

Aus dem Institut für Medizinische Biometrie und Statistik
der Universität zu Lübeck
Direktor: Univ.-Prof. Dr. rer. nat. Andreas Ziegler

**Optimierung der Vorhersage durch verallgemeinerte
Schätzgleichungen: Wahl der Arbeitskorrelationsstruktur,
neue Deletionsdiagnostiken und die Verbindung zu
fraktionellen Polynomen**

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

– **Aus der Sektion Medizin** –

vorgelegt von
Maren Andrea Vens
aus Rhede i.W.

Lübeck, 2013

1. Berichterstatter: Prof. Dr. rer. nat. Andreas Ziegler

2. Berichterstatterin: Prof. Dr. med. Roland Linder

Tag der mündlichen Prüfung: 26.6.2014

Zum Druck genehmigt. Lübeck, den 26.6.2014

Promotionskommission der Sektion Medizin

Inhaltsverzeichnis

Abbildungsverzeichnis	iv
Tabellenverzeichnis	vi
1. Einleitung, Fragestellung und Aufbau der Arbeit	1
2. Verallgemeinerte Schätzgleichungen	5
3. Wahl der Arbeitskorrelationsmatrix	11
3.1. Allgemeine Kriterien zur Wahl der Arbeitskorrelationsmatrix	12
3.2. Statistische Kriterien zur Wahl der Arbeitskorrelationsmatrix	17
3.3. Modellierung der Arbeitskorrelation: Gibt es einen Weg aus dem Dilemma?	19
4. Verallgemeinerte Schätzgleichungen und Regressionsdiagnostik in longitudinalen klinischen Studien	23
4.1. Die SB-LOT-Studie	25
4.2. Verallgemeinerte Schätzgleichungen	27
4.3. Wahl einer sinnvollen Arbeitskorrelationsstruktur	28
4.4. Regressionsdiagnostik	28
4.4.1. Residuen	29
4.4.2. Einflussreiche Punkte und Hebelpunkte	31
4.5. Ergebnisse	33
4.5.1. Die Standard GEE Analyse	33
4.5.2. Regressionsdiagnostik	35
4.6. Diskussion	39
5. Multivariable fraktionelle Polynome für korrelierte abhängige Variablen unter Verwendung von verallgemeinerten Schätzgleichungen	42
5.1. Fraktionelle Polynome	44
5.2. Die GEE-MFP-Methode	45
5.3. Kriterien zur Güte der Anpassung und Selektion	49
5.4. Die Daten der Framingham Heart Study	49

5.5. Zehnfach Kreuzvalidierung	51
5.6. Ergebnisse	51
5.6.1. Die vollständigen Modelle unter Verwendung des kompletten Datensatzes	52
5.6.2. Die Modelle nach der schrittweisen Prozedur unter Verwendung des kompletten Datensatzes	53
5.6.3. Die zehnfach Kreuzvalidierung	57
5.7. Diskussion	58
6. Diskussion und Ausblick	60
7. Zusammenfassung	68
A. Ergebnisse des GEE-MFP-Algorithmus	71
A.1. Ergebnisse unter Verwendung des vollständigen Datensatzes	72
A.2. Ergebnisse der zehnfach Kreuzvalidierung	74
A.3. Abbildungen zum Vergleich der Modelle unter Verwendung der verschiedenen Gütekriterien	98
Literaturverzeichnis	104
Danksagung	110
Lebenslauf	112
Eigene Publikationen	114
Originalarbeiten	115
Buchbeiträge	116
Vorträge	116
Poster	117

Abbildungsverzeichnis

3.1. Obere und untere Grenzen des Korrelationskoeffizienten einer (2×2) -Tabelle für unterschiedliche marginale Häufigkeiten μ_1 und μ_2	16
4.1. Erwarteter Verlauf der SB-LOT-Studie.	25
4.2. Residuenabbildungen.	36
4.3. Cluster-Cook-Statistik für die SB-LOT-Studie unter Verwendung der autoregressiven Arbeitskorrelationsstruktur erster Ordnung.	37
4.4. Halbnormalabbildungen mit simulierten Einhüllenden.	38
4.5. Halbnormalabbildungen mit simulierten Einhüllenden für Patienten mit einer Cluster-Cook-Statistik $< 0,01$ und zusätzlichem Ausschluss von Patient a) 181, b) 38, c) 41.	38
4.6. Halbnormalabbildung mit simulierten Einhüllenden und Cook-Distanzen nach dem Ausschluss von Patienten mit einer Cluster-Cook-Statistik $> 0,01$ und zusätzlichem Ausschluss der drei Patienten 181, 38 und 41.	39
5.1. Darstellung von verschiedenen Kurvenverläufen fraktioneller Polynome zweiter Ordnung.	45
5.2. Modellbildungsprozess der GEE-MFP-Methode.	46
5.3. Schrittweise Prozedur der GEE-MFP-Methode.	48
5.4. Vergleich der Anpassungen an das Alter unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	55
A.1. Vergleich der Anpassungen an den Körpermasseindex unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	98
A.2. Vergleich der Anpassungen an die Nüchtern-Serum-HDL-Cholesterinkonzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	99

A.3. Vergleich der Anpassungen an die Familiengröße unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	100
A.4. Vergleich der Anpassungen an die Nüchtern-Serum-Triglycerid-Konzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	101
A.5. Vergleich der Anpassungen an das Geschlecht unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	102
A.6. Vergleich der Anpassungen an die Nüchtern-Serum-Glukose-Konzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur.	103

Tabellenverzeichnis

3.1. (2×2) -Häufigkeitstabelle mit Elementen $\mu_{tt'} > 0$ für $t, t' = 1, 2$	15
4.1. Die SB-LOT-Studie: Unterschenkelvolumina (in ml) im Verlauf der Studie.	26
4.2. Schätzer der Korrelationskoeffizienten der SB-LOT-Daten, zusammengefasst über beide Behandlungsgruppen.	26
4.3. Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit der autoregressiven Arbeitskorrelationsmatrix erster Ordnung für SB-LOT gegen Placebo.	33
4.4. Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit verschiedenen Arbeitskorrelationsmatrizen für SB-LOT gegen Placebo.	34
4.5. Schätzer für die Korrelationskoeffizienten für die SB-LOT-Daten unter Verwendung der unstrukturierten Arbeitskorrelationsmatrix.	34
4.6. Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit autoregressiver Arbeitskorrelationsmatrix erster Ordnung für SB-LOT gegen Placebo unter Ausschluss von Patienten anhand ihrer Cluster-Cook-Statistik.	36
5.1. Die gewählten erklärenden Variablen und die zugehörigen Skalierungen.	50
5.2. Ergebnisse der vollständigen Modelle unter Verwendung von QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. . . .	52
5.3. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des kompletten Datensatzes und QIC oder QIC_u als Kriterien der Güte der Anpassung und der Selektion.	53
5.4. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des vollständigen Datensatzes und $BQIC_u$ als Kriterium der Güte der Anpassung und der Selektion.	54
A.1. Ergebnisse der vollständigen Modelle unter Verwendung des vollständigen Datensatzes und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	72

A.2. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des vollständigen Datensatzes und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	73
A.3. Gewählte Transformationen in den vollständigen Modellen unter Verwendung des QIC, QIC_u und $BQIC_u$ in jedem der zehn Kreuzvalidierungsdurchläufe für jede Einflussvariable.	75
A.4. Gewählte Transformationen in den Modellen nach der schrittweisen Prozedur unter Verwendung des QIC, QIC_u und $BQIC_u$ in jedem der zehn Kreuzvalidierungsdurchläufe für jede Einflussvariable. Keine Angabe bedeutet, dass die Variable durch die Anwendung des schrittweisen Verfahrens eliminiert wurde.	76
A.5. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–9 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	77
A.6. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–8 und 10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	78
A.7. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–7 und 9–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	79
A.8. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–6 und 8–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	80
A.9. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–5 und 7–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	81
A.10. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–4 und 6–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	82
A.11. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–3 und 5–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	83
A.12. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–2 und 4–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	84
A.13. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1 und 3–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	85

A.14. Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 2–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	86
A.15. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–9 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	87
A.16. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–8 und 10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	88
A.17. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–7 und 9–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	89
A.18. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–6 und 8–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	90
A.19. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–5 und 7–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	91
A.20. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–4 und 6–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	92
A.21. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–3 und 5–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	93
A.22. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–2 und 4–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	94
A.23. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1 und 3–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	95
A.24. Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 2–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion.	96
A.25. Wahrscheinlichkeiten korrekt vorhergesagter Ergebnisse in der Zielvariablen in den vollständigen Modellen und den Modellen nach der schrittweisen Prozedur in der zehnfach Kreuzvalidierung unter Verwendung von QIC, QIC_u und $BQIC_u$ als Selektions- und Gütekriterien.	97

1. Einleitung, Fragestellung und Aufbau der Arbeit

In zahlreichen biometrischen oder epidemiologischen, allerdings auch bei klinischen, Fragestellungen sind die klassischen Annahmen der Statistik, dies sind insbesondere die Normalverteilung der Zielvariablen und die Unabhängigkeit der Beobachtungen, verletzt.

Ersteres ist der Fall, wenn Zähldaten (z. B. Anzahl epileptischer Anfälle) oder binäre Daten (z. B., ob eine Erkrankung vorliegt oder nicht) als Zielgrößen verwendet werden. Verallgemeinerte lineare Modelle (engl. generalized linear models, GLM) sind ein seit Jahrzehnten genutzter Ansatz zur Regressionsanalyse solcher Zielvariablen (Nelder und Wedderburn, 1972). Bei den GLM spezifiziert der Anwender eine sogenannte Linkfunktion, um die Prädiktoren mit der Zielgröße in Verbindung zu bringen. Zum Beispiel ist eine Wahl für die Modellierung dichotomer Zielgrößen die Logitfunktion, um die Mittelwertstruktur zu modellieren. Ein weiterer Vorteil der GLM ist, dass die Varianzfunktion nicht notwendigerweise explizit spezifiziert werden muss. Bei dichotomen Daten wird beispielsweise automatisch angenommen, dass sie bernoulli-verteilt sind.

Die Unabhängigkeit ist zum Beispiel verletzt, wenn mehrere Messungen an einem Patienten durchgeführt werden, wie in einem Parallelgruppenvergleich mit mehreren Messungen, oder ein Patient mehrfach behandelt wurde, wie in einer Überkreuzstudie. Auch ist es möglich, die Unabhängigkeit dadurch zu verletzen, dass gepaarte Daten erhoben wurden. Also Messungen zum Beispiel am rechten und linken Auge gemacht wurden. Des Weiteren verletzen Familienstudien die Unabhängigkeit in aller Regel. Werden diese Abhängigkeiten ignoriert, kann es zu falschen Schlussfolgerungen kommen. In der Regel überschätzt man dann die Genauigkeit (Sherman und Cessie, 1997) und es kommt zu falsch-positiven Ergebnissen.

Eine Möglichkeit mit diesen Abhängigkeiten umzugehen, liefern die verallgemeinerten Schätzgleichungen (engl. generalized estimating equations, GEE), die von Liang und Zeger (1986) und Zeger und Liang (1986) eingeführt wurden. In diesem Ansatz werden korrelierte Beobachtungen als ein Cluster angesehen. So lassen sich beispielsweise alle Mitglieder einer Familie, alle Nachbehandlungen eines Patienten oder die Werte des rechten und linken Auges eines Probanden als Cluster ansehen. Es wird angenommen, dass zwischen den Beobachtungen eines Clusters eine Verbindung besteht, während Beobachtungen verschiedener Cluster als unabhängig gelten (Ziegler et al., 1996). Im Gegensatz zu Ansätzen, die auf Standardschätzverfahren wie der Maximum-Likelihood-Schätzung oder der Methode der kleinsten Quadrate beruhen, wie die GLM dies tun, und somit die korrekte Spezifikation der kompletten multivariaten Verteilung

benötigen, um ihre statistischen Eigenschaften wie asymptotische Normalität des Schätzers einzuhalten, liegt die Stärke der GEE darin, dass lediglich die Mittelwertstruktur korrekt spezifiziert sein muss. Valide Schätzungen der Regressionskoeffizienten sind mithilfe der GEE auch dann möglich, wenn die Varianzen und Korrelationen fehlerhaft spezifiziert wurden. Selbstverständlich hat auch diese Eigenschaft der Robustheit der GEE Folgen, denn es kann zu einem Verlust in der Effizienz führen.

Genau wie bei den GLM müssen bei den GEE die Linkfunktion und eine Verteilungsannahme, welche dann die Mittelwertstruktur und die Varianzfunktion liefert, angegeben werden. Zusätzlich muss noch eine Korrelationsstruktur, die sogenannte Arbeitskorrelationsstruktur, gewählt werden, da die wahre Korrelationsstruktur in den meisten Anwendungen unbekannt ist. In Kapitel 2 wird die Theorie der GEE dargestellt. Hierbei wird deutlich, dass die Wahl der zugrunde liegenden Korrelationsstruktur einen bedeutenden Einfluss auf die Effizienz und die Validität hat. Daher wird in Kapitel 3 dieser Punkt besonders betrachtet. Hierbei wird sowohl anhand biologisch plausibler als auch statistischer Gründe diskutiert, wann es sinnvoll ist, welche Arbeitskorrelationsstruktur zu wählen. Bisher wurde noch kein solcher Überblick über dieses Thema gegeben und so ausführlich wie in Ziegler und Vens (2010), dem zugrunde liegenden Artikel, diskutiert.

Verallgemeinerte Schätzgleichungen wurden bisher wenig genutzt, um kontrollierte klinische Studien auszuwerten. Um den Nutzen einer solchen Analyse darzulegen, wird in Kapitel 4 eine doppelblinde, Placebo kontrollierte, randomisierte, multizentrische Studie erneut analysiert. Es soll gezeigt werden, dass die Anwendung von GEE in einer solchen Situation Vorteile bietet, da sie nicht, wie die Standardanalysen, nur die letzte Nachuntersuchung und möglicherweise eine Adjustierung zur Messung bei Studienbeginn (Baseline-Messung) berücksichtigt. Es wird auch auf die Erkenntnisse aus Kapitel 3 zurückgegriffen, um die optimale Arbeitskorrelationsstruktur zu wählen. Des Weiteren kommen Regressionsdiagnostiken zum Einsatz, um Ausreißer zu identifizieren. Hierbei wird sowohl auf die klassischen Verfahren zurückgegriffen als auch ein neuer rechenzeitintensiver Prozess, der auf den Abbildungen von Venezuela et al. (2007) basiert, genutzt. Die Ergebnisse dieses Kapitels finden sich auch in Vens und Ziegler (2012).

Bisher wurde immer davon ausgegangen, dass der Zusammenhang der unabhängigen Variablen auf die Zielgröße linearer Natur ist. Sollte zwischen einer Einflussgröße und der Zielvariablen ein nicht linearer Zusammenhang vorliegen, lässt sich also mit den herkömmlichen Methoden kein zufriedenstellendes Ergebnis erzielen. In Kapitel 5 wird

die Möglichkeit, die GEE mit multiplen fraktionellen Polynomen (MFP), einem eleganten Ansatz den nicht linearen Einfluss verschiedener Prädiktoren auf eine Zielvariable bei unabhängigen Beobachtungen zu beschreiben, zu kombinieren anhand einer Familienstudie erläutert. Da es nicht möglich ist, Methoden zur Bestimmung der Güte der Anpassung bei GEE, die auf der Likelihood und somit einer Devianz wie bei den multiplen fraktionellen Polynomen beruhen, heranzuziehen, wird sich zunächst der Kriterien bedient, die bereits in Kapitel 3 beschrieben werden. Diese fanden auch bereits in der Literatur Anwendung bei Cui et al. (2009). Allerdings stellt sich heraus, dass die bisher verwendeten Ansätze in diesem Zusammenhang zu ungünstigen Anpassungen oder auch zu einem Overfitting führen können. Daher wird ein neues Kriterium vorgeschlagen, was in dieser Situation angemessener scheint. Anhand der Daten der Framingham Heart Study (Cupples et al., 2003) und einer zehnfach Kreuzvalidierung werden die Vorteile der Kombination aus GEE und MFP in Verbindung mit dem neu entwickelten Gütekriterium herausgestellt.

Abschließend wird in Kapitel 6 eine Diskussion der vorgestellten Methoden zur verbesserten Anpassung in GEE sowie ein Ausblick und in Kapitel 7 eine Zusammenfassung gegeben.

2. Verallgemeinerte Schätzgleichungen

Die Methode der GEE wurde erstmalig von Liang und Zeger (1986) und Zeger und Liang (1986) eingeführt und seitdem mehrfach modifiziert und erweitert.

Ausgangspunkt seien n unabhängige Cluster $i = 1, \dots, n$, in denen jeweils T abhängige Beobachtungen $t = 1, \dots, T$ vorliegen. Es ist auch möglich, die Methode für ungleiche Clustergrößen T_i zu erweitern. Für jede zufällige Zielvariable y_{it} steht ein p -dimensionaler Vektor \mathbf{x}_{it} unabhängiger Variablen zur Verfügung, welcher möglicherweise auch eine Regressionskonstante enthält. Die Daten werden in dem T -dimensionalen Vektor $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$ und der $(T \times p)$ -Matrix $\mathbf{X}_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{iT})'$ zusammengefasst. Im Kontext der GEE wird die gestapelte Matrix der \mathbf{X}_i Designmatrix für die Mittelwertstruktur oder kurz X -Matrix genannt, selbst dann, wenn die Kovariablen zufällig sind. Die Paare $(\mathbf{y}_i, \mathbf{X}_i)$ seien unabhängig identisch verteilt.

Die Mittelwertstruktur ist gegeben durch

$$\mu_{it} = \mathbb{E}(y_{it} | \mathbf{X}_i) = \mathbb{E}(y_{it} | \mathbf{x}_{it}) = g(\mathbf{x}'_{it} \boldsymbol{\beta}), \quad (2.1)$$

wobei $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ der unbekannte $(p \times 1)$ -Vektor der Mittelwertparameter ist und g die Responsefunktion wie in den GLM bezeichnet. Somit verbindet die Responsefunktion g die bedingte Erwartung von y_{it} gegeben \mathbf{x}_{it} mit den Kovariablen \mathbf{x}_{it} . Es ist zu beachten, dass mit Gleichung (2.1) explizit angenommen wird, dass die Zeitpunkt spezifische Kovariate \mathbf{x}_{it} lediglich einen Effekt auf die Beobachtung y_{it} am Zeitpunkt t hat. Ein möglicher Effekt auf vergangene oder zukünftige Beobachtungen $\mathbf{x}_{it'}$, $t' \neq t$ wird vernachlässigt. Diese Annahme wird in Kapitel 3 noch genauer betrachtet.

