i = \sum_{i=1}^N y_i = T. \quad (3.14)$$

Für die Konstruktion von Inferenzstatistiken kann die Verteilung des Schätzers für ausreichend große Stichproben durch eine Normalverteilung approximiert werden (Prášková und Sen, 2009). Unter Zuhilfenahme dieser approximativen Normalverteilungsannahme können Konfidenzintervalle der Form $\hat{T} \pm z_{1-\alpha/2} SD(\hat{T})$ konstruiert werden, wobei $SD(\hat{T})$ die Standardabweichung von \hat{T} ist und α das Signifikanzniveau bezeichnet. Für die Konstruktion eines Konfidenzintervalls wird eine erwartungstreue Schätzung der Varianz benötigt. Die Varianz von \hat{T} hat folgende Form:

$$V(\hat{T}) = \sum_{i=1}^N \sum_{k=1}^N \Delta_{ik} \frac{y_i y_k}{\pi_i \pi_k}, \quad (3.15)$$

Eine unverzerrte Schätzung der Varianz liefert der Ausdruck (Horvitz und Thompson, 1952):

$$\hat{V}(\hat{T}) = \sum_{i=1}^N \sum_{k=1}^N I_i I_k \hat{\Delta}_{ik} \frac{y_i y_k}{\pi_i \pi_k} = \sum_{i=1}^n \sum_{k=1}^n \hat{\Delta}_{ik} \frac{y_i y_k}{\pi_i \pi_k}, \quad (3.16)$$

⁶Die Abkürzung *pps* steht für die englische Bezeichnung „probability proportional to size“.

wobei $\hat{\Delta}_{ik} = \Delta_{ik}/\pi_{i,k}$ ist. Da $E(I_i I_k \hat{\Delta}_{ik}) = \pi_{i,k} \hat{\Delta}_{ik} = \Delta_{ik}$ gilt, ist $\hat{V}(\hat{T})$ erwartungstreu (Särndal et al., 2003 S. 42 ff.).

Diese Varianzschätzung hat bezüglich ihrer praktischen Umsetzung einen entscheidenden Nachteil. Zur Berechnung werden die gemeinsamen Ziehungswahrscheinlichkeiten aller paarweisen Beobachtungen in der Stichprobe benötigt. Diese sind in der Regel nicht bekannt und müssten approximiert werden. Außerdem macht die Doppelsumme die Berechnung der Varianz rechentechnisch sehr aufwendig (Särndal et al., 2003 S. 421). Die Varianzschätzung nach Horvitz und Thompson liefert zwar eine unverzerrte Schätzung für die Varianz von \hat{T} , aufgrund der beiden genannten Nachteile kommt sie in der Praxis jedoch selten zu Anwendung (Berger und Tillé, 2009).

Eine Approximation der Varianzschätzung für den HT-Schätzers beschreiben Särndal et al. (2003). Da diese Approximation in STATA und SAS implementiert ist, wird die Methode nun näher erläutert.

Mit Hilfe des Prinzips *abhängiger und zufälliger Gruppen* (Dependent Random Groups) leiten Särndal et al. (2003, S. 423 ff.) die vereinfachte Varianzschätzung her. Der Ansatz beruht darauf, dass die Stichprobe s in eine Anzahl A zufälliger disjunkter Gruppen, $s_1, \dots, s_a, \dots, s_A$ geteilt werden kann:

$$s = \bigcup_{a=1}^A s_a. \quad (3.17)$$

Die Partitionierung von s in die zufälligen Gruppen ist nicht willkürlich, vielmehr sollte jede Gruppe dem Stichprobendesign der ursprünglichen Stichprobe entsprechen. Für eine *pps*-Stichprobe mit Zurücklegen sollten die zufälligen Gruppen s_a so gewählt werden, dass sie einer einfachen Stichprobe mit Zurücklegen aus der Grundgesamtheit entsprechen. Das wird erreicht, indem s_1 ohne zurücklegen aus s gezogen wird, gefolgt von s_2 aus $s - s_1$ bis zur Gruppe s_A (Wolter, 2007 S. 22 ff.).

Mit den Beobachtungen aus einer zufälligen Gruppe s_a kann der Totalwert der T der Grundgesamtheit erwartungstreu geschätzt werden. Die Elemente aus der ursprünglichen Stichprobe s wurden mit einer Wahrscheinlichkeit proportional zu Ihrer Größe gezogen. Die Wahrscheinlichkeit für ein Element i aus der Grundgesamtheit in die zufällige Gruppe s_a zu gelangen ist $\pi_{i,a} = (m/n)\pi_i$ für $i = 1, \dots, N$, wobei m die Größe von Gruppe s_a ist. Vereinfachend wird für alle s_a die gleiche Größe angenommen. Für die Stichprobengröße n gilt dann, dass $n = Am$. Mit den Beobachtungen aus Gruppe s_a und den veränderten Inklusionswahrscheinlichkeiten

kann nun eine erwartungstreue Schätzung \hat{T}_a für den Totalwert T berechnet werden:

$$\hat{T}_a = \sum_{i \in s_a} y_i / \pi_{i,a} = (n/m) \sum_{i \in s_a} y_i / \pi_i. \quad (3.18)$$

Dieser hat einem Erwartungswert von:

$$E(\hat{T}_a) = E \left[\sum_{i \in s_a} y_i / \pi_{i,a} \right] = \sum_{i=1}^N E(I_{i,a}) y_i / \pi_{i,a} = T, \quad (3.19)$$

wobei $I_{i,a}$ der Inklusionsindikator ist und anzeigt ob Element i in der zufälligen Gruppe s_a ist. Der Erwartungswert $E(I_{i,a})$ entspricht demnach der Inklusionswahrscheinlichkeit $\pi_{i,a}$. Die Schätzung für den Totalwert der Grundgesamtheit aus allen A Gruppen ist der Mittelwert aus den Schätzungen der zufälligen Gruppen:

$$\hat{T}_{ZG} = 1/A \sum_{a=1}^A \hat{T}_a = \sum_{a=1}^A \sum_{i \in s_a} y_i / \pi_i = \sum_{i=1}^n y_i / \pi_i, \quad (3.20)$$

wobei der Index ZG auf die Schätzung aus den zufälligen Gruppen hinweist. Offensichtlich ist \hat{T}_{ZG} für konstantes m identisch mit der Schätzung für den Totalwert aus Gleichung 3.13. Dementsprechend ist in diesem Fall auch die Varianz von \hat{T} aus Gleichung 3.15 identisch mit der Varianz von \hat{T}_{ZG} .

Es stellt sich die Frage nach dem Praxisbezug dieses Ansatzes, denn mit den Gleichungen 3.13 und 3.20 werden Schätzer mit gleichen Eigenschaften erzielt. Was wird durch die Aufteilung der Stichprobe in zufällige Gruppen gewonnen? Im Gegensatz zu Gleichung 3.13 werden hier nicht mehr die I_i als Zufallsvariablen betrachtet, sondern die Schätzungen \hat{T}_a aus den zufälligen Gruppen s_a . Der entscheidende Vorteil ist, dass die Berechnung der paarweisen Inklusionswahrscheinlichkeiten aus Gleichung 3.16 hinfällig geworden ist. In STATA und SAS werden die zufälligen Gruppen so gewählt, dass A der Stichprobengröße n entspricht. In diesem Fall ist $m = 1$ und die s_a entsprechen den einzelnen Elementen aus der ursprünglichen Stichprobe. Falls die gezogenen Elemente Cluster der Grundgesamtheit sind, entsprechen die zufälligen Gruppen den Clustern der Stichprobe. Bei einem mehrstufigen Stichprobendesign werden in der Regel die Cluster der obersten Stufe als zufällige Gruppe betrachtet (STATA, 2009; SAS, 2004). In den nachfolgenden Ausführungen wird für A ebenfalls $A = n$ angenommen. Im Fall von $A = n$ vereinfacht sich die Ziehungswahrscheinlichkeit zu $\pi_{i,a} = \pi/n$.

Durch das Ziehen ohne Zurücklegen der zufälligen Gruppen sind die \hat{T}_a nicht unabhängig. Wird die Abhängigkeit vernachlässigt, kann die die Varianz von \hat{T} durch eine einfache Vari-

anz eines Mittelwertes approximiert werden:

$$\hat{V}_{ZG}(\hat{T}) = \frac{1}{n(n-1)} \sum_{a=1}^n (\hat{T}_a - \hat{T})^2 = \frac{n}{n-1} \sum_{a=1}^n (y_{\cdot a} - \bar{y}_{\cdot})^2, \quad (3.21)$$

wobei $y_{\cdot a} = \sum_{i \in s_a} w_i y_i$ der gewichteten Summe der Elemente i aus der zufälligen Gruppe s_a und $\bar{y}_{\cdot} = 1/n \sum_{i=1}^n y_{\cdot a}$ dem Mittelwert über alle $y_{\cdot a}$ entspricht.

Der eigentliche Ziehungsmechanismus wird, bei dieser Varianzschätzung vernachlässigt. Die Ziehungen werden behandelt als wären sie durch Ziehen mit Zurücklegen gezogen worden. Dieses Vorgehen führt in der Regel zu einer Verzerrung der Varianzschätzung. Die Vereinfachung geht demnach mit einer Verzerrung der Varianzschätzung einher (Särnadal, 2003 S. 428).

Eine wichtige Frage lautet, ob diese Verzerrung zu ungültigen Inferenzstatistiken führt und ob sie eventuell quantifiziert und korrigiert werden kann. Um die Verzerrung beurteilen zu können wird der Erwartungswert für die vereinfachte Varianzschätzung aus 3.21 betrachtet:

$$\begin{aligned} E(\hat{V}_{ZG}(\hat{T})) &= V(\hat{T}) - 2 \frac{1}{n(n-1)} \sum_{a=1}^n \sum_{b>a}^n Cov(\hat{T}_a, \hat{T}_b) \\ &\quad + \frac{1}{n(n-1)} \sum_{a=1}^n (E(\hat{T}_a) - E(\hat{T}))^2. \end{aligned} \quad (3.22)$$

Aus dieser Gleichung folgt, falls die $\hat{T}_1, \dots, \hat{T}_n$ unabhängig sind und den gleichen Erwartungswert haben, $\hat{V}_{ZG}(\hat{T})$ aus Gleichung 3.21 unverzerrt ist für $V(\hat{T})$. Im Allgemeinen kann nicht von Unabhängigkeit ausgegangen werden. Für große Stichproben mit einem kleinen Auswahlsatz ist die Summe der Kovarianzen aus Gleichung 3.22 jedoch erwartungsgemäß sehr klein und negativ. Auch die Varianz zwischen den Schätzungen aus den zufälligen Gruppen wird für große Stichproben eher als vernachlässigbar eingeschätzt. Für Surveydaten mit einer großen Anzahl an Beobachtungen und kleinem Auswahlsatz ist die Verzerrung in der Regel eher unerheblich und tendiert dazu die Varianz leicht zu überschätzen (Wolter, 2007 S. 33 f.).

Wolter (2007, S. 43) zeigt für die Approximierung der Varianz einer *pps*-Ziehung ohne Zurücklegen (*oz*) durch die Varianz einer *pps*-Ziehung mit Zurücklegen (*mz*), dass die Verzerrung allgemein durch den Ausdruck

$$Bias(\hat{V}_{mz}(\hat{T})) = E(\hat{V}_{mz}(\hat{T})) - V_{oz}(\hat{T}) = \frac{n}{n-1} [V_{mz}(\hat{T}) - V_{oz}(\hat{T})]. \quad (3.23)$$

beschrieben werden kann. Falls $V_{mz}(\hat{T})$ größer ist als $V_{oz}(\hat{T})$ tendiert die Approximation der Varianz durch $\hat{V}_{mz}(\hat{T})$ dazu $V_{oz}(\hat{T})$ zu überschätzen. Dementsprechend ist für Populationen,

in denen ein Stichprobendesign ohne Zurücklegen effizient ist, die durch die Vereinfachung induzierte Verzerrung positiv. Konfidenzintervalle werden folglich konservativ geschätzt, so dass das nominale Konfidenzniveau eher überschritten wird. Inferenzstatistiken, basierend auf der vereinfachten Varianzschätzung behalten ihre Gültigkeit (Wolter, 2007 S. 45). Die Verzerrung kann im Allgemeinen für Stichproben mit einem *pps*-Design nicht exakt quantifiziert werden.

Für den Fall einer einfachen Stichprobe ohne Zurücklegen ist es hingegen möglich die Verzerrung aus Gleichung 3.23 genau zu berechnen. Eine einfache Stichprobe mit Zurücklegen hat per Definition gleiche Ziehungswahrscheinlichkeiten für die Elemente der Grundgesamtheit und bildet einen Spezialfall der *pps*-Ziehung. Die Ziehungswahrscheinlichkeiten für die Elemente i der Grundgesamtheit entsprechen dem Auswahlatz ($\pi_i = n/N$) der Stichprobe s . Den Elementen aus den zufälligen Gruppen s_a werden analog zum *pps*-Design neue Ziehungswahrscheinlichkeiten zugeordnet: $\pi_{i,a} = \pi/n = 1/N$. Die Verzerrung der Schätzung durch $\hat{V}_{mz}(\hat{T})$ vereinfacht sich zu folgendem Ausdruck:

$$E(\hat{V}_{mz}(\hat{T})) - V_{oz}(\hat{T}) = (1-f)V_{mz}(\hat{T}), \quad (3.24)$$

wobei $f = n/N$ der Auswahlatz der Stichprobe ist (Särndal, 2003 S. 427). Durch die Multiplikation mit der Endlichkeitskorrektur - dem Faktor $(1-f)$ - kann eine unverzerrte Schätzung der Varianz berechnet werden. Die Schätzung der Varianz aus Gleichung 3.22 erfolgt durch:

$$\hat{V}_{zG}(\hat{T}) = \frac{(1-f)}{n(n-1)} \sum_{i=1}^n (\hat{T}_a - \hat{T})^2. \quad (3.25)$$

Auch hier wird offensichtlich, dass die Verzerrung für sehr kleine Auswahlätze vernachlässigbar ist (Särndal, 2003 S. 427).

Zusammenfassend werden die wichtigsten Erkenntnisse über die designbasierte Totalwert-schätzung wiederholt. Der Totalwert einer endlichen Grundgesamtheit kann durch den HT-Schätzer aus Gleichung 3.13 unverzerrt geschätzt werden. Gleichung 3.16 nach Horvitz und Thompson liefert eine unverzerrte Schätzung für die Varianz des Totalwerts. Sie ist jedoch wenig praktikabel, da sie die Kenntnis der unbekanntenen gemeinsamen Ziehungswahrscheinlichkeiten voraussetzt. Mit der Bildung abhängiger zufälliger Gruppen aus der Stichprobe s wird ein alternativer Schätzer für den Totalwert gebildet (Geichung 3.18). Es wurde gezeigt, dass dieser Schätzer äquivalent zum HT-Schätzer ist. Die Varianzschätzung auf Basis der zufälligen Gruppen kommt ohne die Ziehungswahrscheinlichkeiten der zweiten Stufe aus. Da diese Varianzschätzung den Ziehungsmechanismus ohne Zurücklegen ignoriert, tendiert sie

dazu die Varianz zu überschätzen. Für große Stichproben mit kleinem Auswahlsatz ist die Verzerrung allerdings vernachlässigbar. Dieses Prinzip der Varianzschätzung wird in SAS und STATA für designbasierte Auswertungen von Surveydaten angewendet.

3.2.2 Pseudo-Maximum-Likelihood-Ansatz

Neben dem Totalwert ist von Interesse, ob das Merkmal y_i eventuell durch die Ausprägungen anderer Variablen in der Grundgesamtheit erklärt werden kann. So sollen zum Beispiel aus den gewonnenen Daten des BGS mögliche erklärende Faktoren für Gesundheitsmerkmale gefunden werden. Eine Möglichkeit hierfür bieten modellbasierte Verfahren. Falls dieziehungswahrscheinlichkeiten mit dem Merkmal y_i korreliert sind, eignen sich modellbasierte Verfahren nicht. Die Schätzung der Parameter könnte dann verzerrt sein (Pfeffermann und Sverchkow, 2009).

Ein statistisches Modell, das den Zusammenhang zwischen einer abhängigen Variable und zusätzlichen Kovariablen beschreibt scheint auf den ersten Blick mit dem designbasierten Ansatz nicht zusammenzupassen. Mit dem Pseudo-Maximum-Likelihood-Ansatz (PLM-Ansatz) kann eine Brücke zur designbasierten Schätzung von Modellparametern in einer endlichen Grundgesamtheit geschlagen werden (Skinner, 1998 S. 80 ff.).

Zur Erinnerung, im rein designbasierten Ansatz wird kein Bezug zu einem hypothetischen, übergeordneten Modell einer Superpopulation hergestellt. Das Modell der Superpopulation wird jedoch als sogenanntes *Arbeitsmodell* herangezogen, um die Scorefunktion der endlichen Grundgesamtheit zu definieren (Skinner, 1989, S. 83). Als Beispiel dient erneut das lineare Modell aus Gleichung 3.1:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i.$$

In der Definition des Zielparameters liegt der entscheidende Unterschied zur modellbasierten Analyse. Der Zielparameter der Schätzung in der designbasierten Analyse ist β_N der endlichen Population. Zum Vergleich: im modellbasierten Ansatz soll der Parameter β aus dem hypothetischen Modell der Superpopulation geschätzt werden. Die Inferenz im modellbasierten Ansatz lässt Schlüsse über Zusammenhänge zu, die nicht nur auf die endliche Zielpopulation der gezogenen Stichprobe zutreffen. Die designbasierte statistische Inferenz lässt Schlüsse über Zusammenhänge innerhalb der definierten endlichen Grundgesamtheit zu (Binder und Roberts, 2009; Heeringa et al., 2010 S. 184 f.).

Aus Abschnitt 3.1 ist bekannt, dass die Scorefunktion der Populationslikelihood, ausgewertet am Populationsparameter β_N gleich Null ist. Aus der Gleichung 3.5 für die Populations-

funktion:

$$\mathbf{U}_N(\beta_N) = \sum_{i \in U} \mathbf{x}_i (y_i - \mathbf{x}_i' \beta_N) = \sum_{i \in U} \mathbf{u}_i(\beta_N) = \mathbf{0}$$

wird ersichtlich, dass $\mathbf{U}_N(\beta_N)$ auch als ein Totalwert der Scorebeiträge aller Elemente der endlichen Grundgesamtheit aufgefasst werden kann. Im PLM-Ansatz ist $\mathbf{U}_N(\beta_N)$ eine deskriptive Populationsgröße. Sie kann mit Hilfe designbasierter Verfahren erwartungstreu geschätzt werden. Auch wenn das angenommene Arbeitsmodell den Zusammenhang zwischen y_i und \mathbf{x}_i nicht korrekt beschreibt, kann $\mathbf{U}_N(\beta_N)$ unverzerrt geschätzt werden. Die Formulierung eines Arbeitsmodells dient folglich dem Zweck die funktionale Form der Scorebeiträge für den Totalwert $\mathbf{U}_N(\beta_N)$ zu definieren. Die Form des Totalwerts aus diesem likelihoodbasierten Ansatz wird auch als Schätzgleichung bezeichnet (Binder und Patac, 1994; Godambe und Thompson, 2009).

Der Totalwert kann mit dem im Unterabschnitt 3.2.1 beschriebenen Prinzip nach Horvitz und Thompson geschätzt werden. Nachfolgend soll der Index p auf den Randomisierungsprozess der zufälligen Stichprobenziehung hinweisen.

Der designbasierte Schätzer $\hat{\beta}_p$ für β_N ist die Lösung der Schätzgleichung (Skinner, 1989 S. 82; Binder und Roberts, 2009 S. 43):

$$\hat{\mathbf{U}}_p(\hat{\beta}_p) = \sum_{i \in U} I_i w_i \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\beta}_p) = \sum_{i \in U} I_i w_i \mathbf{u}_i(\hat{\beta}_p) = \mathbf{0}. \quad (3.26)$$

wobei $\hat{\beta}_p$ durch die Nullstelle von $\hat{\mathbf{U}}_p(\hat{\beta}_p)$ definiert ist. Im einfachen linearen Modell entspricht der PML-Schätzer $\hat{\beta}_p$ dem gewichteten KQ-Schätzer einer linearen Regression:

$$\hat{\beta}_p = \left(\sum_{i \in s} w_i \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in s} w_i \mathbf{x}_i' y_i. \quad (3.27)$$

Aus Unterabschnitt 3.2.1 ist bekannt, dass der HT-Schätzer erwartungstreu für den Totalwert einer endlichen Population ist. Gilt dies jedoch auch für die designbasierte Schätzung $\hat{\beta}_p$? Der geschätzte Populationsparameter $\hat{\mathbf{U}}_p(\hat{\beta}_p)$ ist eine Funktion vom geschätzten Parameter $\hat{\beta}_p$. Zur Untersuchung der Eigenschaften von $\hat{\beta}_p$ wird das Prinzip der Taylor-Linearisierung verwendet. Die Taylor-Linearisierung von $\hat{\mathbf{U}}_p(\hat{\beta}_p)$ an der Stelle $\hat{\beta}_p = \beta_N$ hat die Form:

$$\hat{\mathbf{U}}_p(\hat{\beta}_p) \approx \hat{\mathbf{U}}_p(\beta_N) + (\hat{\beta}_p - \beta_N) \hat{\mathbf{S}}_{xx}, \quad (3.28)$$

wobei der Term $\hat{\mathbf{S}}_{xx}$ die Matrix der partiellen Ableitung der Schätzfunktion $\hat{\mathbf{U}}_p(\hat{\boldsymbol{\beta}}_p)$ bezüglich des Parametervektors $\boldsymbol{\beta}_N$ ist. Im linearen Modell gilt für $\hat{\mathbf{S}}_{xx}$:

$$\hat{\mathbf{S}}_{xx} = \sum_{i \in U} I_i w_i \mathbf{x}_i \mathbf{x}_i' \quad (3.29)$$

Für große N wird angenommen, dass $E(\hat{\mathbf{S}}_{xx}) \approx \mathbf{S}_{xx} = \sum_{i \in U} \mathbf{x}_i \mathbf{x}_i'$ gilt. Die Matrix $\hat{\mathbf{S}}_{xx}$ konvergiert für große N in Wahrscheinlichkeit gegen die Konstante Matrix \mathbf{S}_{xx} (Binder und Roberts, 2009 S. 44). Daraus ergibt sich

$$\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_N \rightarrow \mathbf{S}_{xx}^{-1} \hat{\mathbf{U}}_p(\boldsymbol{\beta}_N). \quad (3.30)$$

Für eine asymptotische Erwartungstreue des geschätzten Parameters $\hat{\boldsymbol{\beta}}_p$ muss gelten, dass $\hat{\mathbf{U}}_p(\boldsymbol{\beta}_N)$ einen Erwartungswert von Null hat:

$$E_p(\hat{\mathbf{U}}_p(\boldsymbol{\beta}_N)) = \sum_{i \in U} E_p(I_i) w_i \mathbf{u}_i(\boldsymbol{\beta}_N) = \sum_{i \in U} \mathbf{u}_i(\boldsymbol{\beta}_N) = \mathbf{0}. \quad (3.31)$$

Der Erwartungswert von $\hat{\mathbf{U}}_p(\boldsymbol{\beta}_N)$ entspricht der Schätzfunktion $\mathbf{U}_N(\boldsymbol{\beta}_N)$ aus Gleichung 3.5, die per Definition einen Wert von Null annimmt. Die designbasierte Schätzung $\hat{\boldsymbol{\beta}}_p$ ist daher asymptotisch unverzerrt für den Parametervektor $\boldsymbol{\beta}_N$ der endlichen Population (Binder und Roberts, 2009 S. 44).

Aus Gleichung 3.30 lässt sich die asymptotische Varianz für $\hat{\boldsymbol{\beta}}_p$ herleiten. Da \mathbf{S}_{xx}^{-1} eine konstante Matrix ist, geht dieser Term in quadratischer Form in die Berechnung der Varianz ein. Für die Varianz von $\hat{\boldsymbol{\beta}}_p$ gilt dann

$$\mathbf{V}_p(\hat{\boldsymbol{\beta}}_p) = \mathbf{S}_{xx}^{-1} \mathbf{V}_p(\hat{\mathbf{U}}_p(\boldsymbol{\beta}_N)) \mathbf{S}_{xx}^{-1}. \quad (3.32)$$

Diese quadratische Form wird allgemein als Sandwichform bezeichnet. Da die Matrix \mathbf{S}_{xx}^{-1} und der Parametervektor $\boldsymbol{\beta}_N$ nicht bekannt sind, werden für die Schätzung der Varianz ihre Stichprobenpendants $\hat{\mathbf{S}}_{xx}^{-1}$ und $\hat{\boldsymbol{\beta}}_p$ verwendet:

$$\hat{\mathbf{V}}_p(\hat{\boldsymbol{\beta}}_p) = \hat{\mathbf{S}}_{xx}^{-1} \hat{\mathbf{V}}_p(\hat{\mathbf{U}}_p(\hat{\boldsymbol{\beta}}_p)) \hat{\mathbf{S}}_{xx}^{-1}. \quad (3.33)$$

Der Mittelteil $\hat{\mathbf{V}}_p(\hat{\mathbf{U}}_p(\hat{\boldsymbol{\beta}}_p))$ muss durch designbasierter Verfahren geschätzt werden. Da $\mathbf{U}_p(\hat{\boldsymbol{\beta}}_p)$ im PLM-Ansatz als Totalwert einer endlichen Grundgesamtheit verstanden wird, kann die Varianz mit der in Unterabschnitt 3.2.1 vorgestellten Approximation geschätzt werden. Gemäß

Gleichung 3.21 wird die Varianz wie folgt geschätzt:

$$\hat{\mathbf{V}}_p \left(\hat{\mathbf{U}}_p(\hat{\boldsymbol{\beta}}_p) \right) = \frac{n}{(n-1)} \sum_{i=1}^n (\mathbf{u}_{\cdot i} - \bar{\mathbf{u}}_{\cdot}) (\mathbf{u}_{\cdot i} - \bar{\mathbf{u}}_{\cdot})'. \quad (3.34)$$

Angenommen die Einheiten i sind Cluster aus denen jeweils eine Anzahl n_i Beobachtungseinheiten j mit $j = 1, \dots, n_i$ gezogen wurden. Dann entspricht $\mathbf{u}_{\cdot i} = \sum_{j=1}^{n_i} w_j \mathbf{u}_{ij}$ der gewichteten Summe über die Scorebeiträge der Elemente j aus Cluster i , und $\bar{\mathbf{u}}_{\cdot} = 1/n \sum_{i=1}^n \mathbf{u}_{\cdot i}$ dem Mittelwert über alle $\mathbf{u}_{\cdot i}$. Wurden die Cluster durch einfaches Ziehen ohne Zurücklegen gezogen, kann die geschätzte Varianz gemäß Gleichung 3.25 zusätzlich mit einer Endlichkeitskorrektur $(1-f)$ multipliziert werden, um eine unverzerrte Schätzung zu erhalten.

Diese Form der linearisierten Varianzschätzung wird standardmäßig in STATA und SAS verwendet (STATA, 2009; SAS, 2004). In beiden Programmen stehen auch Resamplingmethoden zur Varianzschätzung zur Verfügung. Die Idee dieser Methoden besteht darin, aus der Stichprobe s nach einem definierten Prinzip wiederholt neue Stichproben zu ziehen und jeweils den Parameter der Grundgesamtheit zu schätzen. Anschließend wird die Varianz aus der Verteilung dieser Schätzungen berechnet. Resamplingmethoden gehen über den Rahmen dieser Arbeit hinaus. Eine Beschreibung bieten zum Beispiel Rust und Rao (1996).

Der beschriebene Ansatz der likelihoodbasierten Schätzgleichung kann auf beliebig komplexe Regressionsmodelle angewendet werden. Je nach Beschaffenheit der zu modellierenden Daten sind zum Beispiel nichtlineare oder gemischte Modelle für die Formulierung des Arbeitsmodells denkbar. Die im nächsten Kapitel vorgestellten Methoden zur Schätzung des Zeiteffekts im BGS nutzen das Prinzip der Schätzgleichungen. Da im Fokus dieser Arbeit auch die Praktikabilität der vorgestellten Verfahren steht, wird für jede Methode deren Anwendung in STATA und SAS diskutiert.

4 Methoden zur Schätzung des Zeiteffekts

Für die im BGS erhobenen Gesundheitsmerkmale wird der zeitliche Effekt zwischen den beiden Erhebungswellen, unter Berücksichtigung der Kontrollvariablen Alter und Geschlecht geschätzt. Im zusammengeführten Datensatz existieren zwei Quellen der Korrelation: die Intraclusterkorrelation innerhalb der Sample Points und die zeitliche Korrelation innerhalb der Längsschnittprobanden.

Welche Methode zur Schätzung des Zeiteffekts ist geeignet? Geeignete Methoden zur Auswertung dieser Daten unverzerrte Punktschätzung für die Parameter garantieren. Ist dieses Kriterium erfüllt, folgt als nächstes eine Betrachtung der Varianzschätzung. Hierbei sollten die auf den geschätzten Varianzen basierenden Inferenzstatistiken zu gültigen Aussagen führen. Werden die Varianzen unterschätzt, können statistische Test fälschlicherweise zu signifikanten Ergebnissen führen. Werden sie hingegen überschätzt können signifikante Unterschiede unentdeckt bleiben. Eine möglichst effiziente Varianzschätzung ist demnach das Ziel.

In den nächsten drei Abschnitten werden das Generalisierte Lineare Modell (GLM), die Methode der Generalisierten Schätzgleichung (Generalized Estimating Equation - GEE) und das Generalisierte lineare gemischte Modell (Generalized Linear Mixed Model - GLMM) vorgestellt. Es werden jeweils das Arbeitsmodell und die sich daraus ergebende Schätzgleichung für den designbasierten PML-Ansatz erläutert. Außerdem wird jeweils die Implementierung in SAS und STATA vorgestellt.

Die Notation des vorherigen Abschnitts wird wieder aufgegriffen. In dem zweistufigen Design werden n Cluster (Sample Points) gezogen. Das Cluster i mit $i = 1, \dots, n$ besteht aus n_i Beobachtungseinheiten (Probanden). Die Einheit j mit $j = 1, \dots, n_i$ wird an $t = (1, 2)$ Zeitpunkten beobachtet. Für eine übersichtlichere Notation wird in den folgenden Ausführungen von einem balancierten Datensatz ausgegangen, in dem alle Beobachtungseinheiten zu beiden Zeitpunkten vorhanden sind.⁷

⁷Für den unbalancierten Fall müsste jeder Einheit j ein spezifischer Zeitindex t_j zugeordnet werden mit $t_j = (1)$ für Probanden die nur im BGS98 enthalten sind, $t_j = (2)$ für Probanden die nur im BGS98 enthalten sind und $t_j = (1, 2)$ für die Längsschnittprobanden.

4.1 Generalisiertes lineares Modell

Mit dem GLM-Ansatz können sowohl lineare, als auch nichtlineare Modelle geschätzt werden. Für eine stetige, abhängige Variable entspricht dieser Ansatz einem linearen Regressionsmodell und für binäre abhängige Merkmale der logistischen Regression.

4.1.1 Arbeitsmodell und Schätzgleichung

Das Arbeitsmodell für den Erwartungswert einer abhängigen Variable y_{ijt} kann im GLM-Ansatz wie folgt ausgedrückt werden:

$$\mu_{ijt} = E(y_{ijt}) = g(\mathbf{x}'_{ijt}\boldsymbol{\beta}) \quad (4.1)$$

Die Responsefunktion $g(\cdot)$ entspricht in der linearen Regression der Identität $\mathbf{x}'_{ijt}\boldsymbol{\beta}$ und in der logistischen Regression dem Ausdruck (Fahrmeir et al., 2007 S. 198 f.)

$$\mu_{ijt} = P(y_{ijt} = 1 | \mathbf{x}_{ijt}) = \frac{\exp(\mathbf{x}'_{ijt}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_{ijt}\boldsymbol{\beta})}. \quad (4.2)$$

Für das lineare Modell entspricht der GLM-Ansatz dem herangezogenen Beispiel aus Unterabschnitt 3.2.2. Für die logistische Regression hat die Dichtefunktion einer Beobachtung der binären Variable y_{ijt} die folgende Form (Skinner, 1998 S. 83):

$$f(y_{ijt}; \boldsymbol{\beta}) = \mu_{ijt}(\boldsymbol{\beta})^{y_{ijt}} (1 - \mu_{ijt}(\boldsymbol{\beta}))^{1-y_{ijt}}. \quad (4.3)$$

Daraus folgt für den Log-Likelihoodbeitrag der Beobachtung ijt :

$$l(y_{ijt}; \boldsymbol{\beta}) = y_{ijt} \log(\mu_{ijt}) - y_{ijt} \log(1 - \mu_{ijt}) + \log(1 - \mu_{ijt}). \quad (4.4)$$

Aus der Differenzierung bezüglich des Parametervektors $\boldsymbol{\beta}$ ergibt sich der Scorebeitrag der Beobachtung ijt für die logistische Regression:

$$\mathbf{u}_{ijt}(\boldsymbol{\beta}) = (y_{ijt} - \mu_{ijt}(\boldsymbol{\beta})) \mathbf{x}_{ijt} \quad (4.5)$$

Ein Vergleich zu den Likelihoodbeiträgen der linearen Regression aus Gleichung 3.5 zeigt, dass ihre Struktur identisch ist. Der Unterschied liegt lediglich in der Definition des Erwartungswertes μ_{ijt} . Somit ist auch die Form der Schätzgleichung identisch. Im designbasierten Ansatz wird der Populationsparameter $\boldsymbol{\beta}_N$ des GLM gemäß dem vorgestellten PLM-Ansatz

mit der folgenden Schätzgleichung geschätzt:

$$\hat{\mathbf{U}}_p(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^n \sum_{j=1}^{n_j} \sum_{t=1}^2 w_{ijt} \mathbf{x}_{ijt} \left(y_{ijt} - \mu_{ijt}(\hat{\boldsymbol{\beta}}_p) \right) = \mathbf{0}. \quad (4.6)$$

Im Fall der linearen Regression besitzt dieses Gleichungssystem eine geschlossene Lösung. Diese entspricht dem Parametervektor $\hat{\boldsymbol{\beta}}_p$ aus Gleichung 3.27. Die Parameter der logistischen Regression können näherungsweise durch iterativer Verfahren geschätzt werden. Eine Übersicht zu solchen Näherungsverfahren und ihre Implementierung in Statistik-Software liefern zum Beispiel Gould, Ptiblado und Sribney (2006, Kapitel 1).

Aus obiger Schätzgleichung wird ersichtlich, dass im GLM-Ansatz die beschriebene Abhängigkeitsstruktur in den Daten außer Acht gelassen wird. Die Beobachtungen werden über alle Cluster und Zeitpunkte gepoolt und somit als unabhängig behandelt. Dieses Vorgehen liefert dennoch erwartungstreue Schätzungen für der Populationsparameter $\boldsymbol{\beta}_N$. Die Schätzung der Varianz erfolgt mit dem Sandwichschätzer aus Gleichung 3.33 (Skinner, 1989 S. 81 f.).