Wir nehmen an, dass die Mittelwertstruktur korrekt spezifiziert ist. Ferner existiere die wahre Kovarianzmatrix $\boldsymbol{\Omega}_i$ der \mathbf{y}_i gegeben \mathbf{X}_i . Diese Matrix muss allerdings im Kontext der GEE für die Mittelwertstruktur (GEE1) nicht korrekt spezifiziert sein. Die angenommene Kovarianzmatrix oder Arbeitskovarianzmatrix von \mathbf{y}_i gegeben \mathbf{X}_i wird mit $\boldsymbol{\Sigma}_i$ bezeichnet. Im Regelfall wird die Zerlegung

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{diag}(v_{it}^{1/2}) \mathbf{R}_i(\boldsymbol{\alpha}) \text{diag}(v_{it}^{1/2}) \quad (2.2)$$

verwendet, wobei $\text{diag}(v_{it}^{1/2})$ die Diagonalmatrix der bedingten Varianzen

$$v_{it} = \text{Var}(y_{it} | \mathbf{X}_i) = \varphi \cdot h(\mu_{it})$$

von y_{it} gegeben \mathbf{X}_i wie im GLM, φ einen Skalenparameter und h die Varianzfunktion bezeichnet. $\mathbf{R}_i(\boldsymbol{\alpha})$ sei die möglicherweise Cluster spezifische Arbeitskorrelationsmatrix,

die in der Regel in einem q -dimensionalen Nuisance-Parameter α formuliert wird.

Besteht kein Vorwissen über die vorliegende Korrelationsstruktur, ist es möglich, eine unstrukturierte Arbeitskorrelationsmatrix zu wählen (Liu et al., 2000; Ziegler und Vens, 2010). Diese hängt dann von $T_i(T_i - 1)/2$ unabhängigen Parametern ab, die im Modell geschätzt werden müssen. Somit hat α die Dimension $T_i(T_i - 1)/2$. Eine unstrukturierte (engl. unstructured, UNSTR) Korrelationsstruktur ist gegeben durch

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & & & \\ \rho_{1,2} & 1 & & \\ \vdots & \ddots & \ddots & \\ \rho_{1,T_i} & \cdots & \rho_{T_{i-1},T_i} & 1 \end{pmatrix}.$$

Da es sich bei den Korrelationsmatrizen um symmetrische Matrizen handelt, wird jeweils nur die untere Dreiecksmatrix angegeben, aus der sich aufgrund der Symmetrie die obere Dreiecksmatrix konstruieren lässt.

Sind Kenntnisse über die Korrelationsstruktur vorhanden, lässt sich die Arbeitskorrelationsmatrix in bestimmten Situationen über eine Funktion durch weniger Parameter darstellen. Dies führt selbstverständlich auch zu einer Verringerung der zu schätzenden Parameter α . Im Folgenden werden vier häufig verwendete Beispiele sogenannter strukturierter Arbeitskorrelationsmatrizen beschrieben.

Die einfachste Möglichkeit besteht in der Annahme der Unabhängigkeit (engl. independence, IND). Hier wird davon ausgegangen, dass innerhalb eines Clusters keine Korrelationen vorliegen. Der Parameter α muss in diesem Fall nicht geschätzt werden. Die zugehörige Korrelationsmatrix lässt sich durch die Einheitsmatrix

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

abbilden.

Eine austauschbare (engl. exchangeable, EXCH) Korrelationsstruktur beschreibt die Situation, in der die Korrelation zwischen verschiedenen Elementen eines Clusters konstant ist. In diesem Fall muss im Modell der Parameter α durch einen Skalar geschätzt

werden. Die austauschbare Arbeitskorrelationsmatrix wird dargestellt durch

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \vdots & \ddots & \ddots & \\ \rho & \cdots & \rho & 1 \end{pmatrix}.$$

Eine weitere häufig verwendete Annahme ist die autoregressive Korrelationsstruktur erster Ordnung (engl. first-order autoregressive, AR(1)). Hier nimmt die Korrelation zwischen den Elementen eines Clusters mit wachsendem Abstand exponentiell ab. Der Parametervektor lässt sich hier darstellen als $\boldsymbol{\alpha} = (1, \rho, \rho^2, \dots, \rho^{T_i-1})$ mit $|\rho| < 1$. Es muss also ein zusätzlicher Parameter geschätzt werden. Als Korrelationsmatrix ergibt sich

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \vdots & \ddots & \ddots & \\ \rho^{T_i-1} & \cdots & \rho & 1 \end{pmatrix}.$$

Die m -Abhängige Arbeitskorrelationsstruktur (engl. m -dependent working correlation, DEP(m)) wird nicht häufig verwendet. Es handelt sich hierbei um eine Bandmatrix der Breite $2m + 1$. Außerhalb dieses Bandes werden die Korrelationen auf Null gesetzt. Somit sind immer m Beobachtungen abhängig voneinander. Der Parametervektor besteht also aus m Parametern, die zusätzlich geschätzt werden müssen. Als Korrelationsmatrix ergibt sich dann

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & & & & & & & & \\ \rho_1 & 1 & & & & & & & & & \\ \vdots & \ddots & \ddots & & & & & & & & \\ \rho_m & \cdots & \rho_1 & 1 & & & & & & & \\ & \ddots & & \ddots & \ddots & & & & & & \\ & & \rho_m & \cdots & \rho_1 & 1 & & & & & \\ \mathbf{0} & & & \rho_m & \cdots & \rho_1 & 1 & & & & \end{pmatrix}.$$

Weitere Arbeitskorrelationsmatrizen finden sich in Ziegler et al. (1998), Hilbe und Hardin (2002) oder Ziegler (2013).

Die GEE1 lassen sich unter Verwendung der Quasi Generalisierten Pseudo Maximum Likelihood (QGPML) herleiten (Gourieroux und Monfort, 1995). Hierfür werden drei Annahmen benötigt:

1. \mathbf{y}_i gegeben \mathbf{X}_i seien normalverteilt mit Mittelwertstruktur $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iT})'$ und Kovarianzmatrix $\boldsymbol{\Sigma}_i$.
2. Der Nuisance-Parameter $\boldsymbol{\alpha}$ kann konsistent durch $\hat{\boldsymbol{\alpha}}$ geschätzt werden.
3. Weiterhin sei $\tilde{\boldsymbol{\beta}}$ eine initiale Schätzung des Parametervektors $\boldsymbol{\beta}$, so dass die Arbeitskovarianzmatrix $\boldsymbol{\Sigma}_i$ durch $\tilde{\boldsymbol{\Sigma}}_i$ geschätzt werden kann.

Damit lauten die GEE1 Schätzgleichungen

$$\mathbf{u}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i' \tilde{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \quad (2.3)$$

für $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}'$ (Liang und Zeger, 1986). Ein Schätzer $\hat{\boldsymbol{\beta}}_R$, der unter Verwendung der Arbeitskorrelationsmatrix \mathbf{R} die GEE1 Schätzgleichungen (2.3) löst, heisst GEE1 Schätzer für $\boldsymbol{\beta}$. Wird die Einheitsmatrix als Arbeitskorrelationsmatrix verwendet, heißen die Schätzgleichungen Unabhängigkeitsschätzgleichungen (engl. independent estimating equations, IEE), und der resultierende Schätzer $\hat{\boldsymbol{\beta}}_I$ ist der IEE Schätzer für $\boldsymbol{\beta}$. Die Schätzung von $\boldsymbol{\beta}$ erfolgt durch einen Schaukelalgorithmus, also ein zweistufiges Verfahren, in dem wechselweise $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\alpha}}$ bis zur Konvergenz aktualisiert werden.

Folgende Eigenschaften lassen sich für einen GEE1 Schätzer $\hat{\boldsymbol{\beta}}$ unter Verwendung der initialen Schätzung $\tilde{\boldsymbol{\Sigma}}_i$ herleiten (Liang und Zeger, 1986; Gourieroux und Monfort, 1995; Ziegler et al., 1998; Ziegler, 2013):

- Er existiert mit Wahrscheinlichkeit 1, bis auf eine Nullmenge, eindeutig.
- Er ist stark konsistent.
- Er ist asymptotisch erwartungstreu
- und asymptotisch normalverteilt mit Kovarianzmatrix $\mathbf{C} = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$, wobei \mathbf{A} die Fisher-Informationsmatrix und \mathbf{B} den Erwartungswert des äußeren Produkts des Gradienten (engl. outer product gradient, OPG) bezeichnet. Die Matrix \mathbf{C} wird als robuste Kovarianzmatrix oder Sandwich-Matrix bezeichnet.

- Gilt zudem $\Sigma_i = \Omega_i$, ist der GEE1 Schätzer asymptotisch effizient und der Schätzer ist bester asymptotisch normaler Schätzer.

A und B lassen sich stark konsistent schätzen durch

$$\hat{A}(\hat{\beta}) = \sum_{i=1}^n \hat{A}_i = \sum_{i=1}^n (\hat{D}_i \hat{\Sigma}_i^{-1} \hat{D}_i) \quad (2.4)$$

und

$$\hat{B}(\hat{\beta}) = \sum_{i=1}^n \hat{B}_i = \sum_{i=1}^n (\hat{D}_i \hat{\Sigma}_i^{-1} \hat{\Omega}_i \hat{\Sigma}_i^{-1} \hat{D}_i), \quad (2.5)$$

wobei \hat{D}_i , $\hat{\Sigma}_i$ und $\hat{\Omega}_i$ die Schätzer von D_i , Σ und Ω_i bezeichnen.

Die GEE1 ermöglichen die konsistente Schätzung der Mittelwertstruktur, selbst wenn die Kovarianzmatrix fehlspezifiziert ist. In Anwendungen ist auch keine korrekte Spezifikation der Arbeitskorrelationsstruktur zu erwarten. Allerdings gilt, je näher die Arbeitskorrelationsmatrix an der wahren Korrelationsstruktur liegt, desto effizienter ist der Schätzer (Chaganty und Joe, 2004).

Im folgenden Kapitel soll daher die Wahl der Arbeitskorrelationsmatrix näher beschrieben werden.

3. Wahl der Arbeitskorrelationsmatrix

*Ziegler A, Vens M (2010)
Generalized estimating equations: notes on the choice of the working correlation matrix.
Methods of Information in Medicine 49, 421–425; Diskussion 426–432.
(Übersetzte und erweiterte Version)*

Am Ende von Kapitel 2 wurde deutlich, dass Effizienz und Validität des GEE Ansatzes von der Wahl der Arbeitskorrelationsmatrix abhängig sind. So sollte die Arbeitskorrelationsmatrix möglichst nah an der wahren Korrelationsstruktur liegen. Aber wie kann eine solche Arbeitskorrelationsmatrix gewählt werden? In diesem Kapitel sollen verschiedene Aspekte bezüglich Effizienz und Validität des GEE Ansatzes noch einmal aufgegriffen werden. Dabei soll auf die folgenden drei Punkte besonders Wert gelegt und die daraus resultierenden Fragen im Laufe des Kapitels beantwortet werden.

1. Die einfachste Wahl der Arbeitskorrelationsmatrix ist, wie in Kapitel 2 beschrieben, die Einheitsmatrix. Diese Wahl führt zu den IEE. Wann ist es hinreichend, die IEE im Gegensatz zu den komplexeren GEE zu benutzen?
2. Gibt es weitere Argumente, eine spezifische Arbeitskorrelationsmatrix zu wählen? Ist es wirklich die Effizienz, welche die Wahl der Arbeitskorrelationsmatrix vorantreibt? Gibt es Situationen, in denen es ratsam ist, die IEE zu nutzen?
3. Es gibt viele Optionen bei der Wahl der Arbeitskorrelationsmatrix. Angenommen, die Wahl fiel auf ein GEE Modell mit einer anderen Arbeitskorrelationsmatrix als der Einheitsmatrix. Wie kann die am besten geeignete Arbeitskorrelationsmatrix, die zu einem minimalen Verlust der Effizienz führt, gewählt werden?

3.1. Allgemeine Kriterien zur Wahl der Arbeitskorrelationsmatrix

Zunächst einmal scheint die Antwort auf die Frage nach der Arbeitskorrelationsmatrix einfach: Man sollte diejenige Arbeitskorrelationsmatrix wählen, welche möglichst nah an der wahren Korrelationsstruktur liegt. Leider ist die wahre Korrelationsstruktur zu meist unbekannt. Aus diesem Grund folgen Fragen wie: Gibt es Situationen, in denen es ratsam ist, die Unabhängigkeitsstruktur zu wählen? Ist es möglich, die angemessenste Arbeitskorrelationsmatrix auszuwählen?

Die Effizienz der IEE im Vergleich mit den GEE hängt von verschiedenen Faktoren ab. Hierzu zählen unter anderem die Verteilung der Kovariaten, die Größe der Cluster, die Regressionsparameter und die Korrelationen zwischen den abhängigen Variablen.

Die allgemeinsten theoretischen Ergebnisse wurden von Mancl und Leroux (1996) hergeleitet. Sie zeigten für verschiedene Modelle und gleich große Cluster, dass die IEE genauso asymptotisch effizient sind wie die GEE mit einer austauschbaren Arbeitskorrelationsstruktur, wenn eine der folgenden Bedingungen erfüllt ist:

- Die Zielvariablen innerhalb der Cluster sind unabhängig.
- Alle Kovariaten innerhalb der Cluster sind konstant.
- Alle Kovariaten sind Mittelwert balanciert. Das bedeutet, dass die Clustermittelwerte zwischen allen Clustern konstant sind.
- Die Kovariaten sind eine Mischung von solchen, die innerhalb der Cluster konstant sind und solchen, die Mittelwert balanciert sind.

Die erste Bedingung bildet die Unabhängigkeit ab. Sobald die Beobachtungen innerhalb der Cluster unabhängig sind, ist die Korrelation gleich Null und kein GEE kann dem IEE überlegen sein. Der zweite Aspekt ist zum Beispiel bei klinischen Studien mit Parallelgruppenvergleich und wiederholten Messungen erfüllt. Hier erhält der Proband dieselbe Behandlung über die gesamte Dauer der klinischen Prüfung. Adjustierungen für konstante Kovariaten, wie zum Beispiel die Kovariaten zu Studienbeginn (Baseline-Kovariaten), sorgen für Cluster konstante Kovariaten. Die dritte Situation wird beispielsweise in Überkreuzstudien erfüllt. Zum Beispiel erhält ein Proband in einem (2×2) -Überkreuzvergleich beide Behandlungen, allerdings variiert die Anwendungsreihenfolge. Wird die aktive Behandlung mit 1 und das Placebo mit 0 kodiert, ist der Mittelwert $1/2$ für die Behandlung für alle Probanden. Mittelwert balancierte Kovariaten werden auch erreicht, indem die Zeit als Kovariate benutzt wird und alle Probanden an denselben Zeitpunkten untersucht werden.

Im Fall ungleicher Clustergrößen sind die IEE immer noch effizient, wenn die Kovariaten Mittelwert balanciert und Cluster konstant sind. Allerdings kann es auch im Fall der Cluster konstanten Kovariaten zu einem Verlust der Effizienz kommen, selbst wenn die Clustergrößen identisch sind. Für ungleiche Clustergrößen kann dieser Verlust sogar enorm sein. Obwohl analytische Ergebnisse nicht für die gesamte Klasse der GLM verfügbar sind, zeigen Monte-Carlo-Simulationen, dass die vorliegenden Ergebnisse auch auf die Fälle kleiner oder moderater Effektstärken übertragbar sind. Der Verlust der Effizienz wird größer für den Fall kleiner Stichprobenumfänge (Wang und Carey, 2003).

Diese theoretischen Ergebnisse können wie folgt zusammengefasst werden:

1. Sollten die Clustergrößen nicht stark variieren und sind alle Kovariaten Cluster konstant, Mittelwert balanciert, oder eine Mischung aus beiden genannten Möglichkeiten, sind die IEE genauso effizient wie die GEE mit einer nicht unabhängigen Arbeitskorrelationsstruktur.
2. Ungleiche Clustergrößen kommen häufig in Beobachtungsstudien vor. Ein enormer Verlust der Effizienz kann festgestellt werden, sollten in einem solchen Fall die IEE genutzt werden.
3. Im Allgemeinen sollte eine angemessen begründete nicht unabhängige Arbeitskorrelationsstruktur gewählt werden. Intensive Modellierung der Arbeitskorrelationsmatrix liefert jedoch lediglich vernachlässigbare Gewinne in der Effizienz.

Nach der Betrachtung dieser theoretischen Aspekte, werden im Folgenden fünf Erkenntnisse aus der Literatur beleuchtet, die in ein Dilemma bei der Wahl der Arbeitskorrelationsstruktur münden können.

1. Die statistischen Eigenschaften des Schätzers $\hat{\beta}$ sind von der Unverzerrtheit der GEE (Gleichung (2.3)) abhängig. Diese Eigenschaft ist für die IEE immer gegeben. Allerdings ist es möglich, dass der Schätzer im Falle der GEE mit einer nicht unabhängigen Korrelationsstruktur verzerrt ist. Eine hinreichende Bedingung für die Konsistenz ist, dass die abhängige Variable y_{it} am Zeitpunkt t im Cluster i unter Verwendung aller Kovariaten des gesamten Clusters, also $x_{i1}, x_{i2}, \dots, x_{iT}$, korrekt spezifiziert ist (Pepe und Anderson, 1994; Pan et al., 2000). Diese Annahme wird implizit in Gleichung (2.1) gemacht und ist erfüllt, wenn beispielsweise nur Kovariaten benutzt werden, die über die Zeit hinweg betrachtet nicht variieren. Allerdings kann diese Annahme in verschiedenen Anwendungen verletzt sein. Hierzu seien die folgenden drei Beispiele gegeben.
 - In Familienstudien ist es möglich, dass der Stresspegel der Eltern einen Effekt auf den Gesundheitszustand des Kindes hat. Im Gegenzug ist es auch möglich, dass der Stresspegel der Kinder den Gesundheitszustand der Eltern beeinflusst. Hierbei werden die Cluster i durch die Familien repräsentiert und jede Familie i besteht aus den Familienmitgliedern $t = 1, \dots, T_i$.
 - In Studien mit Schulkindern ist es möglich, dass die Belastung anderer Kin-

Tabelle 3.1: (2×2) -Häufigkeitstabelle mit Elementen $\mu_{tt'} > 0$ für $t, t' = 1, 2$.

		y_2		Total
		1	2	
y_1	1	μ_{11}	μ_{12}	μ_1
	2	μ_{21}	μ_{22}	
Total		μ_2	1	

der den Gesundheitsstatus eines bestimmten Kindes beeinflusst. Es kann beispielsweise jede Schulklasse durch ein Cluster i mit den Schulkindern $t = 1, \dots, T_i$ dargestellt werden.

- In Studien zur Zahngesundheit ist es möglich, dass der Gesundheitsstatus benachbarter Zähne einen einzigen Zahn beeinflusst. In einer solchen Studie wäre das Cluster i der Patient oder die Patientin und $t = 1, \dots, T$ seine Zähne.

In jedem dieser Fälle muss die Validität der Annahme $\mathbb{E}(y_{it}|x_{it}) = \mathbb{E}(y_{it}|X_i)$ (siehe auch Gleichung 2.1) analytisch gezeigt werden.

2. Ein großer Vorteil der IEE ist es, dass sie in den meisten Anwendungsfällen, solange die Fallzahl nicht allzu klein ist, konvergieren (Dahmen und Ziegler, 2006). Werden zusätzliche Korrelationsparameter in das Modell eingeschlossen, konvergiert der Algorithmus seltener (Emrich und Piedmonte, 1992).
3. Im Fall dichotomer abhängiger Variablen ist der Parameterraum des Korrelationskoeffizienten beschränkt (Prentice, 1988; McDonald, 1993; Chaganty und Joe, 2004). Betrachten wir hierzu eine (2×2) -Häufigkeitstabelle wie Tabelle 3.1. Da $\mu_1 = \mu_{21} + \mu_{22}$ und $\mu_2 = \mu_{12} + \mu_{22}$ gilt, ist μ_{22} beschränkt durch

$$\max \{0, \mu_1 + \mu_2 - 1\} \leq \mu_{22} \leq \min \{\mu_1, \mu_2\}.$$

Dadurch kann der Korrelationskoeffizient ρ nur Werte aus dem Intervall

$$\max \left\{ -\sqrt{\frac{\mu_1 \mu_2}{\eta_1 \eta_2}}, -\sqrt{\frac{\eta_1 \eta_2}{\mu_1 \mu_2}} \right\} \leq \rho \leq \min \left\{ \sqrt{\frac{\mu_1 \eta_2}{\mu_2 \eta_1}}, \sqrt{\frac{\mu_2 \eta_1}{\mu_1 \eta_2}} \right\}$$

mit $\eta_t = 1 - \mu_t$, $t = 1, 2$, annehmen. Abbildung 3.1 zeigt die oberen und unteren Grenzen des Korrelationskoeffizienten auf der y-Achse zu variierenden marginalen Häufigkeiten μ_2 auf der x-Achse. Hierzu seien zwei Beispiele gegeben. Zunächst seien $\mu_2 = 0,1$ und $\mu_1 = 0,1$ (durchgezogene Linie). Daraus ergibt sich, dass die obere Grenze des Korrelationskoeffizienten ρ gleich 1 und somit nach

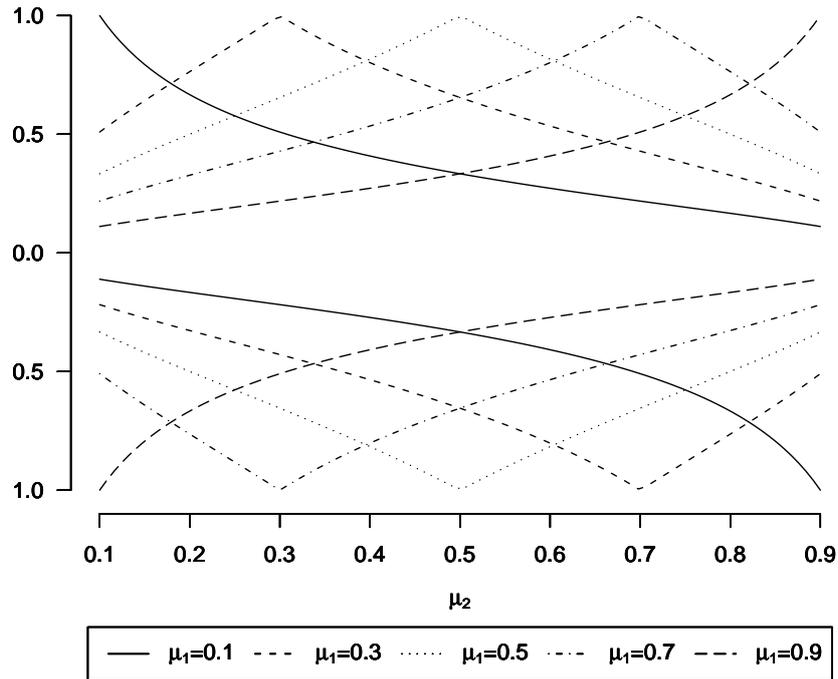


Abbildung 3.1: Obere und untere Grenzen des Korrelationskoeffizienten einer (2×2) -Tabelle wie Tabelle 3.1 für unterschiedliche marginale Häufigkeiten μ_1 und μ_2 . Auf der x-Achse ist die marginale Frequenz von μ_2 und die obere (untere) Grenze über (unter) 0 auf der y-Achse aufgetragen.

oben nicht beschränkt ist, während die untere Grenze beschränkt wird durch $-0,11$. Wird hingegen davon ausgegangen, dass $\mu_2 = 0,1$ und $\mu_1 = 0,9$ (gestrichelte Linie) seien, ergibt sich als obere Grenze ein Wert von $0,11$, während der Korrelationskoeffizient nach unten unbeschränkt ist.

Der annehmbare Bereich des Korrelationskoeffizienten ρ ist die Schnittmenge aller Beschränkungen für alle Paare $(\mu_{it}, \mu_{it'})$ für alle $i = 1, \dots, n$ und $t \neq t'$. Im Allgemeinen ist μ_{it} zudem eine Funktion, die von den Kovariaten x_{it} abhängt, was zu einer noch komplizierteren und stärkeren Serie von Beschränkungen des Korrelationskoeffizienten führt (Chaganty und Joe, 2004). Abhängig vom Bereich von x kann das Intervall des Korrelationskoeffizient recht schmal werden. Weist beispielsweise nur eine einzelne Kovariate einen quantitativen Wertebereich über den gesamten Bereich der reellen Zahlen auf, enthält das Intervall für ρ nur noch den Punkt 0. Somit wird die Absicht der Modellierung einer Arbeitskorrelationsmatrix im Fall binärer Daten vollends bezwungen.

4. Die GEE Schätzer existieren nicht, wenn die Arbeitskorrelationsmatrix $R_i(\alpha)$ des i -ten Clusters den Eintrag $-\frac{1}{T-1}$ auf allen Positionen außerhalb der Hauptdia-

gonalen hat, da diese Korrelationsmatrix nicht mehr regulär ist (Hanley et al., 2000). Um diese Singularität zu vermeiden, kann ein gebundener Schätzer

$$\hat{q}_{\text{bounded}} \geq \max \left\{ -\frac{1}{T-1} + \varepsilon, \hat{q}_{\text{unbounded}} \right\}$$

mit einem sehr kleinen ε genutzt werden. Einige Programmpakete lösen dieses Problem. So wird beispielsweise im SAS-Makro PROC GENMOD $\varepsilon = 0.01$ gewählt. Allerdings ist es in einer solchen Situation sicherer die IEE anstelle der gebundenen GEE zu benutzen (Hanley et al., 2000).

5. Insbesondere sollte ein Augenmerk darauf gelegt werden, mit welcher Methode die Arbeitskorrelationsmatrix geschätzt wird. Nicht mit allen Methoden lassen sich grundsätzlich reellwertige Lösungen erlangen (Crowder, 1995), obwohl auch hier Alternativen bestehen (Wang und Carey, 2003).

3.2. Statistische Kriterien zur Wahl der Arbeitskorrelationsmatrix

Bisher wurden Situationen vorgestellt, in denen die GEE mit einer nicht diagonalen Arbeitskorrelationsmatrix den IEE vorzuziehen sind. In diesem Abschnitt sollen daher statistische Kriterien, die dabei helfen sollen, eine spezifische Arbeitskorrelationsmatrix zu wählen, erläutert werden.

Die Annahme der multivariaten Normalverteilung ist geeignet, um die GEE und ihre Eigenschaften herzuleiten. Allerdings wird in den Modellen zur Herleitung der Schätzgleichungen explizit zwischen der wahren und der angenommenen Verteilung unterschieden. Es ist gerade der Vorteil der GEE, dass die wahre Verteilung nicht korrekt spezifiziert sein muss. Als Konsequenz daraus können Likelihood basierte Maße zur Beurteilung der Modellgüte und Verfahren zur Modellselektion nicht angewendet werden.

Zwei Alternativen werden in der Literatur für die Modellwahl diskutiert. Zum einen werden Bootstrap geglättete Kreuzvalidierungsverfahren für den durchschnittlichen quadratischen Vorhersagefehler (Pan und Connett, 2002) und für die erwartete Vorhersageverzerrung (Pan, 2001b) betrachtet. Diese Verfahren sind jedoch sehr rechenzeitin-

tinsiv und daher wenig praktikabel.

Zum anderen werden Maßzahlen vorgeschlagen, die Ähnlichkeiten mit dem Akaike Informationskriterium (engl. Akaike's information criterion, AIC) aufweisen (Pan, 2001a; Hilbe und Hardin, 2002; Cui und Qian, 2007). Beliebte sind auch Kriterien, die den Unterschied zwischen der geschätzten Fisher-Informationsmatrix, also der geschätzten Modell basierten Kovarianzmatrix $\hat{A}_P = \frac{1}{n} \sum_{i=1}^n \left(\hat{D}_i' \hat{\Sigma}_i^{-1} \hat{D}_i \right)$, und der geschätzten robusten Kovarianzmatrix \hat{C}_R abbilden (Shults und Chaganty, 1998; Pan, 2001a; Hin et al., 2007; Hin und Wang, 2009). In dem Ausdruck zu \hat{A}_P bezeichne $\hat{\Sigma}_i$ die geschätzte Diagonalmatrix der Varianzen v_{it} in Cluster i , \hat{D}_i und $\hat{\Sigma}_i$ basieren auf $\hat{\beta}_R$, welches durch ein GEE Modell mit der Arbeitskorrelationsmatrix R geschätzt wurde. Basierend auf dieser Notation ist \hat{C}_R die geschätzte robuste Varianzmatrix, wenn die Arbeitskorrelationsmatrix R genutzt wurde. Viele Kriterien betrachten die Transformation $\hat{Q} = \hat{A}_P^{-1} \hat{C}_R$.

Zunächst stellte Pan (2001a) das Quasi Likelihood Informationskriterium (engl. quasi likelihood information criterion, QIC) vor. Dieses wurde in vielerlei Hinsicht modifiziert, und es wurden zusätzliche Maßzahlen, die ähnliche Strukturen besitzen, eingeführt.

Ausgangspunkt für die Herleitung des QIC ist die Quasi Likelihood $\varphi Q(\mu)$, wobei φ den Streuungsparameter und $Q(\mu)$ die Quasi Likelihoodfunktion unter der Unabhängigkeitsannahme bezeichnen. Dann wird das Quasi Likelihood Informationskriterium QIC_P nach Pan geschätzt durch

$$\widehat{QIC}_P(\mathbf{R}) = -2\hat{\varphi}_P Q(\hat{\mu}) + 2 \operatorname{tr}(\hat{A}_R^{-1} \hat{C}_R). \quad (3.1)$$

Unter der Verwendung von $\hat{\beta}_R$ werden $\hat{\mu}$ und $Q_P(\hat{\mu})$ geschätzt. Die Quasi Likelihood Funktion wird auch hier unter der Annahme der Unabhängigkeit aller Beobachtungen berechnet. Die Arbeitskorrelationsmatrix wird mithilfe eines zweistufigen Verfahrens bestimmt. Im ersten Schritt wird das beste Modell über den QIC ausgewählt. Hierbei ist die beste Arbeitskorrelationsmatrix diejenige zu dem Modell mit dem geringsten QIC.

Ist das Unabhängigkeitsmodell korrekt, dann gleicht $\hat{Q} = \hat{A}_R^{-1} \hat{C}_R$ asymptotisch der Einheitsmatrix. Entsprechend wäre in diesem Fall $C_1 = \operatorname{tr}(\hat{Q}) \approx p$. Dies führt zu einer vereinfachten Version des QIC, genannt QIC_u , welches durch $\widehat{QIC}_u = -2\hat{\varphi} Q(\hat{\mu}) + 2p$ geschätzt wird. Diese Vereinfachung wird nur durch die Annahme der Unabhängigkeit erreicht. Daher ist QIC_u nur zur Selektion von Variablen der Mittelwertstruktur geeignet. Im Gegensatz dazu kann QIC_P sowohl zur Selektion der Mittelwert- als auch

zur Auswahl der Assoziationsstruktur eingesetzt werden. Wird QIC_P als Gütemaß verwendet, ist dasjenige Modell das Beste, welches QIC_P minimiert. Entsprechendes gilt für QIC_U für die Mittelwertstruktur.

Anstatt $\hat{\beta}_R$ zur Schätzung des QIC zu verwenden, kann dieses auch als $\text{QIC}_I(\mathbf{R})$, welches auf \hat{A}_I , der geschätzten Fisher-Informationsmatrix mit der Identität als Arbeitskorrelationsmatrix (Hilbe und Hardin, 2002; Cui und Qian, 2007), basiert, berechnet werden. $\text{QIC}_P(\mathbf{R})$ und $\text{QIC}_I(\mathbf{R})$ ähneln sich in der Praxis, da $\hat{\beta}$ in beiden Fällen konsistent für β ist. Allerdings können sich die resultierenden besten Modelle unterscheiden. Dieses ist für Anwendungen von Bedeutung, da sich die derzeit in kommerziellen Programmen verfügbaren Implementationen unterscheiden.

Der relevante Teil der Modellierung der Arbeitskovarianzmatrix ist der zweite Term aus Gleichung (3.1). Daher wurde als alternatives Selektionskriterium von Hin und Wang (2009) das Korrelationsinformationskriterium (engl. correlation information criterion, CIC) $\text{CIC}(\mathbf{R}) = C_1 = \text{tr}(\hat{\mathbf{Q}})$ vorgeschlagen, das in den Varianten $\text{CIC}_P(\mathbf{R})$ und $\text{CIC}_I(\mathbf{R})$ geschätzt werden kann.

Eine Reihe weiterer Kriterien ähneln dem CIC: So haben Hin et al. (2007) das Rotnitzky-Jewell Kriterium $\text{RJ}(\mathbf{R}) = \sqrt{\left(1 - \frac{C_1}{p}\right)^2 + \left(1 - \frac{C_2}{p}\right)^2}$ betrachtet, wobei $C_2 = \text{tr}(\mathbf{Q}^2)$.

Die Arbeitskorrelationsstruktur, welche zum Modell mit dem kleinsten CIC oder $\text{RJ}(\mathbf{R})$ führt, wird hierbei jeweils als die beste Korrelationsstruktur gewählt.

Weitere Kriterien sind C_2 und $\text{DBAR} = C_1 - 2C_2 + 1$. Shults und Chaganty (1998) schlagen verallgemeinerte Fehlerquadrate $\text{SC} = \sum_{i=1}^n \hat{\mathbf{f}}_i' \hat{\mathbf{R}}_i^{-1} \hat{\mathbf{f}}_i / (N - p - q)$ mit $N = \sum_{i=1}^n T_i$ vor, p und q sind die Anzahl der zu schätzenden Parameter für die Mittelwert- bzw. die Assoziationsstruktur und $\hat{\mathbf{f}}_i = \text{diag}(h(\hat{\boldsymbol{\mu}}_{it}))^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$.

Ein Vergleich der verschiedenen Maßzahlen QIC, CIC, RJ, SC, DBAR, C_1 und C_2 für GEE1 Modelle geben Hin et al. (2007), Hin und Wang (2009) sowie Shults et al. (2009). Hierbei wird auf Monte-Carlo-Simulationen zurück gegriffen.

3.3. Modellierung der Arbeitskorrelation: Gibt es einen Weg aus dem Dilemma?

Zusammenfassend lässt sich sagen, dass es Situationen gibt, in welchen die IEE genauso effizient sind wie GEE mit einer nicht diagonalen Arbeitskorrelationsmatrix. In anderen Situationen ist es hingegen ratsam, aus Effizienzgründen eine nicht diagonale Korrelationsmatrix zu nutzen. Es wurden statistische Kriterien beschrieben, um die optimale Arbeitskorrelationsmatrix zu finden. Die theoretischen Ergebnisse bezüglich Konsistenz, Konvergenz des Algorithmus, Beschränkungen des Parameterraums sowie die Existenz der Parameter können vor der Benutzung einer nicht diagonalen Arbeitskorrelationsmatrix bewahren. Aus theoretischer Sicht gibt es aber keinen Ausweg aus dem Dilemma „IEE oder nicht IEE“. Allerdings gibt es praktische Wege, dieses Dilemma zu umgehen. Hier sollen drei von ihnen beschrieben werden.