4.1.2 Implementierung in SAS und STATA

In SAS und STATA ist der designbasierte PLM-Ansatz für die lineare und logistische Regression implementiert. In SAS können mit den Prozeduren:

- PROC SURVEYREGRESSION
- PROC SURVEYLOGISTIC

die Parameter der beiden Modellklassen geschätzt werden. Das CLUSTER-Statements erlaubt die Cluster der ersten Ziehungsstufe anzugeben. Eine weitere Clusterung innerhalb der Elemente der ersten Stufe kann nicht berücksichtigt werden. Über das WEIGHT-Statement können zusätzlich die Designgewichte angegeben werden. Die Schätzung der Varianz wird über die Gleichung 3.33 berechnet (SAS, 2004).

In STATA wird das Stichprobendesign über den Befehl svyset definiert. Mit Hilfe der Befehle:

- svy:regress
- svy:logit

können die Parameter der beiden Modellklassen geschätzt werden. Wird im svyset-Befehl nur die Clusterung der ersten Stufe angegeben, sind die Prozeduren von SAS und STATA äquivalent.

In STATA ist es im Gegensatz zu SAS bei einem mehrstufigen Design möglich, auch eine Clusterung innerhalb der Elemente der ersten Stufe zu berücksichtigen. Eine Berücksichtigung mehrerer Ziehungsstufen hat keine Auswirkung auf die Punktschätzung der Parameter. In der Varianzschätzung wird jedoch nicht nur die Variation zwischen den Clustern der ersten Stufe berücksichtigt, sondern zusätzlich die Variabilität zwischen den Clustern der zweiten Stufe.

Der Mittelteil $\hat{\mathbf{V}}_p(\hat{\mathbf{U}}_p(\hat{\beta}_p))$ aus Gleichung 3.33 wird um einen zusätzlichen Term erweitert. Dieser Term beschreibt die Varianz zwischen den Clustern der zweiten Stufe. In einer zweistufigen Clusterstichprobe kann die Varianz wie folgt geschätzt werden (Lehtonen und Pahkinen, 2004 S. 79 f.):

$$\begin{aligned} \hat{\mathbf{V}}_p(\hat{\mathbf{U}}_p(\hat{\beta}_p)) = & (1-f) \frac{n}{(n-1)} \sum_{i=1}^n (\mathbf{u}_{\cdot i} - \bar{\mathbf{u}}_{\cdot\cdot}) (\mathbf{u}_{\cdot i} - \bar{\mathbf{u}}_{\cdot\cdot})' \\ & + f(1-f_i) \frac{n_i}{(n_i-1)} \sum_{j=1}^{n_i} (\mathbf{u}_{\cdot ij} - \bar{\mathbf{u}}_{\cdot i}) (\mathbf{u}_{\cdot ij} - \bar{\mathbf{u}}_{\cdot i})' \end{aligned} \quad (4.7)$$

Der erste Teil dieser Gleichung entspricht, abgesehen von der Endlichkeitskorrektur $(1-f)$, der Varianzschätzung aus Gleichung 3.33. Im zusätzlichen Teil entspricht $\mathbf{u}_{\cdot ij} = \sum_{t=1}^2 w_{ijt} \mathbf{u}_{ijt}$ der gewichteten Summe über die Scorebeiträge von Elemente j aus Cluster i über die beiden Zeitpunkte. Der Term $\bar{\mathbf{u}}_{\cdot i} = 1/n_i \sum_{i=1}^{n_i} \mathbf{u}_{\cdot ij}$ entspricht dem Mittelwert über alle $\mathbf{u}_{\cdot ij}$ in Cluster i . Der Term f_i bezeichnet den Auswahlsatz der zweiten Stufe aus Cluster i . Für $f = 0$ nimmt die Varianzkomponente der zweiten Stufe ebenfalls den Wert Null an. Aus diesem Grund werden die nachfolgenden Ziehungsstufen in STATA nur berücksichtigt, wenn eine Endlichkeitskorrektur in der darüber liegenden Ebene angegeben wird (STATA, 2009 S. 159).

Da im BGS die Samplepoints mit einer Wahrscheinlichkeit proportional zur Größe und nicht durch einfaches Ziehen ohne Zurücklegen gezogen wurden, ist eine Endlichkeitskorrektur in diesem Fall nicht angebracht. In den Abschnitten 3.2.1 und 3.2.2 wurde gezeigt, dass eine Endlichkeitskorrektur nur bei einer einfachen Ziehung ohne Zurücklegen zu einer unverzerrten Varianzschätzung führt.

Da in dieser Arbeit nicht nur eine theoretisch korrekte Methode im Vordergrund steht, sondern auch eine praktikable Lösung für den Anwender gefunden werden soll, werden in der nachfolgenden Simulationsstudie beide Verfahren untersucht. Hier wird sich zeigen, ob die Varianzschätzung aus Gleichung 4.7 trotz der Endlichkeitskorrektur, eine akzeptable Varianzschätzung für das vorliegenden Stichprobendesign hervorbringt.

Das in SAS und STATA implementierte Verfahren, in dem bei der Varianzschätzung nur die Clusterung der obersten Stufe berücksichtigt wird, erhält in der Simulation die Bezeichnung

$WGLM_1$. Die in STATA implementierte Methode mit Berücksichtigung der zweiten Ziehungsstufe, wird in der Simulation mit $WGLM_2$ bezeichnet.

Es werden ebenfalls modellbasierte - ungewichtete - Schätzungen in der Simulation untersucht. Um gültige Inferenzstatistiken zu erhalten, sollten die Varianzen bei einer vorliegenden Abhängigkeit innerhalb der Cluster, robust geschätzt werden. In Stata können mit der Option `vce(cluster clustername)` robuste Standardfehler geschätzt werden. SAS bietet innerhalb der Prozedur GENMOD eine robuste Varianzschätzung. In der Simulation wird die ungewichtete Schätzung der Parameter mit robusten Standardfehlern mit GLM_R bezeichnet.

4.2 Generalisierte Schätzgleichung

Im GEE-Ansatz werden schon während der Parameterschätzung angenommene Korrelationsstrukturen berücksichtigt. In die Schätzgleichung der PML-Schätzung eine so genannte Arbeitskorrelationsmatrix integriert. Diese generalisierte Form der Schätzgleichung geht zurück auf die Arbeit von Liang und Zeger (1986).

4.2.1 Arbeitsmodell und Schätzgleichung

Das Arbeitsmodell für den Erwartungswert einer abhängigen Variable y_{ijt} ist identisch mit dem des GLM-Ansatzes:

$$\mu_{ijt} = E(y_{ijt}) = g(\mathbf{x}'_{ijt}\boldsymbol{\beta}) \quad (4.8)$$

Der Unterschied liegt in der Definition der Schätzgleichung.

Allgemein kann für eine Verteilung aus der Exponentialfamilie gezeigt werden, dass der Scorebeitrag einer Beobachtung k folgende Form hat:

$$\mathbf{u}_{ijt}(\boldsymbol{\beta}) = \mathbf{x}_{ijt} \frac{d_{ijt}}{\sigma_k^2} (y_{ijt} - \mu_{ijt}), \quad (4.9)$$

wobei $d_{ijt} = \delta \mu_{ijt} / \delta \mathbf{x}_{ijt} \boldsymbol{\beta}$ die äußere Ableitung von μ_{ijt} ist und $\sigma_{ijt}^2 = \text{Var}(y_{ijt} | \mathbf{x}_{ijt})$ (Fahrmeir et al., 2007 S. 220 ff.). Wenn die Einheiten der zweiten Ziehungsstufe nur zu einem Zeitpunkt beobachtet werden, ist die Scorefunktion der Grundgesamtheit durch

$$\mathbf{U}_N(\boldsymbol{\beta}_N) = \sum_{i=1}^N \mathbf{X}'_i \mathbf{D}_i \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \quad (4.10)$$

definiert. $\mathbf{D}_i = \text{diag}(d_1, \dots, d_{n_i})$ ist eine Diagonalmatrix, mit $\delta\mu_{ij}/\delta(\mathbf{x}'_{ij}\boldsymbol{\beta})$ als Diagonalelementen. Die Matrix $\Sigma_i = \text{diag}(\sigma_1, \dots, \sigma_{n_i})$ enthält die Varianzen auf der Diagonalen (Molenberghs und Verbeke, 2005 S. 153 ff.).

Für die GEE-Schätzung wird Σ_i mit der Varianzmatrix $V_i = \Sigma_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\Sigma_i^{1/2}$ ersetzt, wobei $\mathbf{R}_i(\boldsymbol{\alpha})$ die erwähnte Arbeitskorrelationsmatrix mit unbekanntem Parametervektor $\boldsymbol{\alpha}$ ist. Ist $\mathbf{R}_i(\boldsymbol{\alpha})$ eine Einheitsmatrix, wird Unabhängigkeit zwischen den Individuen innerhalb eines Clusters angenommen. Die designbasierte Schätzgleichung entspricht dann dem beschriebenen GLM-Ansatz (Carlin et al. 1999).

Mit der Arbeitskorrelationsmatrix wird eine Vermutung über die Korrelationsstruktur in das Modell gegeben. Hierbei gibt es verschiedene Spezifikationsmöglichkeiten für $\mathbf{R}_i(\boldsymbol{\alpha})$. Für die Grundgesamtheit des BGS scheint eine konstante Intraclusterkorrelation plausibel. In diesem Fall muss für $\boldsymbol{\alpha}$ nur ein Wert geschätzt werden. Wird $\boldsymbol{\alpha}$ als konstant angenommen, wird von einer austauschbaren Arbeitskorrelationsmatrix gesprochen.⁸ Für eine austauschbare Korrelation kann der Parameter $\boldsymbol{\alpha}$ durch die Gleichung

$$\hat{\boldsymbol{\alpha}} = 1/n \sum_{i=1}^n \frac{1}{n_i(n_i-1)} \sum_{j \neq k} e_{ij}e_{ik} \quad (4.11)$$

geschätzt werden, wobei $e_{ij} = (y_{ij} - \mu_{ij})/\sqrt{v(\mu_{ij})}$. Der Term $v(\mu_{ij})$ bezeichnet die Varianz von μ_{ij} (Molenberghs und Verbeke, 2005 S. 158). Für die Schätzung des Parameters $\boldsymbol{\alpha}$ wird in einem ersten Schritt eine konsistente Schätzungen für $\boldsymbol{\beta}$ durchgeführt. Für diese erste Schätzung wird in der Regel für die Arbeitskorrelation die Einheitsmatrix eingesetzt. Dies entspricht dem beschriebenen GLM-Ansatz. Mit Hilfe der Anfangswerte $\hat{\boldsymbol{\beta}}^0$ kann $\boldsymbol{\alpha}$ geschätzt werden, um im nächsten Schritt den GEE-Schätzer $\hat{\boldsymbol{\beta}}$ zu schätzen. Diese Abfolge wird in einem iterativen Verfahren wiederholt, bis sich stabile Werte für Schätzungen $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\alpha}}$ ergeben (Liang und Zeger, 1986; Zorn, 2001).

Für eine designbasierte Schätzung der Populationsparameter müssen analog zum GLM-Ansatz die inversen Inklusionswahrscheinlichkeiten in die Scorefunktion der GEE-Schätzung integriert werden. Die Schätzgleichung hat dann folgende Form:

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^n \mathbf{X}'_i \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_p)) = \mathbf{0} \quad (4.12)$$

wobei $\mathbf{W}_i = \text{diag}(w_1, \dots, w_{n_i})$ eine Diagonalmatrix mit den Designgewichten der Individuen $j = 1, \dots, n_i$ aus Cluster i ist (Preisser et al., 2002). Die Schätzung der Varianz erfolgt wieder-

⁸Eine ausführliche Beschreibung weiterer Spezifikationen der Korrelationsstruktur ist zum Beispiel in Molenbergh und Verbeke (2005, S. 157 ff.) zu finden.

um mit der robusten Sandwichschätzung aus Gleichung 3.33. Für den geschätzten Totalwert der Grundgesamtheit $\hat{U}(\hat{\beta}_p)$ wird die obige designbasierte Schätzgleichung eingesetzt. Auch der Parameter α sollte designbasiert, das heißt mit Berücksichtigung der Designgewichte, geschätzt werden (Carilló et al. 2010). Wie sich im nächsten Unterabschnitt zeigen wird, ist dies in SAS nicht möglich.

In der Schätzgleichung 4.12 wird die Abhängigkeit innerhalb der Cluster berücksichtigt, wobei die zeitliche Korrelation unberücksichtigt bleibt. Eine in mehreren Ebenen verschachtelte Form der Arbeitskorrelationsmatrix ist zwar theoretisch vorstellbar, jedoch in SAS und STATA nicht implementiert. Bei der Auswertung des BGS kann demzufolge entweder die zeitliche Abhängigkeit oder die Intraclusterkorrelation berücksichtigt werden kann. Liang und Zeger (1986) zeigen, dass auch unter Fehlspezifikation der Arbeitskorrelation der GEE-Ansatz eine unverzerrte Schätzung liefert. Eine Berücksichtigung der tatsächlich zugrunde liegenden Abhängigkeitsstruktur kann jedoch zu effizienteren Schätzern führen (Carlin et al., 1999; Neuhaus, 1992; Zorn, 2001).

Für die Modellierung binärer abhängiger Daten gibt es im GEE-Ansatz eine zusätzliche Möglichkeit zur Schätzung der Abhängigkeitsstruktur innerhalb der Cluster. Anstelle des Korrelationskoeffizienten wird die Abhängigkeit zwischen den binären Beobachtungen j und k innerhalb des Clusters i über Log-Odds-Ratios, mit Hilfe einer zusätzlichen logistischen Regression geschätzt. Der Parameter α bezeichnet in diesem Fall das Log-Odds-Ratio und nicht den Korrelationskoeffizienten. Für eine austauschbare Struktur der Log-Odds-Ratios ist α ein Skalar. Folglich muss wiederum nur ein zusätzlicher Parameter mit Hilfe von:

$$\text{logit}(P(Y_{ij} = 1 | Y_{ik} = y_{ik})) = \alpha y_{ik} + \log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right) \quad (4.13)$$

geschätzt werden, wobei $\mu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$ ist. Der zweite Term auf der rechten Seite dient in der Schätzung dieser logistischen Regression als offset-Variable. Somit sind die Schätzungen für α und β voneinander abhängig (Diggle et al., 1994 S. 147; Molenbergh und Verbeke, 2005 S. 166). Die Schätzung der beiden Parameter verläuft wie folgt: In einem ersten Schritt wird für den Parametervektor β eine Schätzung unter Unabhängigkeit durchgeführt, das heißt für $\mathbf{R}_i(\alpha)$ wird die Einheitsmatrix eingesetzt. Mit Hilfe dieser Schätzung wird über die logistische Regression aus Gleichung 4.13 der Parameter α geschätzt. Diese Schätzung geht wiederum in die Schätzung des Parametervektors β ein. Bis sich eine stabile Lösung für $\hat{\beta}$ und $\hat{\alpha}$ einstellt, alterniert die Schätzung zwischen der GEE-Schätzung für β und der Schätzung für α . Aufgrund dieses Prozesses wird hier auch von einer *alternierenden logistischen Regression* gesprochen (Diggle et al., 1994 S. 147).

4.2.2 Implementierung in SAS und STATA

In SAS sind GEE-Schätzungen in der Prozedur

- PROC GENMOD

implementiert (SAS, 2004). Eine designbasierte Schätzung wurde hierin jedoch ursprünglich nicht beabsichtigt. Im WEIGHT-Statement können zwar Gewichte angegeben werden, bei der Schätzung der Arbeitskorrelation werden diese jedoch nicht im Sinne einer designbasierten Schätzung berücksichtigt. Die gewichtete Schätzgleichung in SAS entspricht Gleichung 4.12, in der Schätzung von α werden die Gewichte jedoch nicht adäquat berücksichtigt (Carillo et al. 2010).

Im Hinblick auf eine praktische Anwendung im BGS werden in der Simulationsstudie dennoch die Eigenschaften der GEE-Schätzung mit den Designgewichten im WEIGHT-Statement und austauschbarer Arbeitskorrelation untersucht. Es wird die zeitliche Korrelation innerhalb der Längsschnittprobanden berücksichtigt. Diese GEE-Schätzung wird in der Simulation mit $WGEE_Z$ bezeichnet. Außerdem wird die GEE-Schätzung mit Berücksichtigung der Intraclusterkorrelation untersucht. Diese wird in der Simulation mit $WGEE_C$ bezeichnet. Die Clusterung der Daten wird im REPEATED-Statement festgelegt.

Für binäre abhängige Variablen bietet GENMOD zusätzlich die Möglichkeit der alternierenden logistischen Regression. Es besteht die Möglichkeit mit Hilfe der SUBCLUSTER-Option eine Clusterung innerhalb der ersten Clusterstufe anzugeben. In diesem Fall wird in der logistischen Regression für α ein zusätzliches Log-Odds-Ratio für die Individuen innerhalb des Subclusters geschätzt. Diese Option bietet für binäre Merkmale eine gleichzeitige Berücksichtigung beider Korrelationsquellen. Die alternierende logistische Regression mit Berücksichtigung von Gewichten und Subclustern wird in der Simulation mit $WALR$ bezeichnet. Die ungewichteten Pendanten der drei GEE-Methoden werden in der Simulation mit GEE_Z , GEE_C und ARL bezeichnet.

In STATA ist der GEE-Ansatz über den Befehl

- xtgee

implementiert. Über die Option *pweight* können Designgewichte angegeben werden. Mit dieser Option werden designbasierte Schätzungen für die Parameter der Grundgesamtheit durchgeführt. Die Gewichte müssen jedoch innerhalb eines Clusters konstant sein (STATA, 2011 S. 125). Da im BGS die Gewichte weder innerhalb der Sample Points noch innerhalb der Längsschnittprobanden konstant sind, ist die GEE-Implementierung in STATA für die vorliegenden Daten nicht geeignet.