1. In den meisten klinisch-epidemiologischen Studien variieren die Clustergrößen nicht sonderlich. Darüber hinaus sind die unabhängigen Variablen wie Behandlung entweder Cluster konstant oder Mittelwert balanciert. Daher würde die Benutzung der GEE lediglich einen geringen Zugewinn in der statistischen Macht liefern. Alle genannten Argumente führen also zu einer Empfehlung zugunsten der IEE in klinisch-epidemiologischen Studien (Dahmen und Ziegler, 2004).
2. Ein Versuchsleiter würde möglicherweise die optimale Arbeitskorrelationsmatrix für seine Daten basierend auf entweder statistischen Kriterien, von denen einige im weiteren Verlauf näher beschrieben werden, oder aufgrund des biologischen Hintergrunds auswählen. Selbstverständlich ist es möglich, dass die zugrundeliegende Korrelationsstruktur sehr viel komplizierter ist als die angenommene Arbeitskorrelationsmatrix es abbildet. Aus diesem Grund können Sensitivitätsanalysen herangezogen werden, um den Effekt verschiedener Arbeitskorrelationsmatrizen zu untersuchen. Speziell ist es für einen bestimmten Datensatz möglich, dass sich die interessierenden Parameter genauso wie die zugehörigen Standardfehler für verschiedene Arbeitskorrelationsmatrizen sehr ähneln. Ein Beispiel hierfür wird in Kapitel 4 näher betrachtet. Weiterhin lässt sich der Effekt einer spezifischen Kovariatenstruktur auf die Parameterschätzungen für unterschiedliche sinnvolle Modelle mit Hilfe von Monte-Carlo Simulationsstudien untersuchen. Wenn immer möglich, sollte eine spezifische Arbeitskorrelationsstruktur sowohl aufgrund statistischer als auch biologischer Gründe gewählt werden. Während es möglicherweise intuitiv ist, was mit biologischer Plausibilität gemeint ist, müssen

die statistischen Gründe immer noch näher erläutert werden.

Vereinfacht gesagt, kann für Strukturen, ähnlich Haushalten oder Familien, die austauschbare Arbeitskorrelationsstruktur genutzt werden. Für longitudinale Daten sollte die AR(1) Arbeitskorrelationsmatrix gegenüber einer Bandmatrix, wie beispielsweise der Dreibandmatrix, also der 1-Abhängiger Arbeitskorrelationsstruktur, vorgezogen werden (Wang und Carey, 2003). Des Weiteren lässt sich sagen, dass m -Abhängige Korrelationsstrukturen biologisch nicht plausibel sind. Für schwach abhängige dichotome Zielvariablen könnten die IEE das Modell der Wahl sein. In den meisten räumlichen Studien wird die AR(1) Struktur eine sinnvolle Wahl sein. Dennoch können periodische Korrelationsstrukturen oder einfache Sinus Kurven in Baumstudien möglich sein (Baradat et al., 1996). Einfach ausgedrückt wird ein großer Baum eine negative Korrelation zu seinen Nachbarn, die meist kleiner sind, zeigen. Die nächsten Bäume hingegen werden wieder ähnlich groß sein wie der Erste. Daraus folgt eine positive Korrelation zwischen dem ersten und dem dritten Baum. Als Ganzes betrachtet kann es so zu einer periodischen Korrelationsstruktur kommen.

Chaganty und Joe (2004) haben einen einfachen, aber eleganten Ansatz beschrieben, um festzulegen, ob die Abhängigkeit zwischen den abhängigen Variablen schwach, moderat oder stark ist und wie eine spezifische Arbeitskorrelationsmatrix gewählt werden kann. Wenn alle oder die meisten möglichen T -dimensionalen Vektoren von Nullen und Einsen auftreten, ist die Abhängigkeit nicht stark. Die Abhängigkeit ist stark, wenn die Häufigkeiten sich um die Vektoren mit allen Einträgen gleich 0 oder 1 konzentrieren. In allen anderen Fällen geben die Odds Ratios und Korrelationen Aufschluss darüber, ob die Abhängigkeit schwach oder moderat ist. Für dichotome abhängige Variablen sollte ein Korrelationskoeffizient zwischen 0,1 und 0,3 als moderat betrachtet werden und ein Korrelationskoeffizient von 0,4 und größer deutet auf eine starke Abhängigkeit hin. Die typische obere Grenze des Korrelationskoeffizient ist um die 0,7 und nicht nah bei 1. In ähnlicher Weise wird ein Odds Ratio für den Assoziationsparameter zwischen 1 und 3 als schwache, zwischen 3 und 10 als moderate, und über 10 als starke Abhängigkeit betrachtet. Mit diesen Korrelationskoeffizienten oder Odds Ratios, die ohne die Kovariaten berechnet wurden, kann die Arbeitskorrelationsmatrix als eine feste Arbeitskorrelation in der GEE Analyse spezifiziert werden.

Alternativ können auch zunächst die IEE benutzt werden. Geschätzt werden dadurch die marginalen Häufigkeiten, die die Grenzen des Korrelationskoeffizienten für jedes Paar von Clustern kontrollieren, und zum Schluss fällt die Entscheidung auf einen angemessenen Korrelationskoeffizienten in der Mitte der Grenzen.

3. Der letzte Ansatz, um mit den Restriktionen für dichotome abhängige Variablen umzugehen, ist ein zweistufiges Verfahren. Im ersten Schritt wird die optimale Arbeitskorrelationsmatrix gewählt (Pan, 2001a; Pan und Connett, 2002; Cui und Qian, 2007; Hin und Wang, 2009). Dies geschieht mithilfe der in Kapitel 3.2 vorgestellten statistischen Ansätze. In einem zweiten Schritt werden die Parameterrestriktionen untersucht. Wenn das optimale Modell im Sinne des ersten Schritts die Parameterrestriktionen verletzt, wird das zweitbeste Modell genutzt, sofern dieses die Restriktionen einhält. Ansonsten wird das drittbeste Modell gewählt und so weiter.

4. Verallgemeinerte Schätzgleichungen und Regressionsdiagnostik in longitudinalen klinischen Studien

Vens M, Ziegler A (2012)
*Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials:
a case study.*
Computational Statistics and Data Analysis 56, 1232–1242.
(Übersetzte und erweiterte Version)

Der GEE Ansatz wurde in vielerlei Hinsicht modifiziert und erweitert. Diese Erweiterungen umfassen zum Beispiel Ansätze zur Analyse von Daten mit fehlenden Werten (Ziegler et al., 2003), Ansätze zur Fallzahlplanung (Dahmen und Ziegler, 2004) oder Regressions- bzw. Deletionsdiagnostik (Ziegler et al., 1995; Preisser und Qaqish, 1996). Deletionsdiagnostik wird zur Identifikation von Ausreißern oder einflussreichen Probanden genutzt. Allerdings haben diese Erweiterungen bisher wenig Anwendung gefunden, unter anderem, weil nur wenig Software vorhanden ist.

Insbesondere wurden GEE bisher selten benutzt, um kontrollierte klinische Studien zu untersuchen. Das Ziel dieses Kapitels ist daher zweigeteilt. Erstens soll gezeigt werden, dass die Anwendung der GEE in einer Studie mit wiederholten Messungen eine interessante Alternative oder zumindest eine Ergänzung zur Standardanalyse, die lediglich die letzte Nachuntersuchung und möglicherweise eine Adjustierung zur Baseline-Messung beinhaltet, sein kann. Zweitens soll gezeigt werden, dass die Regressionsdiagnostik als Sensitivitätsanalyse die GEE Analyse begleiten sollte.

Zur Illustration wird eine doppelblinde, Placebo kontrollierte, randomisierte, multizentrische Studie (SB-LOT-Studie) erneut analysiert. In dieser Studie soll der Ödem protektive Effekt des vasoaktiven Medikamentes SB-LOT, einem Coumarin/Troxerutin Kombinationspräparat, bei Patienten, die an einer chronischen venösen Insuffizienz nach Kongestionsverminderung der Beine leiden, untersucht werden. Die Primäranalyse war eine Baseline adjustierte Kovarianzanalyse (ANCOVA) zwischen den beiden Behandlungsgruppen (Vanscheidt et al., 2002). Eine Sekundäranalyse mittels GEE, die eine Differenz des Anstiegs des durchschnittlichen Volumens der Unterschenkel in den beiden Gruppen aufzeigen soll, wird durchgeführt. In Anlehnung an Kapitel 3 wird gezeigt, dass die plausibelste Arbeitskorrelationsstruktur die autoregressive Korrelationsstruktur für diese Studiensituation ist.

Nachdem die einflussreichsten Patienten mithilfe der Regressionsdiagnostik von der Analyse ausgeschlossen wurden, ändert sich die Gesamtaussage nicht. Zur gleichen Zeit verbessert sich die Güte der Anpassung, die mithilfe von Halbnormalabbildungen untersucht wurde, erheblich.

Zusammenfassend lässt sich sagen, dass die Anwendung der GEE in einer longitudinalen klinischen Studie durchaus eine Alternative zur Standardanalyse, die üblicherweise nur die letzte Nachuntersuchung betrachtet, darstellen. Allerdings sollten die Techniken der Regressionsdiagnostik die GEE Analyse begleiten, um als Sensitivitätsanalyse zu dienen.

4.1. Die SB-LOT-Studie

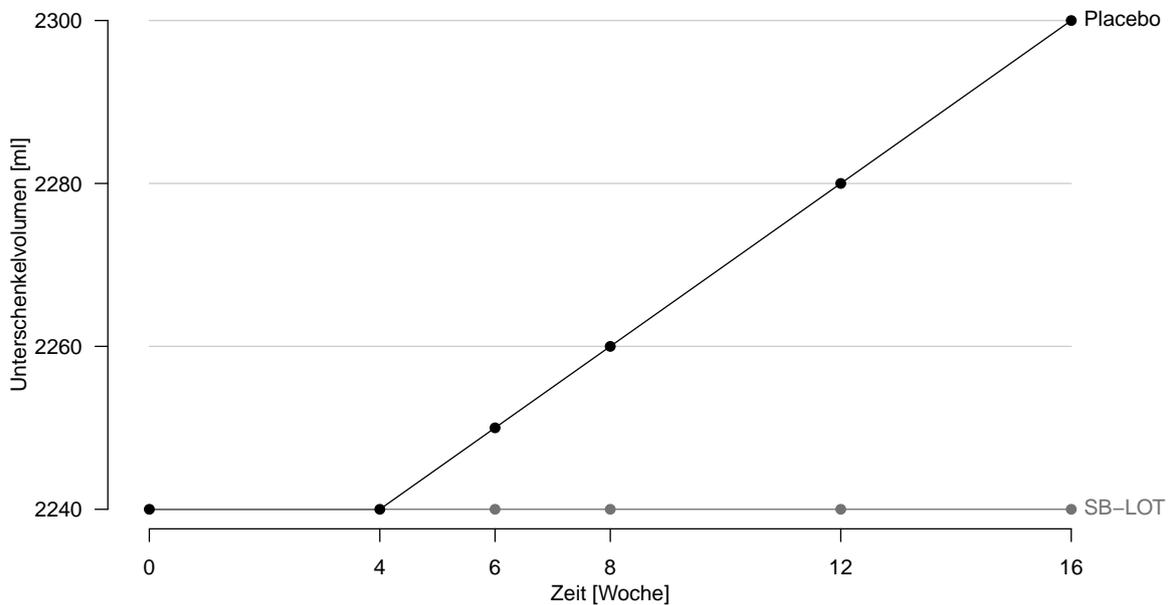


Abbildung 4.1: Erwarteter Verlauf der SB-LOT-Studie. Zum Zeitpunkt der Randomisierung (Woche 0) werden in beiden Gruppen identische Unterschenkelvolumina erwartet. Die medizinischen Kompressionsstrümpfe werden nach Woche 4 entfernt. Nach Woche 4 sollte das Unterschenkelvolumen in der Placebo-Gruppe linear ansteigen während es in der SB-LOT-Gruppe konstant bleiben sollte.

Zur Veranschaulichung des Nutzens von GEE und der Regressionsdiagnostik wird ein Parallelgruppen-Design mit wiederholten Messungen herangezogen. In dieser doppelblinden, Placebo kontrollierten, randomisierten Multicenter-Studie wird der Ödem protektive Effekt von SB-LOT bei Patienten, die unter chronischer venöser Insuffizienz nach Kongestionsverminderung der Beine leiden, untersucht (Vanscheidt et al., 2002). Zu Beginn der Studie (Baseline) wurden 226 Patienten in eine der beiden Gruppen randomisiert. Die Behandlungsgruppe erhielt für die ersten vier Wochen medizinische Kompressionsstrümpfe und SB-LOT (90 mg Coumarin und 540 mg Troxerutin pro Tag), die Kontrollgruppe erhielt in diesen Wochen ebenfalls Kompressionsstrümpfe und ein Placebo-Präparat. Nach den ersten vier Wochen wurden die Kompressionsstrümpfe im Rahmen der ersten Nachuntersuchung bei allen Patienten entfernt. In den folgenden 12 Wochen der Studie erhielten die Patienten der Behandlungsgruppe SB-LOT und der Kontrollgruppe das Placebo-Präparat. Nachdem die Kompressionsstrümpfe entfernt wurden, erwarteten die Prüfer einen Anstieg des Unterschenkelvolumens bei den fol-

Tabelle 4.1: Die SB-LOT-Studie: Unterschenkelvolumina (in ml) im Verlauf der Studie.

SB-LOT (Patienten pro Gruppe $n_1=113$)						
Woche	0 ¹	4	6	8	12	16
Mittelwert	2283,3	2235,5	2258,4	2256,7	2261,8	2245,8
Standardabweichung	365,0	371,7	373,4	363,5	369,9	355,4
Minimum	1409,0	1386,0	1406,0	1404,0	1415,5	1430,5
Maximum	3443,5	3408,5	3435,0	3408,0	3410,0	3405,0
Placebo (Patienten pro Gruppe $n_2=113$)						
Woche	0 ¹	4	6	8	12	16
Mittelwert	2260,2	2245,7	2290,1	2294,7	2294,5	2295,9
Standardabweichung	359,0	350,6	350,8	353,4	356,4	354,1
Minimum	1204,0	1194,5	1234,5	1242,0	1299,5	1246,0
Maximum	3205,0	3395,5	3304,5	3398,5	3394,0	3268,5

¹ unter Verwendung von Kompressionsstrümpfen

Tabelle 4.2: Schätzer der Korrelationskoeffizienten der SB-LOT-Daten, zusammengefasst über beide Behandlungsgruppen.

Nachbeobachtung	Woche 4	Woche 6	Woche 8	Woche 12	Woche 16
Woche 4	1,0000	0,9594	0,9539	0,9593	0,9403
Woche 6	0,9594	1,0000	0,9973	0,9973	0,9971
Woche 8	0,9539	0,9973	1,0000	0,9973	0,9973
Woche 12	0,9593	0,9973	0,9973	1,0000	0,9973
Woche 16	0,9403	0,9971	0,9973	0,9973	1,0000

genden Nachuntersuchungen in der Placebo-Gruppe, während das Unterschenkelvolumen der Probanden in der Behandlungsgruppe möglichst konstant bleiben sollte (siehe Abb. 4.1).

Die Patienten wurden fünf Mal nachuntersucht: 4, 6, 8, 12 und 16 Wochen nach der Initiierung der medikamentösen Therapie. Somit sind die Intervalle zwischen den Nachuntersuchungen nicht immer gleich groß. Der primäre Endpunkt zur Bestimmung der Wirksamkeit war das Unterschenkelvolumen, welches mit Hilfe der Plethysmographie gemessen wurde. Die Primäranalyse war eine Baseline adjustierte Kovarianzanalyse (ANCOVA) der Differenz zwischen des Unterschenkelvolumens bei der letzten Untersuchung und des Volumens zu Beginn der Studie. Eine Auswertung soll unter Intention-to-treat erfolgen. Tabelle 4.1 zeigt das Unterschenkelvolumen (in ml) im Verlauf der Studie. Wie bereits erwartet, ist die Korrelation zwischen den Beobachtungen zu verschiedenen Zeitpunkten hoch (siehe Tabelle 4.2).

Die Sekundäranalyse, also die Analyse mittels der GEE, zielt darauf ab, einen Unterschied im Anstieg des Unterschenkelvolumen zu detektieren, indem die Eigenschaft der wiederholten Messungen der Daten zunutze gemacht wird. Da die Korrelation zwischen den Messungen verschiedener Zeitpunkte als Nuisance-Parameter betrachtet wird, repräsentiert die erneute Analyse der Daten einen typischen Aufbau für eine GEE1 Analyse.

4.2. Verallgemeinerte Schätzgleichungen

Für die erneute Analyse der SB-LOT-Daten wird das Unterschenkelvolumen zum Zeitpunkt t als abhängige Variable y_{it} genutzt. Da die Zeitpunkte x_{it}^{Time} nicht äquidistant sind, werden sie mit den entsprechenden Werten der Nachuntersuchungswoche, beginnend mit $t = 0$ für den Beginn der Studie, kodiert. Die Gruppen werden mit 1 für die Placebo-Gruppe und 0 für die SB-LOT-Gruppe kodiert. Da Teilnehmer beider Gruppen in den ersten vier Wochen Kompressionsstrümpfe tragen, wird erwartet, dass das Unterschenkelvolumen in beiden Gruppen gleich ist (siehe Abbildung 4.1). Allerdings sollte der Anstieg in den darauffolgenden Wochen der Studienteilnahme unterschiedlich sein. Daher ist das Modell für den Mittelwert gegeben durch

$$\mathbb{E}(y_{it}) = \beta_1 + \beta_2 x_{it}^{\text{Treat}} + \beta_3 x_{it}^{\text{Time}} + \beta_4 x_{it}^{\text{Treat}} \cdot x_{it}^{\text{Time}}, \quad t = 1, \dots, 5, \quad (4.1)$$

wobei

$$x_{it}^{\text{Treat}} = \begin{cases} 1 & \text{für Placebo} \\ 0 & \text{für SB-LOT} \end{cases} \quad \text{für alle } t$$

sowie

$$\begin{aligned} x_{i1}^{\text{Time}} &= 4, \\ x_{i2}^{\text{Time}} &= 6, \\ x_{i3}^{\text{Time}} &= 8, \\ x_{i4}^{\text{Time}} &= 12 \text{ und} \\ x_{i5}^{\text{Time}} &= 16. \end{aligned}$$

4.3. Wahl einer sinnvollen Arbeitskorrelationsstruktur

Verglichen werden sollen die Unabhängigkeitsstruktur, die autoregressive Struktur erster Ordnung, die austauschbare Struktur, sowie die 1-Abhängiger Struktur und der Ansatz keine Struktur, also kein Vorwissen einzubringen.