4.3 Generalisierte lineare gemischte Modelle

Die Methoden GLM und GEE gehören in die Klasse der marginale Modelle. In dieser Modellklasse messen die geschätzten Parameter durchschnittliche („population averaged“) Effekte in der Grundgesamtheit (Neuhaus, 1992; Liang und Zeger, 1986).

Für die Modellierung von Daten aus mehrstufigen Stichprobendesigns schlagen zum Beispiel Rabe-Hesketh und Skrondal (2006) und Peffermann et al. (1998) die Verwendung von gemischten Modellen vor. In dieser Modellklasse werden nicht marginale, sondern subjektspezifische Effekte, bedingt auf zufällige Effekte geschätzt.

4.3.1 Arbeitsmodell und Schätzgleichung

Um die Korrelationsstruktur des BGS zu berücksichtigen können cluster- und individuen-spezifische zufällige Effekte in Form von zufälligen Intercepts in das Arbeitsmodell integriert werden. Diese zufälligen Intercepts beschreiben die unbeobachtete Heterogenität zwischen den Clustern beziehungsweise zwischen den Probanden. Diese unbeobachtete Heterogenität ist per Annahme verantwortlich für die Korrelation der Responsevariable y_{ijt} . Bedingt auf die zufälligen Effekte wird Unabhängigkeit zwischen den Beobachtungen angenommen. Das Arbeitsmodell für den bedingten Erwartungswert von y_{ijt} eines GLMM mit zufälligen Intercepts kann durch

$$\mu_{ijt} = E(y_{ijt} | \gamma_i, \mathbf{v}_{ij}) = g(\mathbf{x}'_{ijt} \boldsymbol{\beta} + \gamma_i + \mathbf{v}_{ij}) \quad (4.14)$$

ausgedrückt werden, wobei γ_i der clusterspezifische und \mathbf{v}_{ij} der individuen-spezifische zufällige Intercept ist. Für die zufälligen Effekte wird eine Normalverteilung angenommen mit $\gamma_i \sim N(0, \sigma_\gamma^2)$ und $\mathbf{v}_{ij} \sim N(0, \sigma_v^2)$. Der Parametervektor $\boldsymbol{\beta}$ beinhaltet die festen Effekte der Kovariablen \mathbf{x}_{ijt} auf die abhängige Variable y_{ijt} , bedingt auf die zufälligen Effekte. Im linearen Fall sind diese bedingten Effekte auf Grund der linearen Responsefunktion identisch mit den marginalen Effekten aus dem GLM- und GEE-Ansatz (Rabe-Hesketh und Skrondal (2006)).

Für das logistische Regressionsmodell trifft dies nicht zu. Die bedingten Effekte sind durch die logistische Responsefunktion von der Varianz der zufälligen Effekte abhängig (Rabe-Hesketh und Skrondal, 2006). Neuhaus (1992) zeigt, dass die marginalen Effekte des GLM- und des GEE-Ansatzes betragsmäßig kleiner sind als die bedingten Effekte. Der Unterschied ist umso größer, je größer die Varianz der zufälligen Effekte ist. Da die Schätzung der durchschnitt-

lichen Zeiteffekte in der Grundgesamtheit im Vordergrund steht wird in der Simulation der GLMM-Ansatz nur für den linearen Fall untersucht.

Die Schätzung der festen Parameter β erfolgt über die marginale Verteilung $f(\mathbf{y}_i|\beta)$ aus der die zufälligen Parameter γ_i über die Methode des „Herausintegrierens“ (integrating out) eliminiert werden. Bedingt auf die zufälligen Effekte werden die Beobachtungen der Zielvariable y_{ijt} als unabhängig angenommen. Die marginale Likelihoodfunktion kann dann durch (Molenbergh und Verbeke, 2005 S. 267)

$$L(\beta, \sigma_\gamma^2, \sigma_v^2) = \prod_{i=1}^n \int \left[\prod_{j=1}^{n_i} \int \left[\prod_{t=1}^2 f(y_{ijt}|\beta, \gamma_i, \mathbf{v}_{ij}) \right] f(\mathbf{v}_{ij}) d\mathbf{v}_{ij} \right] f(\gamma_i) d\gamma_i \quad (4.15)$$

dargestellt werden, wobei $f(\mathbf{v}_{ij})$ die Dichtefunktion des individuenspezifischen und $f(\gamma_i)$ die des clusterspezifischen zufälligen Effekts ist.

Im Allgemeinen gibt es keinen analytischen Ausdruck für die Lösung eines mehrdimensionalen Integrals dieser Art. Für die Lösung werden in der Regel numerische Näherungsverfahren verwendet.⁹ Der Log-Likelihoodbeitrag einer Einheit ij der zweiten Ziehungsstufe kann durch

$$\begin{aligned} l_{ij}(\beta, \sigma_\gamma^2|\gamma_i) &= \log \left[\int \left[\prod_{t=1}^2 f(y_{ijt}|\beta, \gamma_i, \mathbf{v}_{ij}) \right] f(\mathbf{v}_{ij}) d\mathbf{v}_{ij} \right] \\ &= \log \left[\int \exp \left[\sum_{t=1}^2 l(y_{ijt}|\beta, \gamma_i, \mathbf{v}_{ij}) \right] f(\mathbf{v}_{ij}) d\mathbf{v}_{ij} \right] \end{aligned} \quad (4.16)$$

ausgedrückt werden, wobei $l(y_{ijt}|\beta, \gamma_i, \mathbf{v}_{ij})$ der Logarithmus der bedingten Dichtefunktion $f(y_{ijt}|\beta, \gamma_i, \mathbf{v}_{ij})$ ist (Rabe-Hesketh und Skrondal, 2006). Dieser Logik folgend ist der Log-Likelihoodbeitrag eines Clusters i :

$$l_i(\beta, \sigma_\gamma^2, \sigma_v^2) = \log \left[\int \exp \left[\sum_{j=1}^{n_i} l_{ij}(\beta, \sigma_\gamma^2|\gamma_i) \right] f(\gamma_i) d\gamma_i \right]. \quad (4.17)$$

Die gesamte Log-Likelihood hat die Form:

$$l(\beta, \sigma_\gamma^2, \sigma_v^2) = \sum_{i=1}^n l_i(\beta, \sigma_\gamma^2, \sigma_v^2). \quad (4.18)$$

Dieses rekursive Prinzip kann um beliebig viele Stufen erweitert werden (Rabe-Hesketh und Skrondal, 2006). Das designbasierte Pendant zu Gleichung (4.15) für eine PLM-Schätzung

⁹Eine Übersicht solcher Verfahren liefern zum Beispiel Molenberghs und Verbeke (2005 S. 273 ff.)

mit separater Gewichtung in jeder Stufe lautet:

$$L(\beta_p, \sigma_\gamma^2, \sigma_v^2) = \prod_{i=1}^n \left\{ \int \left[\prod_{j=1}^{n_i} \int \left[\prod_{t=1}^2 f(y_{ijt} | \beta_p, \gamma_i, v_{ij})^{w_{t|ij}} \right] f(v_{ij}) dv_{ij} \right]^{w_{j|i}} f(\gamma_i) d\gamma_i \right\}^{w_i} \quad (4.19)$$

wobei w_i die inverse Ziehungswahrscheinlichkeit von Cluster i ist. Der Term $w_{j|i}$ ist die Inverse der Wahrscheinlichkeit Individuum j zu ziehen, bedingt auf die vorherige Ziehung von Cluster i . Der Term $w_{t|ij}$ bezeichnet die Inverse der bedingten Wahrscheinlichkeit die Einheit j zu Zeitpunkt t zu beobachten. Aus den Gleichungen 4.17 und 4.18 ist ersichtlich, dass durch die Logarithmierung der Pseudo-Likelihood die einzelnen Gewichte an verschiedenen Stellen eingehen, und sich nicht zu einem globalen Designgewicht zusammenfassen lassen. Eine Gewichtung mit dem Produkt der einzelnen Gewichte, wie im GLM- und GEE-Ansatz, ist in diesem Ansatz nicht möglich (Pfeffermann und Sverchkov, 2009 S. 481, Pfeffermann et al. 1998). Rabe-Hesketh und Skrondal (2006) merken an, dass dies auch für ursprünglich selbstgewichtende Designs zutrifft.

Die Differenzierung der Log-Pseudolikelihood liefert bezüglich des Populationsparameters wieder die Schätzgleichung. Die Varianz von β_p kann analog zu den vorangegangenen Methoden über die Sandwichform aus Gleichung 3.33 geschätzt werden.

Hinsichtlich der praktischen Umsetzung dieses Ansatzes für die designbasierte Auswertungen könnten die erforderlichen getrennten Inklusionswahrscheinlichkeiten problematisch sein. In veröffentlichten Datensätzen werden oftmals lediglich globale Designgewichte mitgeliefert (Pfeffermann et al., 1998).

Sind die getrennten Gewichte bekannt, empfehlen zum Beispiel Pfeffermann et al. (1998) und Asparouhov (2006) eine Skalierung der Gewichte, um eine Verzerrung der geschätzten Parameter zu vermeiden. Die Eigenschaft der geschätzten Parameter bleibt hierbei unbeeinflusst von einer Skalierung der Gewichte der ersten Ziehungsstufe. Die Gewichte weiterer Ziehungsstufen hingegen sollten mit Hilfe einer Konstanten λ transformiert werden. Für die Berechnung von λ werden in Pfeffermann et al. (1998) zwei Methoden vorgeschlagen. Methode 1 beinhaltet eine Normierung der Gewichte auf die effektive Stichprobengröße der jeweiligen Ziehungsstufe. Für Gewichte der zweiten Stufe kann die Normierungskonstante $\tilde{\lambda}_i$ durch

$$\tilde{\lambda}_i = \frac{\sum_{j=1}^{n_j} w_{j|i}}{\sum_{j=1}^{n_j} w_{j|i}^2} \quad (4.20)$$

quantifiziert werden. Das skalierte Gewicht der zweiten Stufe ist dann $w_{j|i}^* = w_{j|i} \tilde{\lambda}_i$. Methode 2 beinhaltet eine Normierung der Gewichte auf die Stichprobengröße n_i von Cluster i . Die Konstante $\bar{\lambda}_i$ ist dann durch

$$\bar{\lambda}_i = \frac{n_i}{\sum_{j=1}^{n_j} w_{j|i}} \quad (4.21)$$

definiert. Das skalierte Gewicht der zweiten Stufe ist dann $w_{j|i}^\circ = w_{j|i} \bar{\lambda}_i$. Für Gewichte weiterer Stufen kann die Skalierung analog durchgeführt werden. In Simulationstudien zur Untersuchung der Eigenschaften der geschätzten Parameter in Abhängigkeit zur Skalierungsmethode hat sich eine Präferenz für Methode 2 eingestellt (Rabe-Hesketh und Skrondal, 2006; Pfeffermann et al., 1998). Bisher hat sich jedoch keine Standardmethode zur Skalierung der Gewichte etabliert (Carle, 2009). Eine praxisrelevante Empfehlung für die Anwendung GLMM-Ansatzes im BGS wird daher erschwert.

4.3.2 Implementierung in SAS und STATA

In SAS können mit Hilfe der Prozedur

- PROC GLIMMIX

gemischte Modelle geschätzt werden (SAS, 2004). Das WEIGHT-Statement wurde jedoch nicht für eine Designgewichtung konzipiert (Chantala, Blanchette und Suchindran, 2011). Die Gewichte können demnach nicht getrennt in die PML-Schätzung integriert werden.

Für STATA wurde das Zusatzpaket GLLAMM entwickelt (Rabe-Hesketh, Skrondal und Pickles, 2004), mit dem gemischte Modelle mit Designgewichtung geschätzt werden können. Die Skalierung der Gewichte ist hierbei dem Benutzer überlassen. In der aktuellen Version 12 von STATA können lineare gemischte Modelle mit Hilfe des Befehls

- xtmixed

designbasiert geschätzt werden (STATA, 2011). Eine Skalierung der Gewichte ist hier direkt implementiert. Der Anwender muss lediglich mit Hilfe der Option *pwscale()* spezifizieren nach welcher Skalierungsmethode die Gewichte transformiert werden sollen. In der nachfolgenden Simulation wird die GLMM-Schätzung unter Berücksichtigung der Designgewichte mit *WGLMM* und das ungewichtete Pendant mit *GLMM* gekennzeichnet.

4.4 Zusammenfassung

Die beschriebenen Methoden unterscheiden sich hinsichtlich ihrer Komplexität. Die GLM-Methoden gehen bei der Parameterschätzung von unabhängigen Beobachtungen aus. Durch die robuste Sandwichform der Varianzschätzung sind zudem gültige Inferenzstatistiken zu erwarten. Da die vorliegende Abhängigkeitsstruktur ignoriert wird, können die Schätzer ineffizient sein.

In der GEE-Schätzung werden vermutete Korrelationsstrukturen über eine Arbeitskorrelationsmatrix in die Schätzgleichung aufgenommen. Hierbei können in SAS und STATA lediglich Abhängigkeiten in einer Ebene berücksichtigt werden. Eine Ausnahme bildet die alternierende logistische Regression in SAS, in der eine zusätzlich eine Abhängigkeit innerhalb von Subclustern berücksichtigt werden kann.

Im GLMM-Ansatz kann mit Hilfe von zufälligen Intercepts die Abhängigkeitsstruktur der BGS-Daten berücksichtigt werden. Für eine designbasierte Schätzung ist die Kenntnis der Inklusionswahrscheinlichkeiten verschiedener Ziehungsstufen notwendig. Um dem Anwender diese Schätzmethode zu ermöglichen müssen diese im BGS-Datensatz zur Verfügung gestellt werden.

In der nachfolgenden Simulationsstudie werden die Eigenschaften der Schätzmethoden für den Zeiteffekt unter verschiedenen Abhängigkeitsszenarien in den Daten untersucht.

5 Simulationsstudie

In der Simulationsstudie wird mit einem Monte-Carlo Experiments untersucht, inwieweit sich die Punktschätzungen Zeiteffekts zwischen den vorgestellten Methoden in Bezug auf ihre Verzerrung und ihre Effizienz unterscheiden. Zunächst werden Vorüberlegungen zur Simulation der Grundgesamtheit, zur Abhängigkeitsstruktur und zur Stichprobenziehung angestellt.

5.1 Vorüberlegungen

Im Mittelpunkt der Analyse steht die Auswirkung der Abhängigkeitsstruktur auf die Eigenschaften der Schätzungen des Zeiteffekts. Aus diesem Grund werden für die Simulation verschiedene Intraclusterkorrelationen mit unterschiedlichen zeitlichen Korrelationen innerhalb der beobachteten Individuen kombiniert. Um die Datenlage im BGS widerzuspiegeln, werden dichotome und metrische Merkmale für eine endliche Grundgesamtheit simuliert. Im Fall von metrischen Merkmalen wird zusätzlich zwischen symmetrischen und rechtsschief verteilten Variablen unterschieden. In der Realität können die metrischen Merkmale selten durch eine Normalverteilung approximiert werden. Im BGS werden oft rechtsschiefe Verteilungen bei den abgefragten Merkmalen beobachtet. Dies ist für die Simulation interessant, da in der Praxis mitunter abhängige Variable nicht transformiert werden, um sie zu normalisieren. Ob dieses Vorgehen Auswirkungen auf die Schätzung des Zeiteffekts hat, wird ebenfalls untersucht.

5.1.1 Modell der Grundgesamtheit

Die Korrelationsstruktur innerhalb der abhängigen Variablen wird mittels zufälliger cluster-spezifischer und zufälliger individuenspezifischer Effekte erzeugt.

Im linearen Fall werden die Merkmale y_{ijt} in der Grundgesamtheit durch das nachfolgende

Modell generiert:

$$y_{ijt} = \beta x_t + \sigma_c v_i + \sigma_z \gamma_{ij} + \varepsilon_{ijt}. \quad (5.1)$$

Die binären Merkmale y_{ijt} werden mit einem latenten Schwellenwertmodell (Rabe-Hesketh und Skrondal, 2006) generiert:

$$y_{ijt}^* = \beta x_t + \sigma_c v_i + \sigma_z \gamma_{ij} + \varepsilon_{ijt} \quad (5.2)$$

$$y_{ijt} = \begin{cases} 1 & \text{wenn } y_{ijt}^* > 0 \\ 0 & \text{wenn } y_{ijt}^* \leq 0 \end{cases}$$

wobei die Beobachtung ijt entsprechend den vorangegangenen Kapiteln der Einheit j aus Cluster i zum Zeitpunkt t entspricht. Der Parameter β ist der interessierende Zeiteffekt. Die Kovariable x_t ist ein Zeitindikator mit $x_t = 0$, für Beobachtungen vom ersten Zeitpunkt und $x_t = 1$ für Beobachtungen, die vom zweiten Zeitpunkt stammen. Der clusterspezifische zufällige Effekt ist zusammengesetzt aus der multiplikativen Konstanten σ_c und v_i mit $v_i \sim N(0, 1)$. Der gesamte zufällige Effekt $\sigma_c v_i$ ist dann ebenfalls normalverteilt mit einem Erwartungswert von 0 und einer Varianz von σ_c^2 . Der individuenspezifische, zufällige Effekt ist zusammengesetzt aus σ_z und γ_{ij} mit $\gamma_{ij} \sim N(0, 1)$. Der gesamte zufällige Effekt $\sigma_z \gamma_{ij}$ ist dann ebenfalls normalverteilt mit einem Erwartungswert von 0 und einer Varianz von σ_z^2 . Die multiplikativen Konstanten σ_c und σ_z werden in der Simulation variiert, um die Abhängigkeitsstruktur innerhalb der Merkmale y_{ijt} zu kontrollieren. Die Erzeugung der Abhängigkeitsstrukturen wird im nächsten Unterabschnitt erläutert.