4.4. Regressionsdiagnostik

Ungewöhnliche Daten, im Sinne von Ausreißern, können die Anpassung eines Regressionsmodells stark beeinflussen. Die Regressionsdiagnostik ermöglicht die Identifizierung solcher Datenpunkte. Ungewöhnliche Werte in der abhängigen Variable werden üblicherweise als Ausreißer bezeichnet, solche in den unabhängigen Variablen werden auch Hebelpunkte genannt. Der Effekt dieser Beobachtungen wird am Besten ermittelt, indem die Änderung der Parameterschätzungen, sobald eine Beobachtung von der Analyse ausgeschlossen wurde, analysiert wird. Die dazugehörigen Statistiken werden auch Deletionsdiagnostiken genannt. In der GEE Analyse müssen zwei Arten von Ausreißern und Hebelpunkten betrachtet werden. Zum einen können sie auf Clusterebene auftreten, zum anderen auf Ebene einer einzelnen Beobachtung innerhalb eines Clusters.

Regressionstechniken, die im linearen Modell (Cook und Weisberg, 1982; Belsley et al., 2004) oder in GLM (Thomas und Cook, 1989) verwendet wurden, wurden für die GEE erweitert. Statistiken zur Güte der Anpassung (engl. goodness of fit, GOF) für die logistischen GEE wurden in einer Reihe von Artikeln vorgeschlagen, eine Zusammenfassung ist gegeben in Evans und Li (2005). Ziegler et al. (1995), Preisser und Qaqish (1996) oder Ziegler und Armingier (1996) haben verschiedene Statistiken zur Identifikation von Ausreißern mithilfe von Residuen oder der modifizierten Cook-Statistik und zur Identifikation von Hebelpunkten mithilfe einer modifizierten Hutmatrix vorgeschlagen. Venezuela et al. (2007) entwickelten einen graphischen Ansatz zur Ausreiseridentifikation mithilfe von Halbnormalabbildungen mit simulierten Einhüllenden. Venezuela et al. (2011) beschrieben Maßzahlen des lokalen Einflusses bei GEE. Eine weitere Residuenabbildung, um die Güte der Anpassung eines Modells basierend auf GEE zu messen, wurde von Oh et al. (2008) vorgestellt.

4.4.1. Residuen

Als Residuum bezeichnet man die Differenz zwischen der vorhergesagten Größe und dem Messwert. In der Literatur werden verschiedene Residuen genutzt.

Auch wenn die Kovarianzmatrix Ω_i fehlspezifiziert ist, können Residuen genutzt werden, um systematische Variationen aufzudecken, die durch einen oder mehrere Regressoren verursacht wurden. In Abbildungen werden die Residuen gegen eine unabhängige Variable oder die vorhergesagten Werte aufgetragen, um Muster zu erkennen und gegebenenfalls Ausreißer auszuschließen.

Das geschätzte ordinäre Residuum $\hat{\varepsilon}_i$ ist gegeben durch

$$\hat{\varepsilon}_i = \mathbf{y}_i - \hat{\boldsymbol{\mu}}_i \quad \text{und} \quad \hat{q}_i = \hat{\varepsilon}_i' \hat{\varepsilon}_i \quad (4.2)$$

Dieses multivariate Residuum und seine quadratische Form sind von Interesse, wenn die Fehler ε_{it} , also die Elemente von $\boldsymbol{\varepsilon}_i$, korreliert sind.

Das standardisierte Residuum s_{it} von y_{it} und seine quadratische Form q_i^s lassen sich über

$$s_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\hat{\sigma}_{it}}} = \frac{\hat{\varepsilon}_{it}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_i]_{tt}}} \quad \text{und} \quad \hat{q}_i^s = \hat{\varepsilon}_i' \hat{\boldsymbol{\Sigma}}_i \hat{\varepsilon}_i \quad (4.3)$$

berechnen, wobei $[\hat{\boldsymbol{\Sigma}}_i]_{tt}$ das t -te Diagonalelement von $\hat{\boldsymbol{\Sigma}}_i$ beschreibt. Sollte die Kovarianzmatrix korrekt spezifiziert sein, kann das α -Quantil der Standardnormalverteilung zur Detektion von Ausreißern in der abhängigen Variable genutzt werden, wenn das geschätzte Residuum \hat{s}_{it} normalverteilt ist. Zur Identifizierung von Clustern kann die χ^2 -Verteilung mit T Freiheitsgraden angewendet werden.

Für die Definition der modifizierten studentisierten Residuen wird die Cluster spezifische Hutmatrix

$$\mathbf{H}_i = \hat{\boldsymbol{\Sigma}}_i^{-1/2} \hat{\mathbf{D}}_i (\hat{\mathbf{D}}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\mathbf{D}}_i)^{-1} \hat{\mathbf{D}}_i' \hat{\boldsymbol{\Sigma}}_i^{-1/2'} \quad (4.4)$$

benötigt, wobei $\hat{\boldsymbol{\Sigma}}_i^{1/2}$ die Wurzel aus $\hat{\boldsymbol{\Sigma}}$ bezeichnet, welche $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^{1/2} \hat{\boldsymbol{\Sigma}}^{1/2'}$ erfüllt. Für ein einfaches lineares Regressionsmodell reduziert sich die modifizierte Hutmatrix \mathbf{H} zu $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Das modifizierte studentisierte Residuum s_{it}^* von y_{it} und seine quadrati-

sche Form q_i^{s*} sind gegeben durch

$$s_{it}^* = \frac{\hat{\varepsilon}_{it}}{\sqrt{[\hat{\Sigma}_i]_{tt} (1 - [\mathbf{H}_i]_{tt})}} \quad \text{und} \quad q_i^{s*} = \hat{\varepsilon}_i' \left[\hat{\Sigma}_i^{1/2} (\mathbf{I} - \mathbf{H}_i) \hat{\Sigma}_i^{1/2} \right]^{-1} \hat{\varepsilon}_i, \quad (4.5)$$

wobei $[\mathbf{H}_i]_{tt}$ das t -te Diagonalelement der Hutmatrix \mathbf{H}_i bezeichnet.

In vielen statistischen Paketen sind das Pearson Residuum und das standardisierte Pearson Residuum implementiert, die durch

$$\hat{\varepsilon}_{it}^P = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{h(\mu_{it})}} \quad \text{und} \quad \hat{\varepsilon}_{it}^{Ps} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{h(\mu_{it})(1 - [\mathbf{H}_i]_{tt})}}, \quad (4.6)$$

definiert sind, wobei h die Varianzfunktion darstellt.

Halbnormalabbildungen mit simulierten Einhüllenden können sowohl der Identifikation von Ausreißern als auch zur Feststellung der Güte der Anpassung dienen, auch wenn die Verteilung der Residuen unbekannt ist. Der Algorithmus, der von Venezuela et al. (2007) vorgeschlagen wurde, um eine Halbnormalabbildung mit simulierten Einhüllenden im GEE Kontext zu konstruieren, unterscheidet sich vom Originalalgorithmus von Atkinson (1981). Genauer gesagt basieren die Berechnungen in Atkinson (1981) auf einem Jack-knife-Residuum, während Venezuela et al. (2007) die Nutzung des allgemeinen standardisierten Residuum vorschlagen. Der sich daraus ergebende Algorithmus von Venezuela et al. (2007) lautet wie folgt:

1. Schätze das GEE Modell mit den Originaldaten.
2. Für $k = 1$ bis K (Standardeinstellung $K = 25$):
 - (a) Für jeden Cluster i , $i = 1, \dots, n$, simuliere einen $(T \times 1)$ Vektor abhängiger Variablen unter Verwendung des geschätzten Mittelwertvektors und der Korrelationsmatrix basierend auf dem Modell, das den Originaldaten angepasst wurde.
 - (b) Für die simulierten abhängigen Variablen passe unter Verwendung derselben Kovariaten dasselbe Modell an.
 - (c) Schätze die standardisierten Residuen s_{it} und ordne diese. Sei s_{mk} der m -te geordnete absolute Wert des standardisierten Residuums, welches zum k ten

Durchlauf gehört, $m = n \cdot T$.

- (d) Berechne das Minimum der kleinsten absoluten Werte der Residuen und bezeichne es mit $|s|_{1,k}$. Analog berechne den Median $|s|_{0,5,k}$ und das Maximum $|s|_{n \cdot T,k}$ der absoluten Werte der Residuen.
3. Trage $|s|_{1,k}, |s|_{2,k}, \dots, |s|_{n \cdot T,k}$ und die geordneten absoluten Werte des standardisierten Residuum der Originaldaten gegen die Halbnormal-Scores

$$\Phi^{-1} \left(\frac{l + nT - \frac{1}{8}}{2nT + \frac{1}{2}} \right)$$

auf, wobei $\Phi(\cdot)$ die kumulative Verteilungsfunktion der Standardnormalverteilung ist.

Eine wichtige Annahme dieses Ansatzes sollte erwähnt werden. Obwohl die GEE1 wegen einer möglichen Missspezifikation der Arbeitskorrelationsmatrix angewendet werden, werden die standardisierten Residuen berechnet und die simulierten Einhüllenden unter der Annahme, dass das gesamte Modell korrekt spezifiziert ist, generiert.

Um Individuen mit großem Einfluss auf die Halbnormalabbildung mit simulierten Einhüllenden zu identifizieren, wird eine Leave-one-out-Strategie angewendet. Im Speziellen wurde in jeder Iteration eine Person aus der Analyse entfernt und eine Halbnormalabbildung mit simulierten Einhüllenden nach obigem Algorithmus mit den übriggebliebenen $n - 1$ Individuen erstellt. Individuen wurden als einflussreich auf die Abbildung bezeichnet, wenn die Entfernung dieses Individuums eine starke Veränderung im Verlauf der Halbnormalabbildung mit simulierten Einhüllenden nach sich zog. Um eine objektive Bewertung der einzelnen Halbnormalabbildungen mit simulierten Einhüllenden zu gewährleisten, wurden alle Abbildungen von zwei erfahrenen, verblindeten Gutachtern unabhängig voneinander beurteilt.

4.4.2. Einflussreiche Punkte und Hebelpunkte

Die geschätzten Regressionskoeffizienten sind wichtig für die Interpretation der Ergebnisse. Daher ist von Interesse, inwiefern sich diese ändern, sobald der i -te Cluster gelöscht wurde. Sei $\hat{\beta}_j$ der j -te Parameterschätzer des Modells, welches unter Verwendung der kompletten Daten geschätzt wurde. Sei $\hat{\beta}_{j(i)}$ der j -te Parameterschätzer des Modells,

welches geschätzt wurde, nachdem Cluster i von der Analyse ausgeschlossen wurde. Dann ist

$$\text{DFBetaC}_{j(i)} = \hat{\beta}_j - \hat{\beta}_{j(i)} \quad (4.7)$$

der Effekt, den Cluster i auf die Schätzung von β_j hat.

Beobachtungen, die den angepassten Wert \hat{y} stark beeinflussen, heißen Hebelpunkte. Im Fall der GEE hat die $(nT \times nT)$ -dimensionale Block diagonale modifizierte Hutmatrix den Rang p . Daher sollte $\text{spur}(\mathbf{H}_i)$ idealerweise p/n gleichen und der Einfluss $[\mathbf{H}_i]_{tt}$ der t -ten Beobachtung von Cluster i sollte idealerweise $p/(nT)$ sein. Wenn die Werte \mathbf{H}_i oder $[\mathbf{H}_i]_{tt}$ mindestens zwei oder dreimal höher sind als der jeweils ideale Wert, ist die Vorhersage für \mathbf{y}_i und y_{it} stark durch die dazugehörigen Daten beeinflusst. Zur Identifizierung von Hebelpunkten kann eine Linienabbildung, auch Indexabbildung genannt, benutzt werden. Hier ist der Hebel jedes Clusters gegen seine Cluster ID aufgetragen, um zu erkennen, welches Cluster einen besonders großen Hebel und somit Einfluss haben könnte.

Cluster mit einem starken Einfluss auf die Parameterschätzung können durch die Berechnung der modifizierten Cook-Statistik identifiziert werden. Speziell bedeutet dies, dass ein Konfidenzellipse für β basierend auf der robusten Varianzmatrix gegeben ist durch $(\hat{\beta} - \beta)' \hat{\mathbf{C}}^{-1} (\hat{\beta} - \beta)$. Zur Berechnung der Cook-Statistik wird der unbekannte wahre Parameter β durch seinen Schätzer aus dem Modell, in dem der i -te Cluster nicht betrachtet wurde, ersetzt. Daher ist die modifizierte Cook-Statistik gegeben durch

$$\widehat{\text{Cook}}_i = \frac{1}{p} (\hat{\beta} - \hat{\beta}_{(i)})' \hat{\mathbf{C}}^{-1} (\hat{\beta} - \hat{\beta}_{(i)}) \quad (4.8)$$

Ein Cluster wird als einflussreich bezeichnet, wenn die dazugehörige Cook-Statistik einen gewissen Schwellenwert, zum Beispiel $\chi_p^2(1 - \alpha)$, überschreitet. Es ist zu beachten, dass die Definition der Cook-Statistik aus Gleichung (4.8) nicht auf der korrekten Spezifikation der wahren Kovarianzmatrix $\mathbf{\Omega}_i$ beruht. Aus diesem Grund ziehe ich diese Definition von Ziegler und Arminger (1996) der Definition von Hammill und Preisser (2006) und Venezuela et al. (2007) vor, die die Modell basierte Kovarianzmatrix, das heißt die Fisher-Informationsmatrix \mathbf{A} , anstelle von \mathbf{C} als Gewicht benutzen. Weitere Regressionsdiagnostiken wurden in Jung (2008) vorgestellt.

Die Berechnung der verschiedenen regressionsdiagnostischen Verfahren braucht viel Rechenzeit. Daher sind in den vergangenen Jahren verschiedene Algorithmen vorgeschlagen worden, um die Geschwindigkeit der Berechnung zu erhöhen (Wei und Fung, 1999; Preisser und Perin, 2007; Preisser et al., 2008). Einige häufig verwendete regres-

Tabelle 4.3: Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit der autoregressiven Arbeitskorrelationsmatrix erster Ordnung für SB-LOT gegen Placebo. S.E.: Standardfehler; KI: Konfidenzintervall.

Parameter	Schätzer	S.E.	95%-KI	<i>p</i> -Wert
Regressionskonstante β_1	2241,25	35,62	[2171,44; 2311,06]	<0.0001
Behandlung β_2	4,08	45,65	[-91,28; 99,44]	0,9332
Zeit β_3	0,00	0,90	[-1,75; 1,76]	0,9992
Behandlung \times Zeit β_4	2,64	1,21	[0,27; 4,99]	0,0288

sionsdiagnostische Verfahren sind in wenigen Standardpaketen vorhanden. Diese umfassen auch SAS/IML Makros (Preisser und Garcia, 2005; Hammill und Preisser, 2006) und PROC GENMOD (SAS ver. 9.2).

4.5. Ergebnisse

Zunächst werden die Ergebnisse der GEE Analyse dargestellt. Anschließend wird auf die Verbesserungen durch die Regressionsdiagnostik eingegangen.

4.5.1. Die Standard GEE Analyse

Im ersten Schritt wurden die SB-LOT-Daten unter Verwendung der AR(1) Arbeitskorrelationsstruktur, wie von Wang und Carey (2003) vorgeschlagen, analysiert (siehe Tabelle 4.3). Analog zum ANCOVA Modell von Vanscheidt et al. (2002) zeigte die Intention-to-treat-Analyse unter Verwendung der GEE eine Überlegenheit von SB-LOT über dem Placebo, da der Parameter β_4 auf dem 5%-Niveau signifikant war ($p = 0,0288$). Die Differenz im Unterschenkelvolumen zwischen SB-LOT und Placebo stieg um 2,64 ml pro Woche (95%-Konfidenzintervall 0,27-4,99 ml pro Woche). Dies führte am Ende der Beobachtungszeit, also nachdem die medizinischen Kompressionsstrümpfe 12 Wochen nicht mehr getragen wurden, zu einer Gesamtdifferenz von 31,68 ml (95%-Konfidenzintervall: 3,24-59,88 ml).

Eine Annahme in den Standard GEE Paketen ist, dass die Nachuntersuchungen für Arbeitskorrelationsstrukturen mit zeitlich spezifischer Struktur in äquidistanten Zeitspannen unternommen werden sollen (siehe auch Kapitel 3 oder Ziegler und Vens (2010)).

Tabelle 4.4: Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit verschiedenen Arbeitskorrelationsmatrizen für SB-LOT gegen Placebo. S.E.: Standardfehler; KI: Konfidenzintervall; QIC: Quasi Likelihood Informationskriterium.

Korrelation ¹	$\hat{\beta}_4$	S.E.	95%-KI	p -Wert	$\hat{\rho}^2$	QIC
Unabhängig	2,51	1,27	[0,02; 5,00]	0,0484	-	1145,5576
Autoregressiv	2,64	1,21	[0,27; 5,00]	0,0288	0,9771	1145,2996
Austauschbar	2,51	1,27	[0,02; 5,00]	0,0484	0,9598	1145,5776
1-Abhängiger	3,06	1,60	[-0,07; 6,19]	0,0557	0,5773	1158,3245
Unstrukturiert	-4,26	3,72	[-11,55; 3,03]	0,2519	Tabelle 4.5	1148,2517

¹ gewählte Arbeitskorrelationsstruktur

² Schätzer des Arbeitskorrelationskoeffizienten

Tabelle 4.5: Schätzer für die Korrelationskoeffizienten für die SB-LOT-Daten unter Verwendung der unstrukturierten Arbeitskorrelationsmatrix.

Nachbeobachtung	Woche 4	Woche 6	Woche 8	Woche 12	Woche 16
Woche 4	1,0000	0,9710	0,9482	0,9530	0,9210
Woche 6	0,9710	1,0000	0,9850	0,9924	0,9568
Woche 8	0,9482	0,9850	1,0000	0,9891	0,9576
Woche 12	0,9530	0,9924	0,9891	1,0000	0,9828
Woche 16	0,9210	0,9568	0,9576	0,9828	1,0000

In der SB-LOT-Studie wurden die Nachuntersuchungen allerdings zeitlich nicht äquidistant vorgenommen. Besonders waren die Zeitpunkte 4 und 6 Wochen als auch die Beobachtungen nach 8 und 12 Wochen benachbart. Daher ist ein interessanter Aspekt, ob das AR(1) Modell eine sinnvolle Wahl als Arbeitskorrelationsstruktur in diesem Beispiel ist. Es sei erwähnt, dass die ungleichen Zeitspannen nicht die Konsistenz der Parameterschätzungen betreffen, aber eventuell deren Effizienz verändern können. Da alle unabhängigen Variablen in einem Cluster entweder konstant innerhalb des Clusters (Absolutglied, Behandlung) oder Mittelwert balanciert (Zeitpunkt, Zeitpunkt \times Behandlung) sind, sollten die IEE sehr ähnlich den GEE sein (siehe Kapitel 3 oder Ziegler und Vens (2010)). Weitere Arbeitskorrelationsmatrizen könnten ebenso gewählt werden, zum Beispiel die 1-Abhängige oder die austauschbare Arbeitskorrelationsstruktur. Es ist zu beachten, dass die Interpretation aller dieser Arbeitskorrelationsstrukturen aufgrund der nicht äquidistanten Zeitspannen zwischen den Beobachtungen schwierig werden könnte.