Für die symmetrisch verteilten Merkmale werden die unabhängig und identisch Verteilten Fehlerterme ε_{ijt} aus Gleichung 5.1 mit Hilfe einer Standardnormalverteilung generiert. Für die schiefen Merkmale werden die ε_{ijt} durch eine Log-Normalverteilung generiert mit $\exp(\varepsilon_{ijt}) \sim N(0, 0, 48)$. Der Störterm ε_{ijt} hat dann eine Varianz von 1 und einen Erwartungswert von 1,27. Die rechtsschiefe Verteilung des Merkmals wird durch die Verteilung des Fehlerterms ε_{ijt} erzeugt. Im Fall des binären Merkmals sind die ε_{ijt} ebenfalls unabhängig und identisch verteilt und werden durch eine logistische Verteilung mit Erwartungswert 0 einer Varianz von $\pi^2/3$ generiert.

5.1.2 Abhängigkeitsstruktur in der Grundgesamtheit

Die Intraclusterkorrelationen innerhalb der Cluster (ICC_c) und die zeitliche Korrelation (ICC_z) können mit einer Varianzzerlegung berechnet werden. Für die Modelle aus den Gleichungen 5.1 und 5.2 ergeben sich folgende Varianzen und Kovarianzen für die abhängige Variable y_{ijt} bzw. die latente abhängige Variable y_{ijt}^* :

$$\begin{aligned} Cov(y_{ijt}, y_{ist}) &= E[y_{ijt} y_{ist}] = E[\sigma_c v_i + \sigma_z \gamma_{ij} + \varepsilon_{ijt}] E[\sigma_c v_i + \sigma_z \gamma_{is} + \varepsilon_{ist}] \\ &= \sigma_c^2 E[v_i^2] = \sigma_c^2 \\ Cov(y_{ij1}, y_{ij2}) &= E[y_{ij1} y_{ij2}] = E[\sigma_c v_i + \sigma_z \gamma_{ij} + \varepsilon_{ij1}] E[\sigma_c v_i + \sigma_z \gamma_{ij} + \varepsilon_{is2}] \\ &= \sigma_c^2 + \sigma_z^2 \\ Var(y_{ijt}) &= E[y_{ijt}, y_{ijt}] = \sigma_c^2 + \sigma_z^2 + \sigma_\varepsilon^2 \end{aligned}$$

wobei $Cov(y_{ijt}, y_{ist})$ die paarweise Kovarianz zwischen den Individuen innerhalb des Clusters i zum Zeitpunkt t ist. $Cov(y_{ij1}, y_{ij2})$ stellt die Kovarianz innerhalb eines Individuum ij dar und $Var(y_{ijt})$ ist die Gesamtvarianz über alle Beobachtungen ijt .

Aus den Varianzkomponenten ergeben sich für ICC_c und ICC_z (Rabe-Hesketh und Skrondal, 2006):

$$ICC_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_z^2 + \sigma_\varepsilon^2} \quad (5.3)$$

$$ICC_z = \frac{\sigma_c^2 + \sigma_z^2}{\sigma_c^2 + \sigma_z^2 + \sigma_\varepsilon^2} \quad (5.4)$$

Der Term σ_ε^2 ist die Varianz des Störterms ε_{ijt} . Da σ_ε^2 in der Simulation vorgegeben ist, gibt es für jede Kombination von ICC_{clus} und ICC_{zeit} eine eindeutige Zuordnung für σ_c und σ_z . Als Richtwerte für die Stärke der Abhängigkeit werden die tatsächlich beobachteten Korrelationskoeffizienten aus dem BGS herangezogen. Innerhalb der Gemeinden wurde im Mittel eine ICC_c von 0,02 beobachtet. Zusätzlich werden der kleinste (0,001) und der größte (0,06) beobachtete Wert in der Simulation verwendet. Innerhalb der Längsschnittprobanden ist das Niveau der Abhängigkeit bedeutend höher. Hier wurde im Mittel eine zeitliche Korrelation von 0,2 beobachtet. Die obere und untere Grenze bilden die Werte 0,1 und 0,4. In der Simulation werden beide Korrelationen miteinander kombiniert, so dass neun Abhängigkeitsszenarien entstehen.

Die Varianz des Fehlerterms ε_{ijt} wurde für die Simulation im linearen Modell auf 1 und im

logistischen Modell auf $\pi^2/3$ festgelegt. Da σ_c und σ_z in den Gleichungen 5.3 und 5.4 von der Varianz des Fehlerterms ε_{ijt} abhängen, unterscheiden sich die Werte zwischen den beiden Modellklassen. Tabelle 5.1 fasst die verschiedenen Kombinationen der Korrelationen zusammen und bildet die zugehörigen Standardfehler der zufälligen Effekte für das lineare und das logistische Modell ab.

ICC_c	ICC_z	Linear		Logistisch	
		σ_c	σ_z	σ_c	σ_z
0,001	0,100	0,038	0,331	0,063	0,601
0,020	0,100	0,149	0,298	0,271	0,540
0,060	0,100	0,258	0,211	0,468	0,382
0,001	0,200	0,032	0,499	0,059	0,905
0,020	0,200	0,159	0,474	0,288	0,860
0,060	0,200	0,274	0,418	0,496	0,759
0,001	0,400	0,034	0,816	0,063	1,480
0,020	0,400	0,183	0,796	0,331	1,444
0,060	0,400	0,316	0,753	0,573	1,366

Tabelle 5.1: In der Simulation verwendete Kombinationen von ICC_c und ICC_z und dazugehörige Standardfehler der zufälligen Effekte für metrische und dichotome Merkmale

5.1.3 Generierung der Grundgesamtheit

Die generierte Grundgesamtheit besteht aus $M = 1000$ verschiedenen großen Clustern. Die Clustergrößen N_i werden mit Hilfe der zufälligen Effekte v_j aus den beiden Modellgleichungen 5.1 und 5.2 erzeugt:

$$N_i = 1000 \exp(v_i^*) \quad (5.5)$$

wobei
$$v_i^* = \begin{cases} -1.5 & \text{wenn } v_i \leq -1.5 \\ 2.5 & \text{wenn } v_i \geq 2.5 \end{cases}$$

Die v_i^* sind identisch mit den Realisationen der clusterspezifischen zufälligen Effekte v_i aus den Gleichungen 5.1 und 5.2, wobei die v_i^* von unten bei -1.5 und von oben bei 2.5 trunziert sind. Diese Vorgehensweise zur Erzeugung der Clustergrößen ist an die Simulationsstudie von Pfeffermann et al. (1998) angelehnt. Die Trunkierung sorgt dafür nicht zu kleine Cluster zu generieren und die Cluster nicht zu groß werden zu lassen. Die Größen liegen in einem Intervall von mindestens 223 und höchstens 12.182 Individuen pro Cluster. Die erwartete Clustergröße liegt bei ungefähr 1.600 Individuen. Die erzeugten Clustergrößen folgen einer schiefen Vertei-

lung. Grundgesamtheit umfasst insgesamt 1.645.701 individuelle Beobachtungseinheiten mit jeweils einer Beobachtung zum Zeitpunkt $t = 1$ und $t = 2$.

Die clusterspezifischen zufälligen Effekte sind mit der Clustergröße assoziiert. Dieser Zusammenhang ist umso stärker, desto mehr die zufälligen Effekte variieren beziehungsweise desto größer σ_c ist. Zusätzlich wird in der Grundgesamtheit festgelegt welche Individuen bei einer Befragung zum Merkmal y_{ijt} eine Antwort geben. Die sogenannte Non-Response soll mit der Ausprägung y_{ijt} assoziiert sein. Dies ist ein durchaus realistischer Zusammenhang für Gesundheitsbefragungen. In der Praxis wurde beobachtet, dass Probanden mit schlechterer Gesundheit im Mittel seltener bereit sind an Befragungen teilzunehmen als gesündere Probanden (Thelfeld, Stolzenberg und Bellach, 1999). Zunächst wurde eine Schichtvariable s generiert:

$$s = \begin{cases} 1, & \text{wenn } \gamma_{ij} > 0 \\ 2, & \text{wenn } \gamma_{ij} \leq 0. \end{cases}$$

Die Schichtzugehörigkeit eines Individuums ist an die Realisierung des individuenspezifischen Effekts γ_{ij} gekoppelt. Des Weiteren wird eine zufällige binäre Responsevariable R mit $r_{ij} \sim \text{Bin}(1, p_s)$ generiert, wobei p_s die Wahrscheinlichkeit eines Individuums ist, an einer Befragung zum Merkmal y_{ij1} teilzunehmen. In Schicht 1 wird p_s auf 0,4 und in Schicht 2 auf 0,6 festgelegt, sodass im Mittel eine Teilnahmewahrscheinlichkeit von 50% zu erwarten ist. Innerhalb der Schichten ist die Teilnahme zufällig.

Um die unbalancierte Datenstruktur im BGS auch in der Simulation nachzuahmen wird zusätzlich eine Variable für die Wiederteilnahme der Individuen generiert. Die Wiederteilnahmewahrscheinlichkeit wird auf 50% festgelegt und mit einer binären Zufallsvariable B mit $b_{ij} \sim \text{Bin}(1, 0,5)$ generiert. Der binäre Indikator wird zufällig über alle Individuen generiert, deren Response zum ersten Zeitpunkt 1 ist. Die tatsächliche Wiederteilnahme kann zwischen den Clustern variieren, im Mittel liegt sie erwartungsgemäß bei 50%. Innerhalb der Cluster werden die Merkmale entsprechend der Modellgleichungen 5.1 und 5.2 generiert.

Mit den in Tabelle 5.1 dargestellten Werten für σ_{Clus} und σ_{Zeit} werden die verschiedenen Abhängigkeitsstrukturen innerhalb der Merkmale y_{ijt} erzeugt. Im linearen Modell wird der Parameter β auf 0,5 und für das logistische Regressionsmodell auf 0,25 (dies entspricht einem Odds Ratio von $\exp(\beta) = 1,284$) festgelegt.

Diese für die Simulation festgelegten Zeiteffekte entsprechen den in Unterabschnitt 3.1 beschriebenen Parametern einer hypothetischen Superpopulation, durch die die endliche Grundgesamtheit generiert wird.

Für die Beurteilung der Schätzmethode müssen die Effekte der zufällig realisierten, aber festen Grundgesamtheit berechnet werden. Diese können, wie in Abschnitt 3.1 erläutert, vom

wahren Parameter der Superpopulation abweichen. Die Schätzung der Populationsparameter erfolgt mit Hilfe der Maximierung der Populations-Log-Likelihood des jeweiligen Modells. Für die GLM-Schätzung beinhaltet das die Schätzung der Koeffizienten einer linearen beziehungsweise einer logistischen Regression. Für die Untersuchung der GEE-Methode müssen die Koeffizienten der populationsbasierten Schätzgleichung aus Gleichung 4.10 geschätzt werden. Für die Berücksichtigung der zeitlichen Abhängigkeit stellt die Schätzung numerisch kein Problem dar, da der Korrelationskoeffizient α lediglich aus zwei Beobachtungen pro Individuum geschätzt werden muss. Für die Berücksichtigung der Abhängigkeit innerhalb der Cluster war es aufgrund der Größe der Cluster numerisch jedoch nicht möglich, den Zeiteffekt in der Population zu berechnen. Im Vergleich zeigte sich, dass die Effekte der der GLM-Schätzung mit denen der GEE-Schätzung unter Berücksichtigung der zeitlichen Korrelation übereinstimmen. Die Tabelle 5.2 zeigt für die verschiedenen Korrelationen die Populationsparameter für die binären abhängigen Variablen.

ICC_c	ICC_z	GLM	GEE
0,001	0,100	0,230	0,230
0,020	0,100	0,231	0,231
0,060	0,100	0,232	0,232
0,001	0,200	0,212	0,212
0,020	0,200	0,213	0,213
0,060	0,200	0,215	0,215
0,001	0,400	0,180	0,180
0,020	0,400	0,181	0,181
0,060	0,400	0,182	0,182

Tabelle 5.2: Populationsparameter für binäre abhängige Variablen

Anhand dieser Tabelle wird deutlich, dass der Parameter einer logistischen Regression von der Varianz der zufälligen Effekte abhängt. Wie in Abschnitt 4.3 erwähnt unterscheiden sich die marginalen Effekte in der logistischen Regression von den bedingten Effekten eines gemischten Modells. Da die Abhängigkeitsstruktur mit Hilfe zufälliger Effekte generiert wurde, spiegelt sich dies hier wider.

Im linearen Modell ergab sich für die normalverteilte abhängige Variable und die abhängige Variable mit log-normalverteiltem Fehlerterm ein Populationsparameter von 0,498. Auch hier unterscheiden sich die Populationsparameter der GLM-Schätzung nicht von der GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation. Für die Beurteilung der GEE-Schätzung unter Berücksichtigung der Abhängigkeit innerhalb der Cluster werden ebenfalls die Populationsparameter der GLM-Schätzung herangezogen.

5.1.4 Stichprobenziehung

Die Stichprobenziehung erfolgt in einem zweistufigen Design, welches in vereinfachter Form dem Ziehungsmechanismus im BGS entspricht.

In der ersten Stufe werden $m = 120$ Cluster ohne Zurücklegen mit einer Ziehungswahrscheinlichkeit proportional zur Größe ($\pi_i = mN_i / \sum_{i=1}^M N_i$) gezogen. Aus Gleichung 5.5 ist ersichtlich, dass die Clustergrößen in einem positiven Zusammenhang mit der Realisierung des zufälligen Effekts v_i und somit auch zur abhängigen Variable y_{ijt} stehen. Die Ziehungswahrscheinlichkeiten der ersten Stufe sind also ebenfalls mit der abhängigen Variable y_{ijt} korreliert.

In der zweiten Stufe werden aus den beiden Schichten eines Clusters jeweils 40 Individuen durch einfaches Ziehen ohne Zurücklegen gezogen. Die durchschnittliche Responserate liegt bei 50%, sodass im Mittel 40 Beobachtungen pro Cluster zu erwarten sind. Innerhalb der Cluster liegt entsprechend der unterschiedlichen Responseraten für Schicht 1 die erwartete Anzahl an Beobachtungen bei 16 und in Schicht 2 bei 24. Die tatsächlichen Stichprobengrößen können jedoch schwanken.

Die Ziehungswahrscheinlichkeit $\pi_{j|i}^s$ eines Individuums in Schicht s , bedingt auf die vorherige Ziehung von Cluster i , ist $\pi_{j|i}^{(s)} = \frac{n_i^s}{N_i^s}$, wobei N_i^s die Größe von Schicht s in Cluster i ist. Die unbedingte Inklusionswahrscheinlichkeit eines Individuums ist dann $\pi_{ij}^{(s)} = \pi_i \pi_{j|i}^{(s)}$.

Zusätzlich wird die bedingte Teilnahmewahrscheinlichkeit der Individuen zu den beiden Zeitpunkten $\pi_{t|ij}$ benötigt. Da die Non-Response der gezogenen Individuen bereits in $\pi_{j|i}^{(s)}$ berücksichtigt wurde, ist für $t = 1$ $\pi_{t|ij} = 1$ und für $t = 2$ ist $\pi_{t|ij} = 0.5$

Die zusammengesetzte Inklusionswahrscheinlichkeit π_{ijt} einer Beobachtung ijt ist das Produkt aus der Ziehungswahrscheinlichkeit Clusters i , der bedingten Ziehungswahrscheinlichkeit der Individuen ij und der bedingten Teilnahmewahrscheinlichkeit zu den unterschiedlichen Zeitpunkten: $\pi_{ijt} = \pi_i \pi_{j|i} \pi_{ij|t}$. Das Designgewicht einer Beobachtung ijt ist die inverse Inklusionswahrscheinlichkeit $w_{ijt} = 1/\pi_{ijt}$.

Für das im Abschnitt 4.3 beschriebene gemischte Modell werden die Gewichte der drei Stufen getrennt in die Schätzung integriert. Für die Skalierung der Gewichte wird Methode 2 aus Abschnitt 4.3 in der Simulation angewendet. Die Gewichte werden also auf die Stichprobengröße der jeweiligen Stufen normiert. Die folgende Übersicht fasst die Eigenschaften der Grundgesamtheit und der Stichprobenziehung zusammen.

Grundgesamtheit

- Die endliche Population besteht aus $M = 1000$ Clustern, deren Größen N_i mit der Realisation von v_i assoziiert sind.
- Innerhalb der Cluster gibt es zwei Schichten. Die Schichtzugehörigkeit der Individuen hängt von der Realisation des individuenspezifischen zufälligen Effektes γ_{ij} ab.
- Die Teilnahmebereitschaft in Schicht 1 liegt bei 40% und in Schicht 2 bei 60% . Die Wiederteilnahmebereitschaft liegt bei 50%
- Die Beobachtungen werden durch die Modelle der Gleichungen 5.1 beziehungsweise 5.2 generiert.
- Die Abhängigkeitsstruktur der Beobachtungen wird über die Variabilität der zufälligen Effekte in Tabelle 5.1 kontrolliert.