Tabelle 4.4 zeigt die Ergebnisse verschiedener GEE Modelle mit unterschiedlichen Arbeitskorrelationsmatrizen. Gezeigt sind die Schätzer für die Steigung β_4 . Tabelle 4.5 zeigt die geschätzten Korrelationskoeffizienten für das Modell mit der unstrukturier-

ten Arbeitskorrelationsmatrix. Die Schätzer für β_4 waren identisch für IEE und GEE mit einer austauschbaren Arbeitskorrelationsstruktur. Die Schätzer der Modelle mit unabhängiger, AR(1) und austauschbarer Arbeitskorrelationsstruktur waren ähnlich. Das GEE mit der 1-Abhängiger Korrelationsmatrix erzielte den höchsten QIC Wert. Die QIC Werte der IEE und der GEE mit austauschbarer Arbeitskorrelationsmatrix waren identisch, weil die Parameter der beiden Modelle identisch waren. Beide Arbeitskorrelationsstrukturen führten zu ein und demselben Modell, obwohl der Schätzer des Arbeitskorrelationskoeffizienten in der austauschbaren Struktur auf 0,9598 geschätzt wurde und bei den IEE auf 0 gesetzt wird. Diese Ergebnisse wurden aufgrund der theoretischen Ergebnisse von Mancl und Leroux (1996) und Wang und Carey (2003) sowie der Ausführungen in Kapitel 3, Spiess und Hamerle (1996) und Ziegler und Vens (2010) erwartet.

Die Ergebnisse für das Modell mit der unstrukturierten Arbeitskorrelationsmatrix unterschieden sich stark von den anderen. Zum Beispiel stimmte der Parameterschätzer nicht mit den theoretischen Erwartungen für die Steigung β_4 (siehe Abbildung 4.1) und der deskriptiven Statistik (siehe Tabelle 4.1) überein. Diese Ergebnisse sind daher nicht verlässlich und könnten durch eine schlechte Kondition der robusten Kovarianzmatrix erklärt werden. Der kleinste QIC Wert wurde für das Modell der AR(1) Arbeitskorrelationsmatrix beobachtet. Dieses Modell scheint die sinnvollste Wahl zu sein. Die IEE könnten auch eine gute Wahl sein. Beide Ergebnisse bekräftigen die Erkenntnisse aus der Literatur (Wang und Carey, 2003; Dahmen und Ziegler, 2004; Ziegler und Vens, 2010) und Kapitel 3.

4.5.2. Regressionsdiagnostik

Die Mittelwertstruktur für die SB-LOT-Daten (Gleichung 4.1) wurde so gewählt, dass der Clusterhebel H_i identisch für alle Personen einer Behandlungsgruppe war. Große Werte $[H_i]_{tt}$ weisen darauf hin, dass x_{it} einen starken Einfluss auf den angepassten Wert \hat{y}_{it} hat (siehe zum Beispiel Venezuela et al. (2007)). Allerdings ist diese Information nur nützlich, wenn x_{it} einige diskrete oder kontinuierliche Kovariaten beinhaltet, was in dieser Anwendung nicht der Fall ist. Daher ist die Analyse der Hebelpunkte nicht von Interesse für diesen Datensatz. Des Weiteren wurde die Varianzfunktion $h(\mu_{it})$ so definiert, dass sie für alle Beobachtungen y_{it} identisch ist. Da die modifizierte Hutmatrix H_i identisch für alle Personen innerhalb einer Behandlungsgruppe war, waren verschiedene Residuen ebenfalls identisch. Daher sind nur die Betrachtungen der ordinären Residuen und modifizierten Cook-Statistik weiter erwähnenswert.

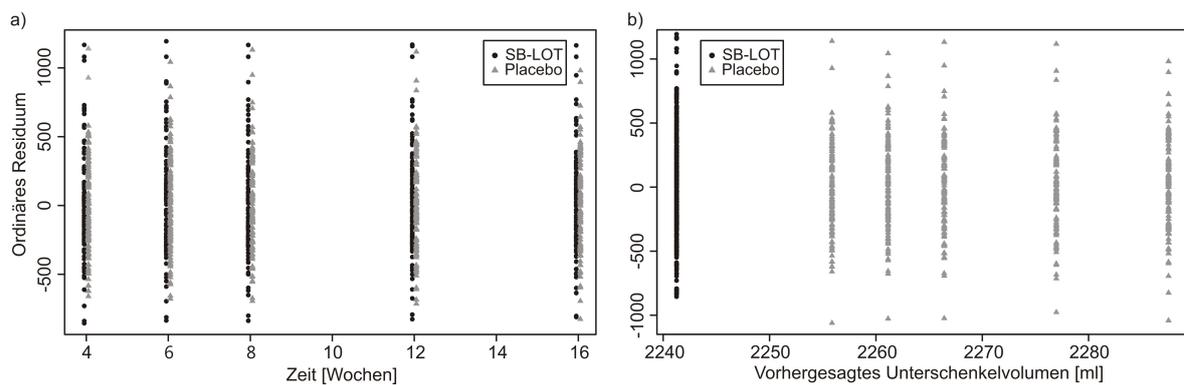


Abbildung 4.2: Residuenabbildungen. a) Ordinäre Residuen über die Zeit (in Wochen). b) Ordinäre Residuen aufgetragen gegen die vorhergesagten Werte des Unterschenkelvolumens (in ml).

Abbildung 4.2 zeigt die ordinären Residuen über die Zeit und gegen die vorhergesagten Werte für das Unterschenkelvolumen aufgetragen. Beide Abbildungen lassen keine spezifische Struktur in den Residuen erkennen. Daher scheint das Modell einigermaßen gut spezifiziert.

Abbildung 4.3 zeigt die Cluster-Cook-Statistik für alle Probanden. Sechzehn Personen hatten eine Cluster-Cook-Statistik, die 0,01 überstieg - der Schwellenwert entspricht in etwa $3p/(n \cdot T)$. Die größte Cluster-Cook-Statistik betrug 0,06. Der dazugehörige Patient hatte große Unterschenkelvolumen mit einem Maximum von 2712,0 ml. Der Proband wurde in die Behandlungsgruppe randomisiert. Die nächstgrößeren Clusterdistanzen betragen ungefähr 0,03. Analog zu dem Probanden mit der größten Cluster-Cook-Statistik hatten auch diese beiden Probanden erhöhte Unterschenkelvolumen mit einem Maximum von 2995,5 ml. Es folgen vier weitere Patienten mit Cluster-Cook-Statistiken zwischen 0,02 und 0,03. Drei von ihnen hatten große Unterschenkelvolumen (zwischen Minimum 3127,0 ml und 3435,0 ml Maximum). Weitere neun Probanden wiesen eine Cluster-Cook-Statistik zwischen 0,01 und 0,02 auf. Sie alle hatten diese erhöhten Cluster-Cook-Statistiken, weil sie entweder sehr hohe oder sehr niedrige Unterschenkelvolumen aufwiesen.

Obwohl diese 16 Studienteilnehmer einen großen Einfluss auf die Parameterschätzungen und vorhergesagten Werte hatten, änderte sich die statistische Signifikanz des Behandlungseffekts nicht, nachdem sie von den Analysen ausgeschlossen wurden (siehe Tabelle 4.6). Sogar der Punktschätzer blieb stabil (Änderung von 2,64 ml auf 2,65 ml) und signifikant ($p = 0,0010$). Ein großer Unterschied wurde jedoch in Bezug auf die Modellanpassung bei Betrachtung des QIC erreicht. Der Wert fiel von 1145,30 auf 1065,21, nachdem Patienten mit einer Cluster-Cook-Statistik $> 0,01$ ausgeschlossen wurden.

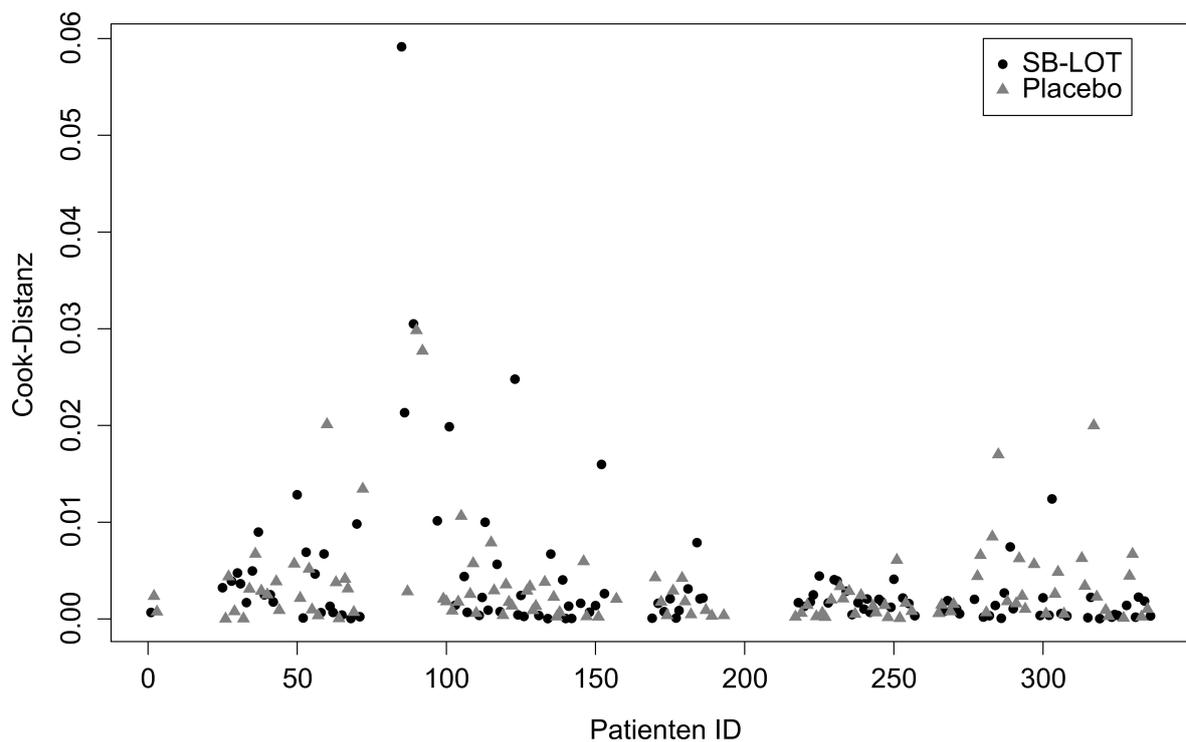


Abbildung 4.3: Cluster-Cook-Statistik für die SB-LOT-Studie unter Verwendung der autoregressiven Arbeitskorrelationsstruktur erster Ordnung.

Halbnormalabbildungen und simulierten Einhüllenden wurden erstellt, um die Güte der Anpassung festzustellen. Die Güte der Anpassung verbesserte sich mit der Anzahl der Probanden, die aufgrund ihrer Cluster-Cook-Statistik ausgeschlossen wurden (Abbildung 4.4). Jedoch hatten immer noch Patienten einen großen Einfluss auf die Halbnormalabbildung mit simulierten Einhüllenden (Abbildung 4.4c)). Um diese Patienten zu identifizieren, wurde ein Leave-one-out-Verfahren benutzt. Genauer gesagt wurde nach dem Ausschluss der Patienten mit einer Cluster-Cook-Statistik $> 0,01$, zusätzlich ein Proband pro Durchlauf ausgeschlossen und die dazugehörige Venezuela Abbildung bewertet.

Mithilfe zweier unabhängiger, verblindeter Gutachter wurden drei Patienten als einflussreich auf die Venezuela Abbildung bezeichnet (Abbildung 4.5). Der erste dieser Patienten (Patienten ID 181) hatte eine Differenz von 695 ml zwischen dem minimalen und maximalen Unterschenkelvolumen. Dieses war bei Weitem die größte Differenz zwischen minimalem und maximalem Unterschenkelvolumen in dieser Studie. Der zweite Proband (Patienten ID 38) hatte ein sehr geringes minimales Unterschenkelvolumen. Alle gemessenen Unterschenkelvolumen dieses Patienten lagen noch unterhalb des ersten Quartils aller minimalen Unterschenkelvolumen. Der dritte Patient (Patienten ID 41) lässt sich genau gegenteilig beschreiben. Hier lagen alle Unterschen-

Tabelle 4.6: Intention-to-treat-Analyse unter Verwendung der verallgemeinerten Schätzgleichungen mit autoregressiver Arbeitskorrelationsmatrix erster Ordnung für SB-LOT gegen Placebo. Es werden Ergebnisse für die Analysen gezeigt, in denen keine Patienten, Patienten mit einer Cluster-Cook-Statistik (CCS) $> 0,02$ und Patienten mit einer Cluster-Cook-Statistik $> 0,01$ ausgeschlossen wurden. S.E.: Standardfehler; KI: Konfidenzintervall; QIC: Quasi Likelihood Informationskriterium

Modell	n	$\hat{\beta}_4$	S.E.	95%-KI	p -Wert	QIC
Vollständige Daten	226	2,64	1,21	[0,27; 5,00]	0,0288	1145,30
CCS $> 0,02$ ausgeschlossen	219	2,31	0,94	[0,47; 4,15]	0,0139	1110,11
CCS $> 0,01$ ausgeschlossen	210	2,65	0,81	[1,06; 4,23]	0,0010	1065,21

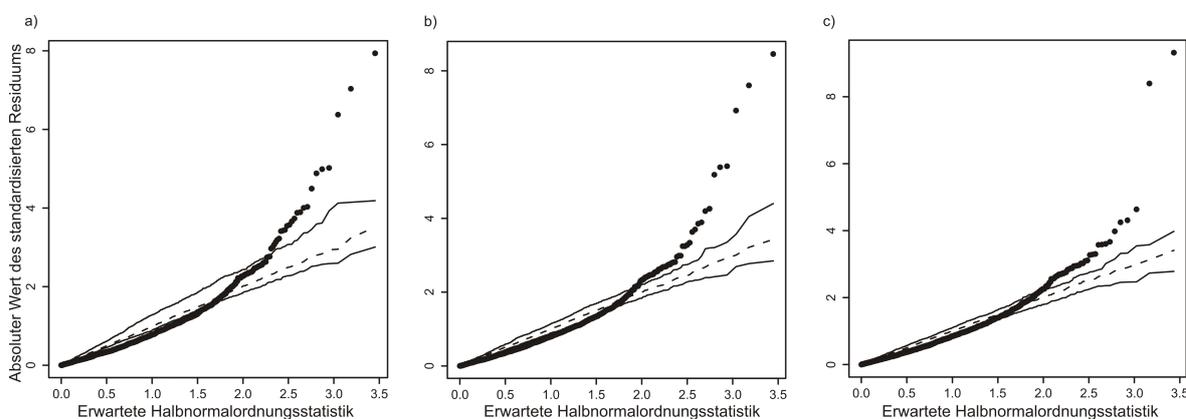


Abbildung 4.4: Halbnormalabbildungen mit simulierten Einhüllenden. a) Für den vollständigen Datensatz, b) für Patienten mit einer Cluster-Cook-Statistik $< 0,02$, c) für Patienten mit einer Cluster-Cook-Statistik $< 0,01$.

kelvolumen über dem dritten Quartil aller gemessenen Unterschenkelvolumen. Nach dem Ausschluss dieser drei Studienteilnehmer verbesserte sich die Halbnormalabbildung mit simulierten Einhüllenden stark, war aber noch nicht perfekt (Abbildung 4.6a) und kein offensichtlich einflussreicher Patient wurde in der Clusterstatistik deutlich (Abbildung 4.6b)). Es liegen immer noch viele Punkte außerhalb der Einhüllenden. Es sei hier darauf hingewiesen, dass die Einhüllenden unter der Annahme, dass das gesamte Modell inklusive der Arbeitskorrelationsmatrix korrekt spezifiziert ist, simuliert wurden. Man erwartet daher gar nicht, dass alle Punkte innerhalb der Einhüllenden liegen. Stattdessen ist es vorzuziehen, die Einhüllenden als ein graphisches Mittel zur Bewertung der Verbesserung der Anpassung zu betrachten. Entscheidend ist, dass der Steigungsparameter nach dem Ausschluss aller einflussreichen Patienten stabil blieb (im Durchschnitt 2,66 ml pro Woche, 95%-Konfidenzintervall: 1,07-4,25 ml pro Woche, $p = 0,0010$).

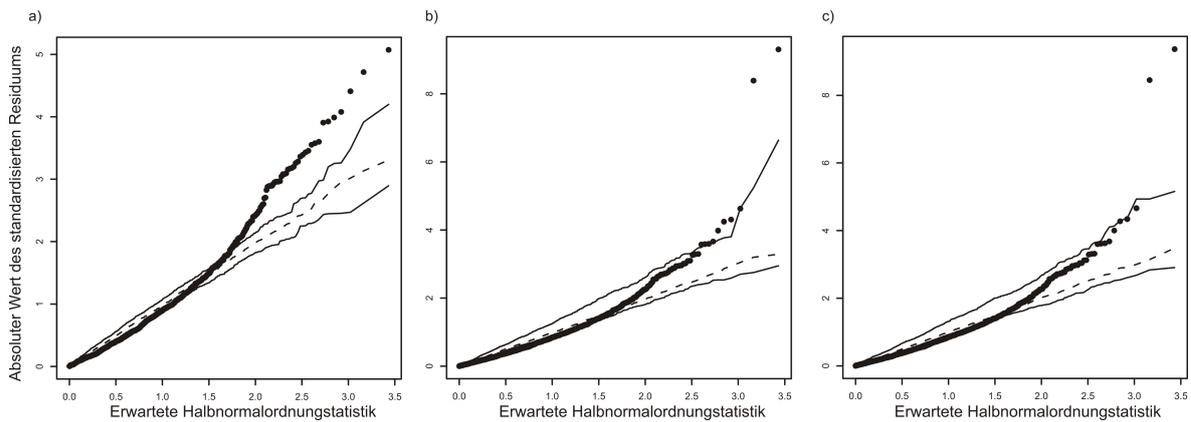


Abbildung 4.5: Halbnormalabbildungen mit simulierten Einhüllenden für Patienten mit einer Cluster-Cook-Statistik $< 0,01$ und zusätzlichem Ausschluss von Patient a) 181, b) 38, c) 41.