Stichprobendesign

- Die Stichproben werden durch ein zweistufiges Verfahren gezogen.
 - In der ersten Stufe werden $m = 120$ Cluster mit einer Wahrscheinlichkeit proportional zur Clustergröße ohne Zurücklegen gezogen.
 - In der zweiten Stufe werden jeweils 40 Individuen durch einfaches Ziehen ohne Zurücklegen aus beiden Schichten gezogen.
 - Die Ziehungswahrscheinlichkeiten sind über v_i und γ_{ij} mit dem Merkmal y_{ijt} korreliert.
-

Abbildung 5.1: Eigenschaften der endlichen Grundgesamtheit und des Stichprobendesigns

5.2 Vergleichskriterien für die Simulationsstudie

Vor der Präsentation der Simulationsresultate, werden Vergleichskriterien vorgestellt, um die Eigenschaften der Schätzungen beurteilen zu können.

Ein geläufiges Kriterium zur Beurteilung von Parameterschätzungen ist die Abdeckungswahrscheinlichkeit, die nachfolgend als Coverage (C) bezeichnet wird. Die Coverage ist der Anteil der Versuche, in denen das Konfidenzintervall der Schätzung den Parameter der Grundgesamtheit β_N enthält, gemessen an der Gesamtzahl aller Simulationsversuche. Die Coverage sollte das festgelegte Konfidenzniveau von 95% nicht unterschreiten, um den Fehler erster Art nicht zu groß werden zu lassen. Sie sollte jedoch auch nicht viel konservativer sein. Wird das geforderte Konfidenzniveau weit überschritten, weist dies auf eine ineffiziente Parameterschätzung hin. Die Varianzen der geschätzten Koeffizienten wären in diesem Fall unnötig groß.

Als weiteres Vergleichskriterium wird die relative Verzerrung (RV) betrachtet. Diese gibt Aufschluss über die durchschnittliche relative Abweichung des geschätzten Parameters $\hat{\beta}$ vom

Populationswert β_N :

$$RV(\hat{\beta}) = \frac{\sum_{k=1}^L \hat{\beta}_k - \beta_N}{\beta_N}, \quad (5.6)$$

wobei L die Anzahl der Versuche in der Simulation ist.

Als drittes Kriterium für die Beurteilung der verschiedenen Schätzmethoden wird die durchschnittliche Standardabweichung (SD) der geschätzten Parameter untersucht. Hierbei würde bei gleicher Abdeckungswahrscheinlichkeit und vergleichbarer relativer Verzerrung die Schätzmethode mit der kleineren Standardabweichung bevorzugt werden.

Als viertes Kriterium wird die Komplexität der Methoden beurteilt. In Hinblick auf eine praktikable Empfehlung für die Schätzung des Zeiteffekts im BGS wird die Methode bevorzugt, die unter sonst gleichen Eigenschaften eine geringere Komplexität aufweist.

In der Simulation wurden aus der endlichen Grundgesamtheit 1.000 Stichproben nach dem beschriebenen Stichprobendesign gezogen und die Zeiteffekte mit Hilfe der in Kapitel 4 behandelten Methoden geschätzt. Die einzige Ausnahme bildet die Simulation des GLMM. Aufgrund der langen Rechendauer wurde die Simulation nur für 500 Stichproben wiederholt.¹⁰

5.3 Resultate

Zunächst werden die stetigen Merkmale betrachtet. Anschließend werden die Simulationsergebnisse der binären abhängigen Variablen vorgestellt.

In Tabelle 5.3 sind die Ergebnisse für die gewichtete Schätzung und in Tabelle 5.4 sind die Resultate der nichtgewichteten Schätzungen für das normalverteilte Merkmal abgebildet. Auffällig ist, dass die gewichtete GEE-Schätzung, in der die Korrelation innerhalb der Cluster berücksichtigt wird ($WGEE_C$), mit steigender ICC_c im Vergleich zu den anderen Methoden sehr verzerrte Ergebnisse liefert. Der wahre Populationsparameter konnte zwar nicht ermittelt werden, es scheint jedoch nicht plausibel, dass der Populationsparameter dieser GEE-Schätzgleichung in dem Ausmaß von der eines einfachen linearen Modells abweicht. Bei der Kombination $ICC_z = 0,4$ und $ICC_c = 0,06$ ist die stärkste relative Verzerrung zu beobachten. Die relative Verzerrung von -2,1 zeigt an, dass durchschnittlich im Vergleich zum Populationsparameter ein doppelt so großer und zudem gegenläufiger Effekt geschätzt wurde.

Auch bei durchschnittlicher Abhängigkeit der Daten ($ICC_c = 0,02$ und $ICC_z = 0,2$) zeigt die $WGEE_C$ -Schätzung im Vergleich zu den anderen Methoden eine deutlich höhere relative Verzerrung, gepaart mit einer Coverage von nur knapp 91%. Für die Schätzung des Zeitef-

¹⁰Die Generierung der Grundgesamtheit und Stichprobenziehung wurden in STATA durchgeführt. Der Programmcode ist im elektronischen Anhang nachzulesen.

fekts scheint die gewichtete GEE-Schätzung mit Berücksichtigung der Korrelation innerhalb der Cluster ungeeignet.

ICC_z	ICC_c	0,001			0,02			0,06		
		Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	$WGLM_1$	0,950	0,001	0,027	0,948	0,001	0,027	0,954	0,002	0,026
	$WGLM_2$	0,951	0,001	0,027	0,949	0,001	0,026	0,954	0,002	0,026
	$WGEE_Z$	0,954	0,001	0,027	0,949	0,001	0,027	0,950	0,001	0,027
	$WGEE_C$	0,950	-0,001	0,026	0,943	-0,016	0,026	0,910	-0,034	0,026
	$WGLMM$	0,940	0,002	0,039	0,948	0,002	0,039	0,946	0,002	0,038
0.2	$WGLM_1$	0,953	0,001	0,027	0,951	0,001	0,027	0,954	0,001	0,027
	$WGLM_2$	0,954	0,001	0,027	0,949	0,001	0,027	0,955	0,001	0,027
	$WGEE_Z$	0,952	0,001	0,027	0,950	0,001	0,027	0,950	0,001	0,027
	$WGEE_C$	0,953	-0,003	0,027	0,909	-0,034	0,027	0,563	-0,112	0,027
	$WGLMM$	0,944	0,003	0,039	0,952	0,002	0,039	0,954	0,002	0,039
0.4	$WGLM_1$	0,950	-0,000	0,029	0,951	-0,000	0,029	0,954	-0,000	0,029
	$WGLM_2$	0,951	-0,000	0,029	0,948	-0,000	0,029	0,954	-0,000	0,029
	$WGEE_Z$	0,952	-0,000	0,028	0,952	-0,000	0,028	0,954	-0,000	0,028
	$WGEE_C$	0,941	-0,019	0,029	0,301	-1,337	0,050	0,000	-2,107	0,063
	$WGLMM$	0,948	0,003	0,041	0,952	0,002	0,041	0,954	0,002	0,040

Tabelle 5.3: Simulationsergebnisse für normalverteilte Merkmale für gewichtete Schätzungen.

Ein Blick auf das ungewichtete Pendant GEE_C in Tabelle 5.4 zeigt, dass hier die Schätzung im Vergleich zu den anderen ungewichteten Methoden sehr ähnliche Eigenschaften hat. Dies lässt darauf schließen, dass die in SAS implementierte Gewichtung in der Prozedur GENMOD nicht geeignet ist, um eine designgewichtete GEE-Schätzung des Zeiteffekts mit Berücksichtigung der Korrelation innerhalb der Sample Points durchzuführen.

Wie in Abschnitt 4.2 angemerkt, gehen die Gewichte in die Schätzung der Korrelationskoeffizienten α der Arbeitskorrelationsmatrix nicht mit ein. Durchschnittlich werden in den Stichproben 40 Individuen pro Cluster beobachtet. Gemäß Gleichung 4.11 beinhaltet die Schätzung der paarweisen Korrelationen die Bildung einer Summe über durchschnittlich $40(40 - 1) = 1.560$ paarweise Beobachtungen $\hat{e}_{ij}\hat{e}_{ik}$ pro Cluster. Eine Vernachlässigung der Gewichte kann hierbei große Auswirkung haben. Dies scheint gemäß der Simulationsergebnisse umso mehr der Fall zu sein, desto größer ICC_c ist.

Anders sieht es bei der gewichteten GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation ($WGEE_Z$) aus. Hier gehen jeweils nur die beiden Beobachtungen eines Individuums

in die Schätzung für α ein. Die Vernachlässigung der Gewichte hat im Vergleich zu den anderen Methoden offenbar keine bedeutende Auswirkung auf die Schätzung des Zeiteffekts.

ICC_c	ICC_z	0,001			0,02			0,06		
		Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	GLM_R	0,958	0,003	0,026	0,960	0,003	0,026	0,961	0,004	0,026
	GEE_Z	0,951	0,003	0,025	0,949	0,003	0,025	0,950	0,003	0,025
	GEE_C	0,951	0,003	0,025	0,947	0,003	0,025	0,946	0,004	0,025
	$GLMM$	0,936	0,005	0,025	0,948	0,005	0,025	0,946	0,005	0,025
0.2	GLM_R	0,967	0,003	0,028	0,971	0,003	0,028	0,967	0,003	0,028
	GEE_Z	0,953	0,003	0,026	0,948	0,003	0,026	0,948	0,003	0,026
	GEE_C	0,951	0,003	0,026	0,950	0,003	0,026	0,947	0,003	0,026
	$GLMM$	0,926	0,004	0,026	0,950	0,005	0,026	0,948	0,005	0,026
0.4	GLM_R	0,978	0,002	0,032	0,981	0,002	0,032	0,981	0,002	0,032
	GEE_Z	0,956	0,002	0,027	0,955	0,002	0,027	0,952	0,002	0,027
	GEE_C	0,944	0,002	0,027	0,947	0,002	0,027	0,947	0,002	0,027
	$GLMM$	0,934	0,005	0,027	0,958	0,004	0,026	0,958	0,004	0,026

Tabelle 5.4: Simulationsergebnisse für normalverteilte Merkmale für ungewichtete Schätzungen.

Ferner ist anhand Tabelle 5.3 zu beobachten, dass die Schätzungen der drei Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$ annähernd identische Eigenschaften aufweisen. Die relative Verzerrung liegt mit einem Wert von höchstens 0,2% für alle Kombinationen von ICC_c und ICC_z in einem vernachlässigbaren Bereich. Ebenso sind die Coverage und die durchschnittlichen Standardfehler der drei Methoden annähernd gleich. In Hinblick auf eine praktische Empfehlung einer der drei Methoden für die Anwendung im BGS liegt der Vorteil bei den beiden Methoden $WGLM_1$ und $WGLM_2$. Sie sind weniger komplex und leichter anzuwenden. Außerdem sind beide Methoden explizit als designbasierte Methoden zur Auswertung von komplexen Surveydaten in STATA beziehungsweise SAS implementiert. Dies ist für $WGEE_Z$ nicht der Fall.

Die gewichtete GLMM-Schätzung ($WGLMM$) zeigt im Vergleich zu diesen drei Methoden eine im Schnitt etwas stärkere relative Verzerrung, und in einigen Fällen eine Coverage knapp unter den geforderten 95%. Die geringere Coverage sollte hier jedoch nicht überbewertet werden. Diese könnte auch auf die kleinere Anzahl an Wiederholungen zurückzuführen sein. Auffällig sind jedoch die mittleren Standardfehler der $WGLMM$ -Schätzung, die über alle Ab-

hängigkeitsszenarien deutlich größer sind als die der drei Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$. Im Hinblick auf eine effiziente Schätzung scheint die $WGLMM$ -Schätzung eher ungeeignet.

Wie in Abschnitt 4.1.2 beschrieben, unterscheiden sich die Schätzungen $WGLM_1$ und $WGLM_2$ hinsichtlich der Schätzung der Varianz. Während in $WGLM_1$ die Varianz mit Hilfe der Gleichung 3.33 geschätzt wurde, kam in $WGLM_2$ die Gleichung 4.7 für die Varianzschätzung zur Anwendung. In $WGLM_2$ wird die Variabilität zwischen den Clustern der ersten Ziehungsstufe und die Variabilität zwischen den gezogenen Individuen innerhalb eines Clusters gemeinsam berücksichtigt. Aufgrund der in STATA notwendigen Endlichkeitskorrektur ist die Anwendung der Gleichung 4.7 für die Varianzschätzung bei einer pps -Ziehung der Cluster jedoch nicht korrekt. Im Vergleich zeigt sich, dass die Standardfehler von $WGLM_1$ und $WGLM_2$ in der Simulation im Mittel identisch sind. Die Inkaufnahme einer nicht angebrachten Endlichkeitskorrektur brachte demnach in der Simulation keine effizientere Schätzung.

Für eine designbasierte Schätzung des Zeiteffekts normalverteilter Zielmerkmale zeichnet sich in der Simulation eine Präferenz für die $WGLM_1$ -Schätzung ab.

Ein Vergleich der Simulationsergebnisse der gewichteten Schätzungen mit denen der ungewichteten Schätzung zeigt, dass die relative Verzerrung der ungewichteten Schätzungen etwas größer ausfällt.¹¹ Die relative Verzerrung ist mit höchstens 0.5% ebenfalls als kaum bedeutsam einzuschätzen.

Eine starke Verzerrung durch die Außerachtlassung des Stichprobendesigns war aufgrund der Eigenschaften der simulierten Grundgesamtheit nicht zu erwarten. Die Responsewahrscheinlichkeit hängt durch die Schichtung innerhalb der Cluster vom Merkmal y_{ijt} ab.

Eine ungewichtete Schätzung der Mittelwerte des Merkmals y_{ijt} zu den beiden Zeitpunkten t_1 und t_2 könnte negativ verzerrt sein, da Individuen mit kleinerer Merkmalsausprägung y_{ijt} in der jeweiligen Stichprobe eher überrepräsentiert werden. Eine Regression des Merkmals y_{ijt} auf die Indikatorvariable x_t beinhaltet eine Schätzung der Mittelwertsdifferenz $\bar{y}_1 - \bar{y}_2$. Da die Wiederteilnahmewahrscheinlichkeit in der simulierten Grundgesamtheit konstant ist, kann davon ausgegangen werden, dass auch die Verzerrung der beiden geschätzten Mittelwerte über die beiden Zeitpunkte konstant bleibt. Eine ungewichtete Schätzung der Mittelwertsdifferenz würde dann zu einer Eliminierung der ursprünglichen Verzerrung führen. Die Annahme, dass die Wiederteilnahme von zum ersten Zeitpunkt beobachteten Merkmalen unabhängig ist, scheint für die Auswertung des BGS eher unrealistisch.

Es fällt auf, dass die ungewichtete GLM-Schätzung mit clusterrobusten Standardfehlern GLM_R

¹¹Ausgenommen ist bei dieser Aussage der Vergleich zwischen $WGEE_C$ und GEE_C , der weiter oben bereits untersucht wurde.

mit wachsender ICC_c und ICC_z im Vergleich zu den anderen ungewichteten Methoden bezüglich der Coverage und der durchschnittlichen Standardfehler zunehmend ineffizient wird. Die drei Methoden GEE_z , GEE_c und $GLMM$ unterscheiden sich hingegen hinsichtlich ihrer Effizienz kaum. Auch im Vergleich zum gewichteten Pendant $WGLM_1$ weist die GLM_R -Schätzung mit wachsender zeitlicher Korrelation größere durchschnittliche Standardfehler auf. Auch wenn eine Gewichtung für die Schätzung des Zeiteffekts vernachlässigbar wäre, kann demnach im Fall eines einfachen Regressionsmodells eine designbasierte Schätzung kleinere Standardfehler erzielen.

In folgender Tabelle sind die Standardfehler von $WGLM_2$, $WGEE_z$, $WGLMM$ und GLM_R in Relation zum durchschnittlichen Standardfehler der $WGLM_1$ -Schätzung für die neun untersuchten Abhängigkeitsszenarien zusammenfassend abgebildet.

ICC_z	0,1			0,2			0,4		
	0,001	0,02	0,06	0,001	0,02	0,06	0,001	0,02	0,06
$WGLM_2$	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
$WGEE_z$	1,00	1,00	1,00	0,99	0,99	0,99	0,97	0,97	0,97
$WGLMM$	1,45	1,45	1,45	1,45	1,45	1,44	1,41	1,41	1,40
GLM_R	0,99	0,99	0,99	1,03	1,03	1,03	1,12	1,12	1,11

Tabelle 5.5: Standardfehler in Relation zu $WGLM_1$ für normalverteiltes Merkmal

Bei einer durchschnittlichen zeitlichen Korrelation von 0,2 verursachte die Verwendung von $WGLMM$ in Bezug auf die Größe der durchschnittlichen Standardfehler einen Effizienzverlust von 41%. Die Verwendung von GLM_R hatte bei einer zeitlichen Korrelation von 0,2 einen Effizienzverlust von 3% zu Folge. Für eine zeitliche Korrelation von 0,4 konnte ein Effizienzverlust von 12% durch die Verwendung von GLM_R anstelle von $WGLM_1$ festgestellt werden. Tabelle 5.6 beinhaltet die Simulationsergebnisse der gewichteten Schätzungen des Zeiteffekts für die stetigen Merkmale mit log-normalverteiltem Fehlerterm. Die Tabelle 5.6 zeigt die Ergebnisse der ungewichteten Schätzungen. Die Resultate sind sehr ähnlich zu den Ergebnissen der normalverteilten Merkmale.