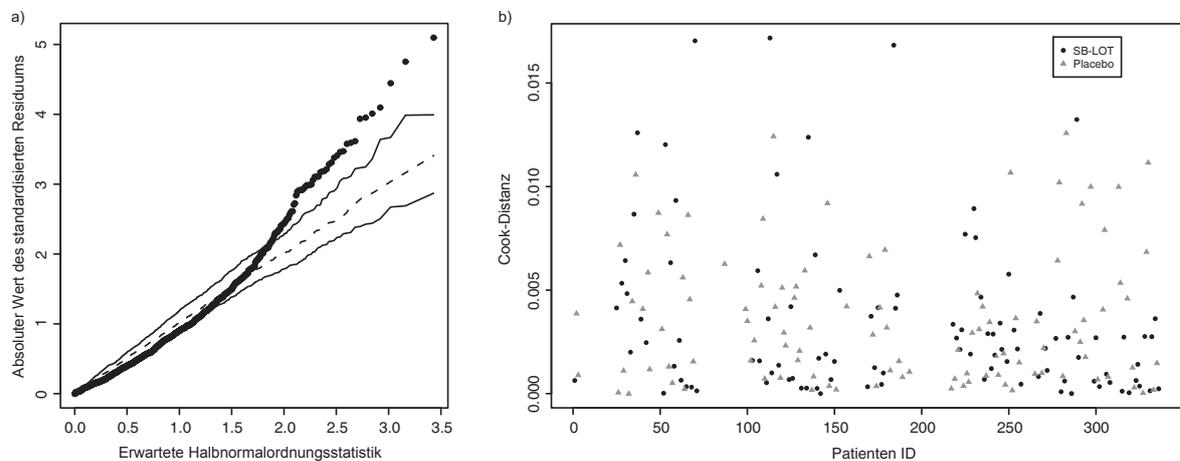


Abbildung 4.6: a) Halbnormalabbildung mit simulierten Einhüllenden und b) Cook-Distanzen nach dem Ausschluss von Patienten mit einer Cluster-Cook-Statistik $> 0,01$ und zusätzlichem Ausschluss der drei Patienten 181, 38 und 41.

4.6. Diskussion

Die Standardanalyse in einer longitudinalen klinischen Studie mit Parallelgruppen betrachtet für gewöhnlich nur den letzten Zeitpunkt. Daher ist die Standardanalyse häufig nur ein gewöhnliches Verfahren zum Vergleich zweier Gruppen. Für kontinuierliche abhängige Variablen sind häufig der t -Test oder der U -Test die Methoden der Wahl. Sollten Bereinigungen bezüglich der Baseline-Messungen durchgeführt werden, wird häufig eine Analyse der Kovarianz (ANCOVA) eingesetzt. Der letzte Ansatz betrachtet den letzten Nachbeobachtungszeitpunkt und den ersten Beobachtungswert zu Beginn der Studie, während die Beobachtungen dazwischen gänzlich ignoriert werden.

Die GEE stellen eine interessante Alternative zu diesen Standardverfahren dar. Hier können alle vorhandenen Zeitpunkte verwendet und somit mehr Informationen über den Verlauf der Studie in Betracht gezogen werden. Mit der GEE Analyse wurde in dieser Studie ein Schätzer für die Differenz des Unterschenkelvolumens zwischen den beiden Behandlungsgruppen auf einer wöchentlichen Basis geliefert. Diese Interpretation konnte nicht aus der ANCOVA von Vanscheidt et al. (2002) gezogen werden. Genauer gesagt, wurde die Überlegenheit des vasoaktiven Medikamentes über dem Placebo gezeigt. Der Effekt des vasoaktiven Medikamentes ist ein Unterschied im Unterschenkelvolumen von ungefähr 2,64 ml pro Woche (95%-Konfidenzintervall: 0,27–4,99 ml pro Woche). Dies führt zu einer Differenz von 31,68 ml (95%-Konfidenzintervall: 3,24–59,88 ml) am Ende der Studie.

Im GEE Ansatz ändert sich die Primäranalyse bezüglich eines kontinuierlichen Endpunkts. Anstelle einen Test für Differenzen in der Regressionskonstanten anzuwenden, ist der Test von Interesse in diesem Fall einer auf den Unterschied in der Steigung. Weil die interessierende Hypothese geändert wird, kann die Steigung im GEE Testansatz als eine erste Sensitivitätsanalyse beim Vergleich mit den Standardmethoden betrachtet werden.

Die autoregressive Arbeitskorrelationsstruktur erster Ordnung (AR(1)) wurde in der Literatur als Arbeitskorrelationsstruktur der Wahl bei longitudinalen Daten vorgeschlagen (Wang und Carey, 2003; Ziegler und Vens, 2010). Es wurde gezeigt, dass die AR(1) Arbeitskorrelationsstruktur zum Modell mit dem kleinsten QIC für die SB-LOT-Daten führt. Wie erwartet sind die IEE nur geringfügig unterlegen (Dahmen und Ziegler, 2004; Ziegler und Vens, 2010).

Eine zweite Sensitivitätsanalyse im Bereich der GEE kann auf der Deletionsdiagnostik basieren. Mit regressionsdiagnostischen Techniken kann eine Modellanpassung weiter optimiert werden. Des Weiteren können einflussreiche Beobachtungen identifiziert werden. Sollten sich die Ergebnisse nach der Entfernung von Hebelpunkten oder einflussreichen Clustern nicht stark verändern, kann man die Studienergebnisse als robust interpretieren. In dem vorliegenden Beispiel haben die regressionsdiagnostischen Ansätze gezeigt, dass sich die Hauptaussage der Studie nach Ausschluss einflussreicher Patienten von der Analyse nicht verändert. Dieses unterstreicht die Validität des Schlusses, dass ein signifikanter Ödem protektiver Effekt von SB-LOT besteht.

In dieser Studie wurden allgemeine deletionsdiagnostische Verfahren, wie Residualabbildungen und Cluster-Cook-Statistiken, verwendet und mit rechenzeitintensiven Halb-

normalabbildungen mit simulierten Einhüllenden kombiniert. Im Speziellen wurde der Leave-one-out-Ansatz auf die Halbnormalabbildungen mit simulierten Einhüllenden erweitert. Zwei Gutachter betrachteten alle Halbnormalabbildungen visuell und identifizierten Patienten mit einem starken Einfluss auf die Abbildung. Diese Patienten hatten ganz spezifische Charakteristiken wie die größte Veränderung, sehr kleine oder sehr große Werte in der Zielvariable Unterschenkelvolumen. Ein Nachteil des Ansatzes ist es, dass alle Halbnormalabbildungen einzeln visuell bewertet werden müssen. Eine automatische Prozedur zur Identifizierung einflussreicher Patienten auf die Abbildung oder die simulierten Einhüllenden würden die Arbeitsbelastung selbstverständlich deutlich senken. Die Analyse von Hebelpunkten war für die SB-LOT-Daten nicht von Bedeutung, da der Hebel für alle Studienteilnehmer innerhalb einer Behandlungsgruppe aufgrund der spezifischen Struktur der Designmatrix identisch war.

Zusammenfassend sollte festgehalten werden, dass die Anwendung der GEE in einer longitudinalen klinischen Studie eine Alternative zur Standardanalyse, die üblicherweise nur den letzten Nachbeobachtungszeitpunkt betrachtet, darstellt. Sowohl die GEE als auch regressionsdiagnostische Verfahren sollten eine Analyse solcher Daten begleiten, um mindestens als Sensitivitätsanalyse dienen zu können.

**5. Multivariable fraktionelle Polynome
für korrelierte abhängige Variablen
unter Verwendung von
verallgemeinerten Schätzgleichungen**

Bisher wurde die Anpassung der GEE durch die genaue Betrachtung der Wahl der Arbeitskorrelationsmatrix und der Einbeziehung regressionsdiagnostischer Verfahren verbessert. All diese Maßnahmen nutzen allerdings nur wenig, wenn zwischen den Einflussvariablen und der Zielvariablen ein nicht linearer Zusammenhang vorliegt, welcher von den GEE bisher nicht modelliert werden kann.

Ein Ansatz solche nicht linearen Zusammenhänge bei unkorrelierten Beobachtungen zu betrachten, sind fraktionale Polynome (engl. fractional polynomials, FPs, Royston und Altman, 1994). Fraktionale Polynome haben gegenüber konventionellen Polynomen, wie Alter und Alter², und Splines deutliche Vorteile (Royston und Sauerbrei, 2008). Auf der einen Seite sind Polynome niedriger Ordnung häufig in ihrer Vielzahl, verschiedene Kurvenverläufe abzubilden, begrenzt. Solche höherer Ordnung resultieren häufig in unerwünschten Kurvenverläufen, wie beispielsweise Oszillationen insbesondere an den Kurvenenden (Sauerbrei und Royston, 1999). Auf der anderen Seite sind Splines schwer zu interpretieren und tendieren zu einer Überanpassung der Daten (Sauerbrei et al., 2007). FPs kombinieren allerdings die wünschenswerten Eigenschaften beider Ansätze: Sie sind flexibel, wenig anfällig gegenüber lokalen Einflüssen und einfach den Daten anzupassen. Eine Vielzahl an flexiblen Kurvenverläufen steht zur Verfügung. Die Erweiterung von univariaten FPs zu einem multivariaten Modell führt dazu, dass FPs auch passend sind für multiple Regressionsmodelle, wie sie häufig, beispielsweise in der Epidemiologie, genutzt werden (Sauerbrei und Royston, 1999).

Die Vorteile der GEE, also die flexible Schätzung korrelierter Regressionsmodelle, und der Vorteil der FPs, das heißt die flexible Modellierung nicht linearer Zusammenhänge zwischen den Einflussvariablen und der Zielvariablen eines Modells, können zu einer Methode kombiniert werden, der GEE-MFP-Methode.

Die theoretischen Eigenschaften einer solchen Kombination wurden von Thompson et al. (2003) analysiert. Ein erster, allerdings sehr einfacher, Algorithmus für nur zwei unabhängige Variablen wurde von Cui et al. (2009) vorgestellt.

Hier soll ein Algorithmus vorgestellt werden, der eine beliebige Anzahl unabhängiger Variablen einbeziehen kann. Der Nutzer kann zudem für jede Einflussvariable einzeln wählen, ob sie linearen oder nicht linearen Einfluss haben kann bzw. haben soll. Diese Unterscheidung ist sinnvoll, da der Ansatz vorsieht, Variablen unterschiedlicher Skalierung gemeinsam zu analysieren. Somit ist es möglich, binären oder kategorialen Variablen einen linearen Einfluss zuzuordnen, da ein nicht linearer Einfluss an dieser Stelle wenig Sinn macht, während für kontinuierliche Einflussvariablen sowohl lineare

als auch nicht lineare Einflüsse Sinn machen können. Die Methode ist zudem um eine schrittweise Prozedur erweitert, sodass lediglich solche Variablen im Modell bleiben, die einen signifikanten Einfluss auf die Zielvariable haben.

Zusätzlich wird in diesem Kapitel gezeigt, dass die Wahl eines Kriteriums zur Güte der Anpassung zu einer Überanpassung führen kann. Insbesondere kommt es unter Verwendung der in Kapitel 3 vorgestellten Quasi Likelihood Informationskriterien (Pan, 2001a,b; Cui et al., 2009) zu unbefriedigenden Kurvenverläufen. Da diese Quasi Likelihood Informationskriterien auf Akaike's Informationskriterium (engl. Akaike's information criterion, AIC), bei welchem der Strafterm unabhängig von der Studiengröße ist, beruhen, kann eine Verbesserung der Log-Likelihood bei großen Fallzahlen leicht erreicht werden. In diesem Kapitel wird daher ein neues Kriterium vorgeschlagen, welches auf dem Bayesianischen Informationskriterium (engl. Bayesian information criterion, BIC) beruht (Schwarz, 1978).

Dieses Kapitel ist wie folgt strukturiert. Zunächst werden die FPs vorgestellt. Anschließend werden der neue GEE-MFP-Algorithmus und das neue Kriterium zur Bestimmung der Güte der Anpassung detailliert erläutert. Zum Schluss werden der Nutzen des Algorithmus und seine Vorzüge anhand der Daten der Framingham Heart Study, wie sie für den „Genetic Analysis Workshop 13“ zur Verfügung gestellt wurden (Cupples et al., 2003), und anhand einer zehnfach Kreuzvalidierung dargestellt.

5.1. Fraktionale Polynome

Es werden fraktionale Polynome erster Ordnung (FP1) und zweiter Ordnung (FP2) betrachtet und in der GEE-MFP-Methode zum Einsatz kommen. Ein FP1 für ein gegebenes Argument $x > 0$ ist gegeben durch x^s , wobei s aus einer Menge von Exponenten $\mathcal{S} = \{-2; -1; -0,5; 0; 0,5; 1; 2; 3\}$ ist und $x^0 = \ln(x)$ gilt (Royston und Altman, 1994). Auf ähnliche Weise wird ein FP2 gebildet. Ein FP2 für $x > 0$ ist gegeben durch den Vektor $x^{\mathbf{s}}$ mit $\mathbf{s} = (s_1, s_2)$ und

$$x^{\mathbf{s}} = x^{(s_1, s_2)} = \begin{cases} (x^{s_1}, x^{s_2}), & \text{if } s_1 \neq s_2 \\ (x^{s_1}, x^{s_1} \ln x), & \text{if } s_1 = s_2 \end{cases}.$$

Es existieren somit 8 FP1 Transformationen und 36 FP2 Transformationen für eine Va-

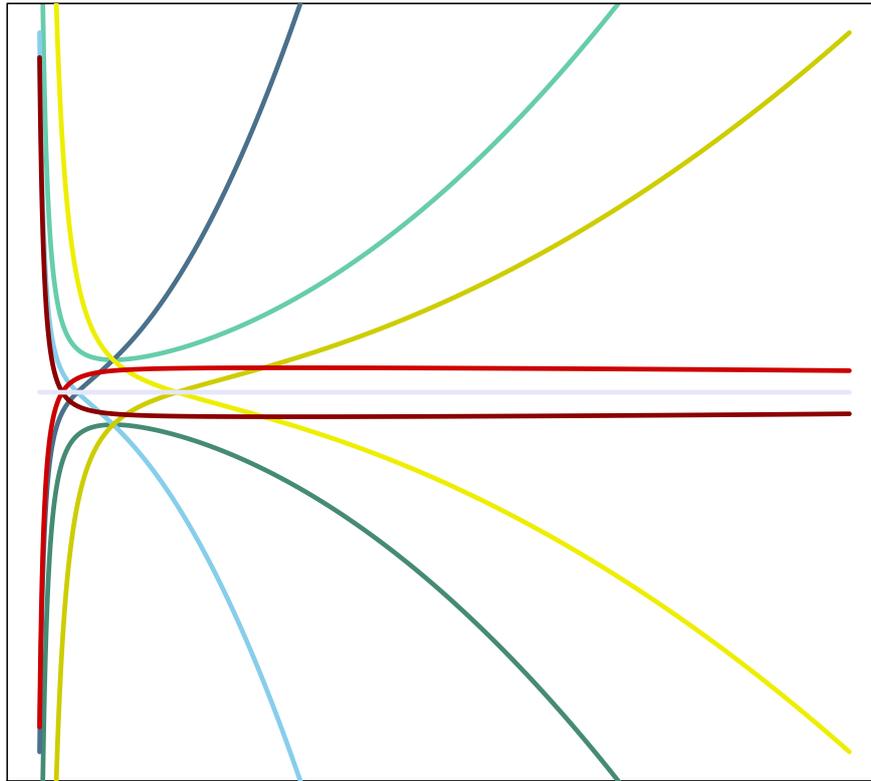


Abbildung 5.1: Darstellung von verschiedenen Kurvenverläufen fraktioneller Polynome zweiter Ordnung.

riable. Die FPs decken einen großen Bereich möglicher Kurvenverläufe ab. Beispiele für diese Verläufe sind in Abbildung 5.1 gegeben.

5.2. Die GEE-MFP-Methode

Hat man zum Ziel, für eine Menge unabhängiger Variablen die beste Transformation zu finden, stößt man schnell an Grenzen. Die Menge möglicher Modelle ist extrem groß. Es existieren bereits 1936 mögliche Modelle für zwei und ungefähr 165 Millionen mögliche Modelle für fünf unabhängige Variablen. Diese Anzahl wird wiederum deutlich erhöht, wenn alle Teilmengen von Variablen betrachtet werden sollen. Daher existiert eine große Notwendigkeit für einen effizienten Algorithmus, der auf einfache Weise das am besten angepasste Modell findet. Ein solcher Algorithmus wird hier vorgeschlagen.