ICC_c		0,001			0,02			0,06		
ICC_z		Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	$WGLM_1$	0,948	0,004	0,026	0,950	0,004	0,026	0,951	0,004	0,026
	$WGLM_2$	0,946	0,004	0,026	0,950	0,004	0,026	0,951	0,004	0,026
	$WGEE_Z$	0,949	0,004	0,026	0,950	0,004	0,026	0,950	0,004	0,026
	$WGEE_C$	0,942	0,002	0,026	0,924	-0,013	0,026	0,881	-0,031	0,026
	$WGLMM$	0,940	0,005	0,038	0,952	0,005	0,038	0,952	0,005	0,037
0.2	$WGLM_1$	0,943	0,004	0,027	0,943	0,004	0,027	0,945	0,004	0,027
	$WGLM_2$	0,943	0,004	0,027	0,943	0,004	0,027	0,945	0,004	0,027
	$WGEE_Z$	0,947	0,004	0,027	0,946	0,004	0,027	0,948	0,004	0,027
	$WGEE_C$	0,940	-0,000	0,027	0,885	-0,031	0,027	0,563	-0,116	0,027
	$WGLMM$	0,932	0,006	0,038	0,950	0,006	0,038	0,954	0,005	0,038
0.4	$WGLM_1$	0,939	0,003	0,029	0,942	0,003	0,029	0,943	0,003	0,029
	$WGLM_2$	0,943	0,003	0,029	0,943	0,003	0,029	0,943	0,003	0,029
	$WGEE_Z$	0,944	0,004	0,028	0,944	0,004	0,028	0,945	0,004	0,028
	$WGEE_C$	0,916	-0,020	0,029	0,309	-1,333	0,050	0,000	-2,104	0,064
	$WGLMM$	0,944	0,007	0,039	0,946	0,007	0,039	0,946	0,006	0,039

Tabelle 5.6: Simulationsergebnisse für Merkmale mit log-normalverteilterm Fehlerterm für gewichteten Schätzungen

Die $WGEE_C$ -Schätzung liefert mit steigender ICC_c erneut sehr verzerrte Schätzungen für den Zeiteffekt, wohingegen das ungewichtete Pendant GEE_C wiederum stabile Resultate liefert. Auch bei durchschnittlicher Abhängigkeit der Daten ($ICC_c = 0.02$ und $ICC_z = 0.2$) zeigt die $WGEE_C$ -Schätzung im Vergleich zu den anderen Methoden einen deutlich höheren Wert für die relative Verzerrung, gepaart mit einer Coverage von nur 88%. Für die Schätzung des Zeiteffekts scheint die gewichtete GEE-Schätzung mit Berücksichtigung der Korrelation innerhalb der Cluster analog zum normalverteilten Fall ebenfalls ungeeignet.

Die Eigenschaften der drei Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$ sind erneut annähernd identisch. Im Gegensatz zum normalverteilten Fall fällt die Coverage für die Merkmale mit log-normalverteilterm Fehlerterm etwas geringer aus. Mit zunehmender zeitlicher Korrelation sinkt sie überdies für die drei Methoden unter das geforderte Konfidenzniveau von 95%.

Verglichen mit Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$ sind die mittleren Standardfehler der $WGLMM$ -Schätzung über alle Abhängigkeitsszenarien deutlich größer. Im Hinblick auf eine effiziente Schätzung scheint die $WGLMM$ -Schätzung demnach erneut eher ungeeignet. Ein Vergleich der beiden Methoden $WGLM_1$, $WGLM_2$ weist keinen bedeutsamen Unterschied auf. Die Inkaufnahme der nicht angebrachten Endlichkeitskorrektur bringt in der Si-

mulation wiederum keinen Gewinn in Bezug auf die Effizienz der Schätzung. Hinsichtlich einer praktikablen Empfehlung für die Schätzung des Zeiteffekts unter Berücksichtigung des Stichprobendesigns fällt die Präferenz zugunsten einer weniger komplexen Schätzung erneut auf $WGLM_1$.

ICC_c		0,001			0,02			0,06		
		Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	GLM_R	0,960	0,005	0,026	0,961	0,005	0,026	0,961	0,005	0,026
	GEE_Z	0,954	0,005	0,025	0,954	0,005	0,025	0,958	0,005	0,025
	GEE_C	0,953	0,005	0,025	0,956	0,005	0,025	0,954	0,005	0,025
	$GLMM$	0,962	0,009	0,025	0,962	0,009	0,025	0,960	0,009	0,025
0.2	GLM_R	0,969	0,004	0,028	0,969	0,004	0,028	0,965	0,005	0,028
	GEE_Z	0,952	0,005	0,026	0,953	0,005	0,026	0,953	0,005	0,026
	GEE_C	0,947	0,004	0,026	0,949	0,004	0,026	0,951	0,005	0,026
	$GLMM$	0,930	0,009	0,026	0,958	0,009	0,026	0,958	0,009	0,026
0.4	GLM_R	0,979	0,003	0,032	0,977	0,003	0,032	0,977	0,004	0,032
	GEE_Z	0,953	0,005	0,027	0,951	0,005	0,027	0,950	0,005	0,027
	GEE_C	0,947	0,003	0,027	0,948	0,003	0,027	0,948	0,004	0,027
	$GLMM$	0,934	0,010	0,027	0,950	0,010	0,026	0,948	0,010	0,026

Tabelle 5.7: Simulationsergebnisse die Merkmale mit log-normalverteilter Fehlerterm für ungewichteten Schätzungen.

Die Simulationsergebnisse der ungewichteten Schätzung des Zeiteffekts zeigen erneut einen leichten Anstieg der relativen Verzerrung. Die höchste relative Verzerrung von 1% wird durch die $GLMM$ -Schätzung bei einer zeitlichen Korrelation von 0.4 verursacht. Analog zum normalverteilten Fall ist die ungewichtete GLM_R -Schätzung mit wachsender ICC_c und ICC_z im Vergleich zu den anderen ungewichteten Methoden bezüglich der Coverage und der durchschnittlichen Standardfehler zunehmend ineffizient. Auch im Vergleich zum gewichteten Pendant $WGLM_1$ weist die GLM_R -Schätzung mit wachsender zeitlicher Korrelation wieder größere durchschnittliche Standardfehler auf. Auch wenn eine Gewichtung für die Punktschätzung vernachlässigbar wäre, kann im Fall eines einfachen Regressionsmodells eine designbasierte Schätzung kleinere Standardfehler erzielen. In Tabelle 5.3 sind die Standardfehler von $WGLM_2$, $WGEE_Z$, $WGLMM$ und GLM_R in Relation zum durchschnittlichen Standardfehler der $WGLM_1$ -Schätzung abgebildet.

ICC_z	0,1			0,2			0,4			
	ICC_c	0,001	0,02	0,06	0,001	0,02	0,06	0,001	0,02	0,06
$WGLM_2$		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
$WGEE_Z$		1,00	1,00	1,00	0,99	0,99	0,99	0,97	0,97	0,97
$WGLMM$		1,42	1,42	1,42	1,42	1,41	1,41	1,37	1,37	1,37
GLM_R		0,99	0,99	0,99	1,03	1,03	1,03	1,12	1,12	1,12

Tabelle 5.8: Standardfehler in Relation zu $WGLM_1$ für die Merkmal mit log-normalverteiltem Fehlerterm

Bei einer durchschnittlichen zeitlichen Korrelation von 0,2 verursachte die Verwendung von $WGLMM$ im Vergleich zu $WGLM_1$ in Bezug auf die Größe der durchschnittlichen Standardfehler, in der Simulation einen Effizienzverlust in von 41%. Die Verwendung von GLM_R hatte bei einer zeitlichen Korrelation von 0,2 einen Effizienzverlust von 3% zu Folge. Für eine zeitliche Korrelation von 0,4 konnte ein Effizienzverlust von 12% durch die Verwendung von GLM_R festgestellt werden. Insgesamt konnte in dieser Simulation kein relevanter Einfluss der Schiefe auf die Schätzung des Zeiteffekts festgestellt werden.

Es folgt die Beschreibung der Simulationsergebnisse für die binären abhängigen Merkmale. Tabelle 5.9 enthält die Ergebnisse für die gewichtete Schätzung und in Tabelle 5.10 sind die Ergebnisse für die ungewichteten Schätzungen abgebildet. Auch hier zeigen sich im Vergleich zur Untersuchung der beiden stetigen Merkmale sehr ähnliche Resultate.

Die $WGEE_C$ -Schätzung und die gewichtete alternierende logistische Regression $WARL$ liefern mit steigender ICC_c ebenfalls sehr verzerrte Schätzungen. Auch bei durchschnittlicher Abhängigkeit der Daten ($ICC_c = 0,02$ und $ICC_z = 0,2$) liefert die $WGEE_C$ -Schätzung in der Simulation, im Vergleich zu den anderen Methoden einen deutlich höheren Wert für die relative Verzerrung (-11%), gepaart mit einer Coverage von knapp 92%. Für die $WARL$ -Schätzung zeigt sich bei mittlerer Abhängigkeit eine relative Verzerrung von knapp -14% und eine Coverage von nur 90%. Im Vergleich dazu liegt die relative Verzerrung der anderen drei Schätzmethoden über alle Abhängigkeitsszenarien hinweg bei höchstens -2,6%.

Die ungewichteten Pendanten GEE_c und ARL liefern wiederum Schätzungen mit akzeptablen Eigenschaften. Für die Schätzung des Zeiteffekts scheinen die gewichtete GEE-Schätzung mit Berücksichtigung der Korrelation innerhalb der Cluster und die gewichtete alternierende Regression ungeeignet zu sein. Die Eigenschaften der drei Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$ sind erneut annähernd identisch. Die Coverage unterschreitet jedoch in allen Kombinationen der Korrelation leicht das geforderte Niveau von 95%.

Ein Vergleich der beiden Methoden WGM_1 und $WGLM_2$ zeigt keinen erheblichen Unterschied auf. Die Verwendung der nicht angebrachten Endlichkeitskorrektur bringt in der Simulation wiederum keinen Gewinn in Bezug auf die Effizienz der Schätzung. Hinsichtlich einer praktikablen Empfehlung für die Schätzung des Zeiteffekts unter Berücksichtigung des Stichprobendesigns, liegt die Präferenz zugunsten einer weniger komplexen und theoretisch fundierten Schätzung abermals bei $WGLM_1$.

		0,001			0,02			0,06		
ICC_c	ICC_z	Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	$WGLM_1$	0,946	-0,026	0,051	0,939	-0,024	0,052	0,944	-0,017	0,053
	$WGLM_2$	0,946	-0,026	0,051	0,939	-0,024	0,051	0,945	-0,017	0,053
	$WGEE_Z$	0,939	-0,025	0,051	0,940	-0,023	0,052	0,947	-0,016	0,053
	$WGEE_c$	0,946	-0,029	0,051	0,932	-0,069	0,052	0,929	-0,106	0,053
	$WALR$	0,942	-0,037	0,051	0,932	-0,080	0,052	0,922	-0,118	0,053
0.2	$WGLM_1$	0,945	-0,019	0,049	0,953	-0,024	0,050	0,941	-0,008	0,051
	$WGLM_2$	0,945	-0,019	0,049	0,954	-0,024	0,050	0,939	-0,008	0,051
	$WGEE_Z$	0,949	-0,018	0,049	0,945	-0,024	0,050	0,941	-0,008	0,051
	$WGEE_c$	0,944	-0,030	0,049	0,917	-0,109	0,050	0,853	-0,222	0,053
	$WALR$	0,947	-0,054	0,049	0,900	-0,139	0,050	0,827	-0,248	0,053
0.4	$WGLM_1$	0,944	-0,021	0,047	0,947	-0,013	0,047	0,941	-0,018	0,048
	$WGLM_2$	0,945	-0,021	0,047	0,945	-0,013	0,047	0,942	-0,018	0,047
	$WGEE_Z$	0,949	-0,021	0,046	0,947	-0,012	0,046	0,938	-0,019	0,047
	$WGEE_c$	0,941	-0,053	0,046	0,858	-0,218	0,048	0,517	-0,623	0,064
	$WALR$	0,855	-0,204	0,046	0,667	-0,450	0,051	0,277	-0,833	0,058

Tabelle 5.9: Simulationsergebnisse für binäre Merkmale für gewichtete Schätzungen.

Die Simulationsergebnisse der ungewichteten Schätzung des Zeiteffekts zeigen im Gegensatz zu den beiden Simulationsergebnissen der metrischen abhängigen Variablen, eine leichte Verringerung der relativen Verzerrung und eine Erhöhung der Coverage.

		0,001			0,02			0,06		
ICC_z	ICC_c	Cov	RV	SD	Cov	RV	SD	Cov	RV	SD
0.1	GLM_R	0,955	-0,015	0,050	0,946	-0,014	0,051	0,954	-0,010	0,052
	GEE_Z	0,949	-0,014	0,049	0,943	-0,014	0,049	0,946	-0,009	0,051
	GEE_c	0,951	-0,015	0,049	0,942	-0,014	0,049	0,950	-0,009	0,050
	ALR	0,948	-0,015	0,049	0,941	-0,014	0,049	0,950	-0,009	0,050
0.2	GLM_R	0,965	-0,007	0,050	0,955	-0,015	0,050	0,947	0,001	0,051
	GEE_Z	0,957	-0,006	0,047	0,942	-0,014	0,048	0,932	0,002	0,049
	GEE_c	0,954	-0,007	0,047	0,940	-0,015	0,048	0,938	0,002	0,048
	ALR	0,953	-0,006	0,047	0,942	-0,014	0,048	0,936	0,002	0,048
0.4	GLM_R	0,972	-0,004	0,050	0,970	0,004	0,050	0,963	-0,005	0,051
	GEE_Z	0,954	-0,004	0,044	0,945	0,004	0,045	0,940	-0,007	0,045
	GEE_c	0,947	-0,004	0,045	0,944	0,004	0,045	0,937	-0,005	0,046
	ALR	0,950	-0,004	0,044	0,942	0,004	0,044	0,936	-0,007	0,045

Tabelle 5.10: Simulationsergebnisse für binäre Merkmale für ungewichtete Schätzungen.

Analog zu den stetigen Variablen zeigt sich, dass die ungewichtete GLM Schätzung mit clusterrobusten Standardfehlern GLM_R mit wachsender ICC_c und ICC_z im Vergleich zu den anderen Methoden bezüglich der Coverage und der durchschnittlichen Standardfehler zunehmend ineffizient wird. Im Vergleich zum gewichteten Pendant $WGLM_1$ weist die GLM_R -Schätzung mit wachsender zeitlicher Korrelation wieder größere durchschnittliche Standardfehler auf.

Auch im Fall eines logistischen Regressionsmodells kann eine designbasierte Schätzung des Zeiteffekts kleinere Standardfehler erzielen, auch wenn die Gewichtung für die Punktschätzung vernachlässigbar ist.

In Tabelle 5.11 sind die durchschnittlichen Standardfehler von $WGLM_2$, $WGEE_Z$ und GLM_R in Relation zum durchschnittlichen Standardfehler der $WGLM_1$ -Schätzung abgebildet.

ICC_z	0,1			0,2			0,4		
	0,001	0,02	0,06	0,001	0,02	0,06	0,001	0,02	0,06
$WGLM_2$	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
$WGEE_Z$	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,99
GLM_R	0,99	0,98	0,98	1,01	1,01	1,01	1,08	1,07	1,07

Tabelle 5.11: Standardfehler in Relation zu $WGLM_1$ für binäre Merkmale

Bei einer durchschnittlichen zeitlichen Korrelation von 0,2 verursachte die Verwendung von GLM_R im Vergleich zu $WGLM_1$ in Bezug auf die Größe der durchschnittlichen Standardfehler des Zeiteffekts einen Effizienzverlust von 1%. Für eine zeitliche Korrelation von 0,4 konnte ein Effizienzverlust von 7% durch die Verwendung von GLM_R festgestellt werden. Dieser Effizienzverlust für binäre Merkmale fiel in der Simulation etwas niedriger aus als für die stetigen Merkmale.

In den Simulationen zeigte sich für die binären und stetigen abhängigen Merkmale über die neun untersuchten Abhängigkeitsszenarien, im Wesentlichen ein sehr ähnliches Bild. Die gewichtete GEE-Schätzung mit Berücksichtigung von ICC_c zeigte eine vergleichsweise starke Verzerrung bei wachsender ICC_c und ICC_z und scheint somit ungeeignet für die Schätzung des Zeiteffekts. Die gewichteten GLMM-Schätzung $WGLMM$ für die stetigen Merkmale, erwies sich aufgrund durchschnittlich größerer Standardfehler als ineffizient.

Zwischen den drei Methoden $WGLM_1$, $WGLM_2$ und $WGEE_Z$ zeigten sich keine bedeutsamen Unterschiede bezüglich ihrer Effizienz und relativen Verzerrung. Basierend auf diesen Ergebnissen kann für die Schätzung des Zeiteffekts die designbasierte GLM-Schätzung $WGLM_1$ für die praktische Anwendung empfohlen werden.

Die Resultate gelten für die generierte endliche Grundgesamtheit und das beschriebene Stichprobendesign der Simulationsstudie. Die Eigenschaften der generierten Grundgesamtheit und die zweistufige Stichprobenziehung wurden der Zielpopulation und dem Stichprobendesign des BGS in vereinfachter Form nachempfunden. Im nächsten Abschnitt werden die beschriebenen Methoden auf einige Testvariablen aus dem BGS für die Schätzung des Zeiteffekts angewandt. Hier wird sich zeigen, ob sich die Resultate der Simulation auch in der Auswertung der BGS-Daten widerspiegeln.

5.4 Beispieldaten aus dem BGS

Im Zuge dieser Arbeit wurden vom RKI Daten der beiden Erhebungszeiträume bereitgestellt. Der kombinierte Datensatz enthält die Beobachtungen aller bisher untersuchten Probanden aus dem BGS98 und der Folgerhebung DEGS1. Die Erhebung der DEGS1-Daten ist noch nicht abgeschlossen. Eine vorläufige Beurteilung der untersuchten Methoden zur Schätzung des Zeiteffekts sollte jedoch auch mit dem unvollständigen Datensatz möglich sein. Die metrischen und binären Testvariablen wurden in anonymisierter Form bereitgestellt.

Neben der Variable Zeit, die hier von primärem Interesse ist, wurden zusätzlich die Kontrollvariablen Alter und Geschlecht in die Modellgleichung integriert. Die Variable Alter wurde hierbei in 10-Jahres-Gruppen unterteilt und als kategoriale Variable in das Modell aufgenom-

men. Tabelle 5.12 zeigt die geschätzten Zeiteffekte und ihre Standardfehler für fünf stetige Merkmale.¹²

Variable	Korrelation ^a		Schätzmethode				
	Zeit	Cluster	GLM_R	$WGLM_1$	$WGLM_2$	$WGEE_Z$	$WGEE_C$
1	0,193	0,010	-2,748** (0,022)	-2,731** (0,030)	-2,731** (0,030)	-2,727** (0,022)	-2,727** (0,029)
2	0,339	0,028	-1,017** (0,020)	-0,959** (0,031)	-0,959** (0,031)	-0,962** (0,020)	-0,967** (0,031)
4	0,113	0,001	-0,060** (0,021)	-0,052* (0,023)	-0,052* (0,023)	-0,050* (0,021)	-0,052* (0,023)
6	0,872	0,022	0,141** (0,012)	0,116** (0,013)	0,116** (0,013)	0,122** (0,007)	0,123** (0,013)
10	0,385	0,012	0,390** (0,018)	0,383** (0,022)	0,383** (0,022)	0,380** (0,018)	0,384** (0,021)

^aArbeitskorrelation aus GEE-Schätzung
* p<0,05, ** p<0,01 (Standardfehler in Klammern)

Tabelle 5.12: Geschätzte Zeiteffekte und Standardfehler für stetige Merkmale aus dem BGS

In der Anwendung zeigte sich die GLMM-Schätzung als nicht praktikabel, da die Schätzung eine sehr lange Rechenzeit in Anspruch nimmt und die zudem Log-Likelihood nicht konvergierte. Aus diesem Grund werden die gemischten Modelle für die Schätzung des Zeiteffekts im BGS nicht weiter betrachtet.

Die beiden Methoden $WGLM_1$ und $WGLM_2$ liefern identische Ergebnisse für die Standardfehler und bekräftigen die in der Simulation festgestellt Äquivalenz beider Methoden zur Varianzschätzung. Im Vergleich zur ungewichteten Schätzung GLM_R zeigen sich erhebliche Unterschiede im geschätzten Zeiteffekt. Eine Missachtung des Stichprobendesigns scheint nicht empfehlenswert. Die modellbasierten Parameterschätzungen können verzerrt sein, wenn das Stichprobendesign vernachlässigt wird (Pfeffermann et al.,1998; Pfeffermann und Sverchkov, 2009).

Die Ergebnisse der gewichteten GEE-Schätzung mit Berücksichtigung der Korrelation innerhalb der Sample Points $WGEE_c$ weisen annähernd identische Standardfehler im Vergleich zur $WGLM$ -Methode auf. Die Punktschätzungen für den Zeiteffekt unterscheiden sich leicht. Die

¹²Die Nummern in der Spalte „Variable“ beziehen sich auf die Bezeichnung der Variablen im bereitgestellten Datensatz. Die Nummern wurden beibehalten um eine doppelte Verblindung der Variablen zu vermeiden.

Standardfehler der gewichteten GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation $WGEE_Z$ vielen im Vergleich zur gewichteten GLM-Schätzung etwas kleiner aus.

Bei sehr hoher zeitlicher Korrelation scheint dieser Effekt zuzunehmen. Für Variable 6 wurde eine zeitliche Korrelation von 0,872 geschätzt. Hier wurde durch die Verwendung von $WGEE_Z$ der Standardfehler im Vergleich zu den anderen gewichteten Methoden nahezu halbiert. Die kleineren Varianzen hatten jedoch keinen Einfluss auf die Testentscheidungen. Die geschätzten Zeiteffekte der $WGEE_Z$ -Schätzungen unterscheiden sich ebenfalls etwas von den $WGLM$ -Schätzungen. Die Abweichung der Punktschätzungen der beiden gewichteten GEE-Schätzungen von den gewichteten GLM-Schätzungen können als unerheblich eingeschätzt werden.

In Tabelle 5.13 sind die geschätzten Zeiteffekte und ihre Standardfehler für fünf binäre Merkmale abgebildet.

Variable	Korrelation ^a		Schätzmethode					
	Zeit	Cluster	GLM_R	$WGLM_1$	$WGLM_2$	$WGEE_Z$	$WGEE_C$	$WARL$
1	0,109	0,005	-0,257** (0,058)	-0,212** (0,071)	-0,212** (0,071)	-0,205** (0,064)	-0,209** (0,072)	-0,202** (0,072)
2	0,149	0,003	0,275** (0,038)	0,285** (0,047)	0,285** (0,047)	0,285** (0,042)	0,284** (0,047)	0,284** (0,046)
3	0,427	0,009	0,028 (0,057)	0,099 (0,070)	0,099 (0,070)	0,118* (0,054)	0,105 (0,070)	0,139* (0,065)
4	0,509	0,023	-0,191** (0,044)	-0,180** (0,060)	-0,180** (0,060)	-0,173** (0,043)	-0,171** (0,060)	-0,165** (0,054)
5	0,576	0,013	-0,108** (0,041)	-0,082 (0,044)	-0,082 (0,044)	-0,033 (0,037)	-0,085 (0,043)	-0,031 (0,038)

^aArbeitskorrelation aus GEE-Schätzung
* p<0,05, ** p<0,01 (Standardfehler in Klammern)

Tabelle 5.13: Geschätzte Zeiteffekte und Standardfehler für binäre Merkmale aus dem BGS

Auch für die binären Merkmale kann festgestellt werden, dass sich die geschätzten Zeiteffekte der gewichteten GLM-Schätzung im Vergleich zur ungewichteten Schätzung GLM_R unterscheiden.

Die gewichtete GEE-Schätzung mit Berücksichtigung der Korrelation innerhalb der Sample Points $WGEE_C$ weisen auch hier annähernd identische Standardfehler im Vergleich zur $WGLM$ -Schätzung auf. Die Punktschätzungen für den Zeiteffekt unterscheiden sich wieder-

um etwas.

Die Standardfehler der gewichteten GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation $WGEE_z$ fallen im Vergleich zur gewichteten GLM-Schätzung etwas kleiner aus. Bei sehr hoher zeitlicher Korrelation scheint dieser Effekt zuzunehmen. Für die Testvariable 4 wurde eine zeitliche Korrelation von 0,5 geschätzt. Hier wurde durch die Verwendung von $WGEE_z$ der Standardfehler im Vergleich zur gewichteten GLM-Schätzung um knapp 40% vermindert. Auf die Testentscheidung hat der kleinere Standardfehler in diesem Fall keine Auswirkung. Die $WGEE_c$ -Schätzung führt für Variable 3 zu einem signifikanten Ergebnis (bei einem Signifikanzniveau von 5%), wobei mit den anderen Methoden der Zeiteffekt als nicht signifikant geschätzt wird. Dies ist auf den größer geschätzten Zeiteffekt in Kombination mit einem kleineren Standardfehler zurückzuführen.

Die Standardfehler der gewichteten alternierenden logistische Regression liegen außer für Testvariable 1 stets zwischen denen der $WGEE_c$ - und $WGEE_z$ -Schätzung. Auch für die Methoden $WARL$ unterscheiden sich die Punktschätzungen für den Zeiteffekt etwas von der gewichteten GLM-Schätzung.

Der größte Unterschied in der Punktschätzung zwischen den gewichteten GEE-Methoden und der gewichteten GLM-Schätzung ist mit einer Differenz von 0,05 für Testvariable 5 bei der $WARL$ -Schätzung zu verzeichnen. Für die Schätzung des Odds-Ratios beinhaltet dies einen Faktor von $\exp(0,05) = 1,05$, um den sich die Odds-Ratios der GLM-Schätzung und der GEE-Schätzung unterscheiden. Der Unterschied in den geschätzten Koeffizienten ist für die untersuchten Testvariablen in Bezug auf die Interpretation demnach nicht erheblich.

Die Untersuchung der binären und stetigen Testvariablen ergibt, dass die designbasierte GLM-Schätzung und die gewichteten GEE-Schätzungen vergleichbare Punktschätzungen für den Zeiteffekt liefern. Die gewichtete GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation liefert für die untersuchten Testvariablen kleinere Standardfehler. Es wurde gezeigt, dass dies zu unterschiedlichen Testentscheidungen führen kann. Effekte, die in den anderen Methoden als nicht signifikant beurteilt werden, könnten mit $WGEE_z$ signifikant sein.

Es ist jedoch fraglich, ob einem solchen Ergebnis in der Praxis Vertrauen entgegengebracht würde. Basierend auf den Erkenntnissen der Simulationsstudie und der hier angestellten Beispieluntersuchungen sollte ein Anwender die designbasierten GLM-Schätzung für die Schätzung des Zeiteffekts im BGS zu nutzen. Dies ist in STATA und SAS implementiert und einfach umzusetzen.

6 Abschließende Bemerkungen

Untersuchungsgegenstand dieser Arbeit war der Vergleich statistischer Methoden für die Analyse des Zeiteffekts im BGS unter Berücksichtigung des komplexen Stichprobendesigns. Die Untersuchung zielte darauf ab praktikable Methoden für die Schätzung des Zeiteffekts zu finden. Komplexe Stichproben können auf Basis modellbasierter und designbasierter Verfahren analysiert werden. Beide Ansätze wurden in dieser Arbeit vorgestellt. Für die designbasierte Analyse wurde zusätzlich der PML-Ansatz eingeführt, der über die Totalwertschätzung einer endlichen Population motiviert wurde. Es wurden die drei Methoden GLM, GEE und GLMM zur designbasierten Schätzung des Zeiteffekts vorgestellt. Die Umsetzbarkeit der vorgestellten Methoden wurde für die Statistik-Software SAS und STATA untersucht.

In einer Simulation wurden die drei Methoden hinsichtlich Verzerrung und Effizienz für stetige und binäre abhängige Merkmale untersucht. Die gewichtete GEE-Schätzung mit Berücksichtigung der Intraclusterkorrelation lieferte hierbei sehr verzerrte Schätzungen. Die gewichtete GLMM-Schätzung erwies sich im Vergleich als ineffizient. Die gewichtete GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation und die gewichtete GLM-Schätzung lieferten annähernd identische Ergebnisse. Basierend auf dieser Simulation wird die designbasierte GLM-Schätzung für die praktische Anwendung empfohlen. Diese ist einfach in der Anwendung und in SAS und STATA für die Auswertung von Surveydaten implementiert.

Bei einer Untersuchung von Beispieldaten aus dem BGS stellte sich die gewichtete GEE-Schätzung mit Berücksichtigung der zeitlichen Korrelation innerhalb der Längsschnittprobanden als effizienter heraus. Für alle Testvariablen hatte der geschätzte Zeiteffekt einen kleineren Standardfehler. Für eine Testvariable führte dies zu einer gegenläufigen Testentscheidung. Es zeigte sich ein signifikanter Zeiteffekt, der durch die anderen Methoden nicht festgestellt werden konnte. Es ist jedoch fraglich, ob einem solchen Ergebnis in der Praxis Vertrauen entgegengebracht würde. Auch die Erkenntnisse aus der Beispieluntersuchung führen zu der Empfehlung die designbasierte GLM-Schätzung für die Schätzung des Zeiteffekts im BGS zu nutzen.

Literaturverzeichnis

- Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods* 35(3), 439–460.
- Berger, Y. G. und Y. Tillé (2009). Sampling with unequal probabilities. In C. Rao (Hrsg.), *Sample Surveys: Design, Methods and Applications*, Volume 29A of *Handbook of Statistics*, Kapitel 2, S. 39 – 54. Elsevier.
- Binder, D. A. und Z. Patak (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association* 89(427), 1035–1043.
- Binder, D. A. und G. Roberts (2009). Design- and model-based inference for model parameters. In C. Rao (Hrsg.), *Sample Surveys: Inference and Analysis*, Volume 29B of *Handbook of Statistics*, Kapitel 24, S. 33 – 54. Elsevier.
- Binder, D. A. und G. R. Roberts (2004). *Design-based and Model-based Methods for Estimating Model Parameters*, Kapitel 3, S. 29–48. *Analysis of Survey Data*. Wiley & Sons, Inc.
- Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology* 9(1), 49.
- Carlin, J. B., R. Wolfe, C. Coffey, und G. C. Patton (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort. *Statistics in Medicine* 18(19), 2655–2679.
- Carrillo, I. A., J. Chen, und C. Wu (2010). The pseudo-gee approach to the analysis of longitudinal surveys. *Canadian Journal of Statistics* 38(4), 540–554.
- Cox, L. H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association* 82(398), 520–524.

- Diggle, P., K. Liang, und S. Zeger (1994). *Analysis of longitudinal data* (corr. print ed.). Number 13 in Oxford statistical science series. Oxford [u.a.]: Clarendon Press.
- Fahrmeir, L., T. Kniep, und S. Lang (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer.
- Godambe, V. und M. E. Thompson (2009). Estimating functions and survey sampling. In C. Rao (Hrsg.), *Sample Surveys: Inference and Analysis*, Volume 29B of *Handbook of Statistics*, Kapitel 26, S. 83 – 101. Elsevier.
- Godambe, V. P. und M. E. Thompson (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review / Revue Internationale de Statistique* 54(2), 127–138.
- Gould, W., J. Pitblado, und W. Sribney (2006, January). *Maximum Likelihood Estimation with Stata, Third Edition* (3 ed.). Stata Press.
- Greene, W. (2008). *Econometric analysis* (6 ed.). Upper Saddle River, NJ [u.a.]: Pearson Prentice Hall.
- Heeringa, S., B. T. West, und P. A. Berglund (2010). *Applied survey data analysis*. Chapman & Hall/CRC.
- Horvitz, D. G. und D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260), 663–685.
- Korn, E. L. und B. I. Graubard (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158(2), 263–295.
- Kott, P. S. (1991). A model-based look at linear regression with survey data. *The American Statistician* 45, 107–112.
- Lehtonen, R. und E. Pahkinen (2004). *Practical Methods for Design and Analysis of Complex Surveys* (2 ed.). Wiley & Sons, Inc.
- Liang, K.-Y. und S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.
- Molenberghs, G. und G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer.

- Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 1(3), 249–273.
- Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, und J. Rasbash (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(1), 23–40.
- Pfeffermann, D. und M. Sverchkov (2009). Inference under informative sampling. In C. Rao (Hrsg.), *Sample Surveys: Inference and Analysis*, Volume 29B of *Handbook of Statistics*, Kapitel 39, S. 455 – 487. Elsevier.
- Potthoff, P., E. Schroeder, U. Reis, und A. Klamert (1999). Ablauf und Ergebnisse der Feldarbeit beim Bundes-Gesundheitssurvey. *Gesundheitswesen Sonderheft* 2, 62–67.
- Preisser, J. S., K. K. Lohmann, und P. J. Rathouz (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine* 21, 3035–3054.
- Prášková, Z. und P. K. Sen (2009). Asymptotics in finite population sampling. In C. Rao (Hrsg.), *Sample Surveys: Inference and Analysis*, Volume 29B of *Handbook of Statistics*, Kapitel 40, S. 489 – 522. Elsevier.
- Rabe-Hesketh, S. und A. Skrondal (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(4), 805–827.
- Rabe-Hesketh, S., A. Skrondal, und A. Pickles (2004, October). Gllamm manual. U.C. Berkeley Division of Biostatistics Working Paper Series 1160, Berkeley Electronic Press.
- RKI (2009). *DEGS - Studie zur Gesundheit Erwachsener in Deutschland: Projektbeschreibung*. Beiträge zur Gesundheitsberichterstattung des Bundes. Robert Koch-Institut.
- RKI (2012). Studienorte. http://www.rki.de/cln_153/nn_1335824/DE/Content/GBE/Erhebungen/Gesundheits surveys/Degs/downloads__links.html, 06.04.2012.
- Rust, K. F. und J. N. K. Rao (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* 5(3), 283–310.
- Särndal, C., B. Swensson, und J. Wretman (2003). *Model assisted survey sampling*. Springer series in statistics. Springer.
- SAS-InstituteInc (2004). *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

- Skinner, C. (1989). *Domain Means, Regression and Multivariate Analysis*, Kapitel 3, S. 59–88. *Analysis of Complex Surveys*. Wiley & Sons, Inc.
- StataCorp. (2009). *Stata Survey Data Reference Manual: Release 11*. Statistical Software. College Station, TX: StataCorp LP.
- StataCorp. (2011). *Stata Longitudinal-Data/Panel-Data Reference Manual Release 12*. Statacorp Lp.
- Stolzenberg, H. (2000). *Bundes-Gesundheitsurvey 1998: Dokumentation des Datensatzes*. Robert Koch Institut.
- Thefeld, W., H. Stolzenberg, und B. M. Bellach (1999). Bundes-Gesundheitssurvey: Response, Zusammensetzung der Teilnehmer und Non-Responder-Analyse. *Gesundheitswesen Sonderheft 2*, 57–61.
- Wolter, K. (2007). *Introduction to variance estimation*. Statistics for social and behavioral sciences. Springer.
- Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science* 45(2), pp. 470–490.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Abschlussarbeit selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe.

Ich erkläre ausdrücklich, dass ich sämtliche in der Arbeit verwendeten fremden Quellen, auch aus dem Internet, als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken u. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken u. Ä. anderer Autorinnen und Autoren (indirektes Zitieren) die Quelle angegeben habe.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend der Prüfungsordnung und/oder der Allgemeinen Satzung für Studien- und Prüfungsangelegenheiten der HU (ASSP) geahndet werden. Die Arbeit wurde in gleicher oder ähnlicher Form bisher bei keiner anderen Institution eingereicht.

Berlin,