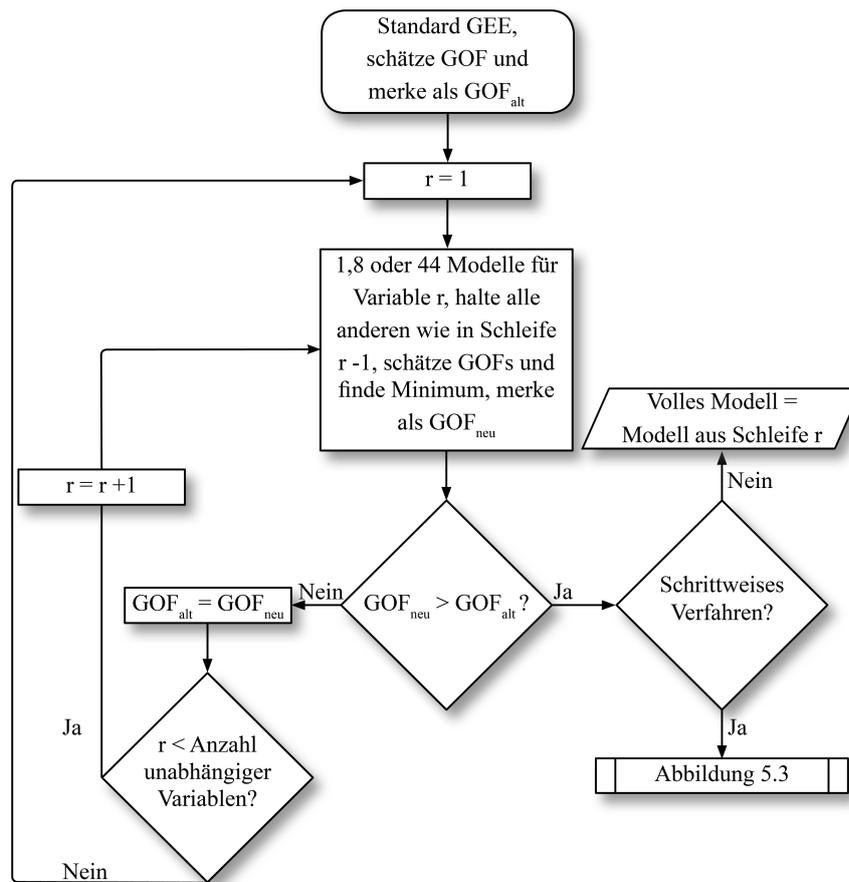


Abbildung 5.2: Modellbildungsprozess der GEE-MFP-Methode.

Zuerst muss für jede Variable spezifiziert werden, ob angenommen wird, dass der Einfluss der Variable linear ist, was vor allem bei binären oder kategoriellen Risikofaktoren Sinn macht, einer FP1 oder entweder einer FP1 oder FP2 Transformation folgt. Da der Algorithmus in SAS implementiert ist, müssen die üblichen Variablen der SAS PROC GENMOD, wie der Name des Datensatzes, die abhängige Variable und ihre Verteilung oder die Linkfunktion, die „class“ Variable, das „repeated subject“ und die Korrelationsstruktur spezifiziert werden. Zusätzlich muss ein Kriterium der Güte der Anpassung (engl. goodness of fit, GOF) gewählt werden. Die möglichen GOFs werden in Kapitel 5.3 näher beschrieben.

Danach beginnt der eigentliche Algorithmus. Die folgenden Schritte 1 bis 3 werden als Modellbildungsprozess bezeichnet (Abbildung 5.2).

1. Schätze ein vollständiges Modell, in dem alle Variablen linearen Einfluss haben.

Merke den zugehörigen GOF.

Setze Schleifenvariable $r = 1$

2. Transformation der unabhängigen Variablen.

Für jede Variable:

- a) Schätze entweder ein (die Variable hat linearen Einfluss), acht (die Variable kann einen FP1 transformierten Einfluss haben) oder 44 (die Variable kann entweder FP1 oder FP2 transformierten Einfluss haben) Modelle und die dazugehörigen GOF Werte, während die Einflüsse der anderen Variablen wie in dem Schritt davor gehalten werden.
- b) Suche den kleinsten GOF Wert.
- c) Behalte die Transformation, die den minimalen GOF Wert erzielte für die Transformationschritte der folgenden Variablen.

3. Nachdem alle Variablen transformiert wurden, vergleiche den letzten GOF Wert mit dem aus der Runde $r - 1$.

- a) Wenn $r = 1$, vergleiche den GOF Wert aus Schritt 2 mit dem aus Schritt 1.
 - i. Wenn beide GOF Werte gleich sind, stoppe und wähle das lineare Modell als finales Modell.
 - ii. Wenn der GOF Wert aus Runde r kleiner ist, setze $r = r + 1$ und springe zu Schritt 2.
- b) Wenn $r \neq 1$, vergleiche den GOF Wert aus Runde r Schritt 2 mit dem aus Runde $r - 1$ Schritt 2.
 - i. Wenn beide Werte gleich sind, stoppe und nehme das Modell aus Schritt $r - 1$ als das finale Modell.
 - ii. Wenn der GOF Wert aus Runde r kleiner ist, setze $r = r + 1$ and springe zu Schritt 2.

Wenn zusätzlich die mögliche schrittweise Prozedur gewählt wurde, geht der Algorith-

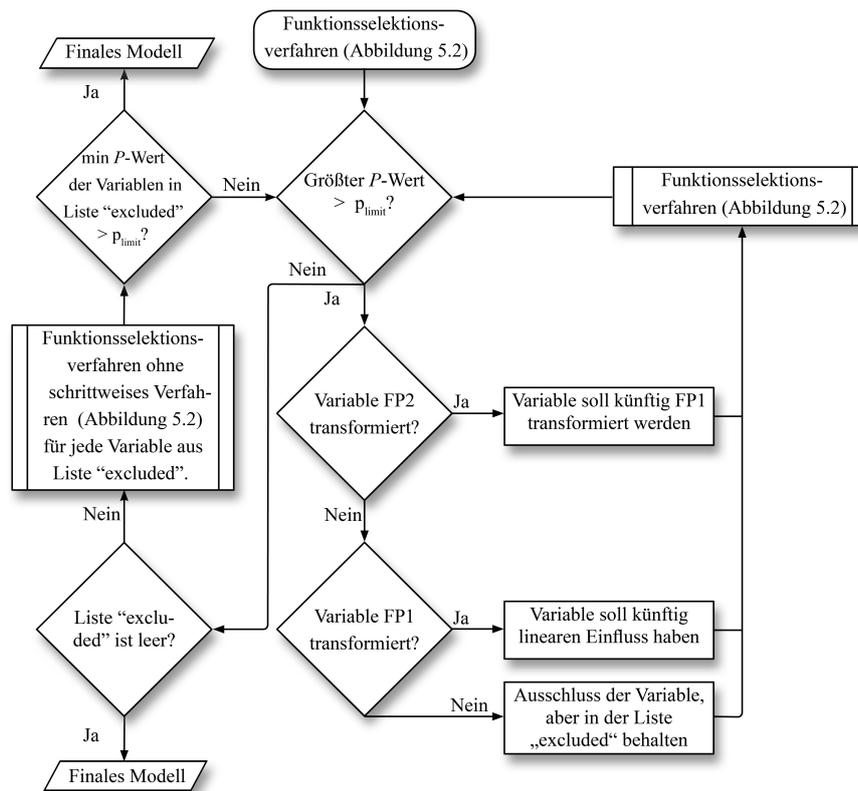


Abbildung 5.3: Schrittweise Prozedur der GEE-MFP-Methode.

mus weiter wie folgt vor (Abbildung 5.3):

1. Finde den schlechtesten p -Wert.
2. Ist dieser größer als ein gegebener Schwellenwert p_{limit} und
 - a) die zugehörige Variable war FP2 transformiert, zwingte sie im folgenden Modellbildungsprozess (Abbildung 5.2) einer FP1 Transformation zu folgen, danach gehe zu Schritt 1.
 - b) die zugehörige Variable war FP1 transformiert, zwingte sie im folgenden Modellbildungsprozess (Abbildung 5.2) einen linearen Einfluss zu haben, danach gehe zu Schritt 1.
 - c) die zugehörige Variable hatte linearen Einfluss, schliesse sie aus dem Modell aus, schreibe sie auf eine Ausschlussliste „excluded“, führe einen Modellbil-

dungsschritt durch und gehe danach zu Schritt 1.

3. Ist er nicht größer als ein gegebener Schwellenwert p_{limit} und

a) die Ausschlussliste ist nicht leer:

i. Nehme jede Variable der Ausschlussliste einzeln in das Modell und führe je einen Modellbildungsprozess durch.

ii. Suche den kleinsten p -Wert unter den im vorigen Schritt inkludierten Variablen. Ist dieser kleiner als ein gegebener Schwellenwert p_{limit} , gehe zu Schritt 1, ansonsten nehme das letzte Modell aus Schritt 2 als finales Modell.

b) die Ausschlussliste ist leer, stoppe und nehme das letzte Modell aus Schritt 2 als finales Modell.

5.3. Kriterien zur Güte der Anpassung und Selektion

Um die Güte der Anpassung zu messen, schlug Pan (2001a) das QIC vor (siehe Kapitel 3), welches eine Erweiterung des AIC darstellt. Beide Varianten des QIC, QIC und QIC_u , stehen in der vorgeschlagenen GEE-MFP-Methode als Kriterium zur Güte der Anpassung und Selektion, also als GOF-Kriterium, zur Verfügung. Bereits Cui et al. (2009) nutzten diese Maße, um die besten FP Transformationen in ihrem Ansatz zu finden. Allerdings ist bekannt, dass die Verwendung des AIC häufig zu Modellen mit großer Anzahl unabhängiger Variablen führt, was die Überanpassung der Daten zur Folge haben kann. Um diesem Problem aus dem Weg zu gehen, wird hier ein Quasi Likelihood Kriterium vorgeschlagen, das auf dem BIC beruht (Schwarz, 1978):

$$\text{BQIC}_u = -2\hat{p} Q(\hat{\mu}) + p \ln(N)$$

mit der Gesamtzahl der Beobachtungen der Studie $N = nT$. Der Strafterm des BQIC_u ist abhängig von der Fallzahl. Dies führt dazu, dass das BQIC_u zusätzliche Parameter im Modell stärker als das QIC bestraft, sobald die Fallzahl größer als acht ist. Somit ist es schwieriger, eine Variable mit einem FP2 transformierten Einfluss ins Modell aufzunehmen, was dazu führt, dass eine Überanpassung vermieden werden kann.

Tabelle 5.1: Die gewählten erklärenden Variablen und die zugehörigen Skalierungen.

Erklärende Variable	Skalierung
Alter (in Jahren)	dividiert durch 10
Körpermasseindex (in kg/m ²)	dividiert durch 10
Serum-Totalcholesterin-Konzentration (in mg/dl)	dividiert durch 100
Nüchtern-Serum-HDL-Cholesterin-Konzentration (in mg/dl)	dividiert durch 100
Familiengröße	dividiert durch 10
Zigaretten pro Tag (kategorisiert)	
Nüchtern-Serum-Glukose-Konzentration (in mg/dl, winsorisiert)	dividiert durch 10
Körperhöhe (in Inches)	dividiert durch 10
Nüchtern-Serum-Triglycerid-Konzentration (in mg/dl, winsorisiert)	dividiert durch 100
Alkoholkonsum (in g/d)	1 addiert, dividiert durch 100
Sex (männlich=0, weiblich=1)	
Bluthochdruckbehandlung (ja=1, nein=0)	

5.4. Die Daten der Framingham Heart Study

Um die GEE-MFP-Methode zu illustrieren, wurden die Daten der Framingham Heart Study mit Erlaubnis der Investigatoren erneut analysiert. Für die Analyse werden die Daten der Kohorte 2, wie sie für den Genetic Analysis Workshop 13 bereit gestellt wurden (Cupples et al., 2003), genutzt.

Die dichotome Zielvariable Bluthochdruck zu irgendeinem Zeitpunkt während der Studie (HBP_x; ja=1, nein=0) wird unter der Verwendung der erklärenden Variablen aus Tabelle 5.1 modelliert. Hergenommen werden die Werte der Einflussvariablen zu Beginn der Studie.

Die Familiengröße (engl. family size, FS) wurde, wie von Khan und Manzoor (2002) vorgeschlagen, in das Modell aufgenommen. Die Anzahl der Zigaretten, die pro Tag von dem Studienteilnehmer geraucht wurden (engl. cigarettes smoked per day, CPD), wird in 6 Kategorien unterteilt: Keine Zigaretten pro Tag (Nichtraucher), < 10, ≤ 20, ≤ 30, ≤ 40 und > 40 Zigaretten pro Tag. Diese Einteilung wurde gewählt, weil die meisten Raucher ihren Konsum in Zehnerschritten angegeben haben.

Sollte die Nüchtern-Serum-Glukose-Konzentration größer sein als ihr Mittelwert plus dreimal die Standardabweichung, in diesem Fall 180 mg/dl, wurde sie winsorisiert und auf diesen Wert gesetzt. Die Nüchtern-Serum-Triglycerid-Konzentration wurde auf ähnliche Weise ab einem Wert größer 475 mg/dl winsorisiert. Zusätzlich wurde 1 zu der Angabe der Gramm Alkohol pro Tag addiert, um Werte größer als null zu garantieren, wie es für die FP Transformationen nötig ist (Royston und Altman, 1994; Royston und Sauerbrei, 2008).

Alle Variablen außer Sex und Bluthochdruckbehandlung, beides binäre Einflussvariablen, wurden vor der Analyse wie von Royston und Sauerbrei (2008) vorgeschlagen durch die Anwendung folgender Schritte

$$lr = \log_{10} [\max(x) - \min(x)],$$

$$\text{scaling} = 10^{\text{sign}(lr) \cdot \text{int}(|lr|)},$$

$$x^* = \frac{x}{\text{scaling}}$$

skaliert.

Die Skalierungen für den Datensatz sind in Tabelle 5.1 gegeben.

5.5. Zehnfach Kreuzvalidierung

Um die Stabilität und Validität des Algorithmus und der Ergebnisse zu zeigen, wurde eine zehnfach Kreuzvalidierung durchgeführt. Hierzu wurden die zur Verfügung stehenden Daten in zehn Teile geteilt. Im Datensatz sind 1627 Personen aus 330 Familien enthalten. Da der Algorithmus die Korrelationsstruktur innerhalb der Familien berücksichtigt, wurden Familienmitglieder nicht in verschiedene Teildatensätze gelöst. In jedem Teildatensatz befanden sich daher 33 Familien. Die Anzahl der Personen in einem Teildatensatz konnte folglich stark variieren. Anschließend wurden jeweils neun der zehn Datensätze zusammengefasst und Modelle unter Verwendung aller Kriterien erstellt. Die Güte der erreichten zehn Modelle für jedes Kriterium wurde dann anhand des Teils untersucht, der nicht in den Modellbildungsprozess und die schrittweise Prozedur eingeschlossen war.

5.6. Ergebnisse

Die GEE-MFP-Methode wurde bei den Daten der Framingham Heart Study unter Verwendung des Logit-Links und einer austauschbaren Arbeitskorrelationsstruktur in Anlehnung an die Ergebnisse aus Kapitel 3 genutzt. Alle Variablen, außer diejenigen für das Geschlecht und die Bluthochdruckbehandlung, welche linearen Einfluss haben sollten, durften eine der 44 FP1 oder FP2 Transformationen annehmen. Zusätzlich sollte

Tabelle 5.2: Ergebnisse der vollständigen Modelle unter Verwendung von QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; Konz.: Konzentration.

Variable	Parameter	QIC oder QIC_u		BQIC_u		
		Schätzer	p -Wert	Parameter	Schätzer	p -Wert
Regressionskonstante		5,46	0,01		-4,49	$3,48 \cdot 10^{-4}$
Alter	age^{-1}	23,45	$2,16 \cdot 10^{-15}$	age	0,62	$< 10^{-15}$
	$\text{age}^{-0,5}$	-34,37				
Körpermasseindex	$\text{bmi}^{0,5}$	5,07	$< 10^{-15}$	bmi	1,49	$< 10^{-15}$
Serum-Totalcholesterin-Konz.	chol^{-2}	0,02	0,98	chol^{-2}	0,07	0,91
Nüchtern-Serum-HDL-Cholesterin-Konz.	hdl^2	2,23	$9,35 \cdot 10^{-6}$	hdl^2	2,18	$9,06 \cdot 10^{-6}$
Familiengröße	fs^3	$-1,80 \cdot 10^{-3}$	$1,72 \cdot 10^{-5}$	fs^3	$-1,80 \cdot 10^{-3}$	$5,15 \cdot 10^{-6}$
Alkoholkonsum	drkt^3	0,94	$4,69 \cdot 10^{-3}$	drkt	0,53	0,14
	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-1,93				
Körperhöhe	hgt^{-2}	-33,46	0,31	hgt^{-2}	-9,09	0,79
Nüchtern-Serum-Triglycerid-Konz.	$\ln(\text{tgw})$	0,43	$9,35 \cdot 10^{-4}$	$\ln(\text{tgw})$	0,43	$9,69 \cdot 10^{-4}$
Nüchtern-Serum-Glukose-Konz.	glucw^{-1}	-16,46	0,05	glucw^{-1}	-17,93	0,03
Zigaretten pro Tag	cpd_k^{-1}	-0,31	0,10	cpd_k^{-1}	-0,26	0,16
Geschlecht	sex	-0,30	0,12	sex	-0,40	0,04
Bluthochdruckbehandlung	hrx	0,66	0,19	hrx	0,69	0,16
		$\text{QIC}=1711,64$ oder $\text{QIC}_u = 1711,98$		$\text{BQIC}_u = 1786,86$		

auch die schrittweise Prozedur mit einem Schwellenwert $p_{\text{limit}} = 0,05$ angewendet werden. Die Daten wurden jeweils für die Kriterien QIC, QIC_u und BQIC_u analysiert und anschließend verglichen.

5.6.1. Die vollständigen Modelle unter Verwendung des kompletten Datensatzes

Die Ergebnisse für die vollen Modelle, also die Modelle nach dem Modellbildungsprozess, sind in den Tabellen 5.2 und A.1 dargestellt. Die Ergebnisse waren identisch für die Maße QIC und QIC_u (Tabellen 5.2 und A.1(a)). Der Wert für das QIC war 1711,64 und für das QIC_u 1711,98 (Tabelle 5.2).

Beispielsweise wurde die Variable für Alter FP2 transformiert. Es wurden dazu die Exponenten $s = (-1; -0,5)$ gewählt. Die Parameterschätzungen für age^{-1} und $\text{age}^{-0,5}$ waren 23,45 und -34,37 (Tabellen 5.2 und A.1(a)). Die zugehörigen 95%-Konfidenzintervalle sind in Tabelle A.1(a) gegeben. Die p -Werte für age^{-1} und $\text{age}^{-0,5}$ waren $6,55 \cdot 10^{-8}$ und $2,52 \cdot 10^{-10}$ (Tabelle A.1(a) vorletzte Spalte). Der kombinierte p -Wert für den Einfluss des Alters auf die Zielvariable Bluthochdruck ist in Tabelle 5.2 vierte Spalte oder in Tabelle A.1(a) letzte Spalte gegeben durch $2,16 \cdot 10^{-15}$.

Lediglich die dichotomen Variablen hatten linearen Einfluss nach dem Modellbildungs-

prozess. Während die Variablen für Alter und den Alkoholkonsum FP2 transformiert wurden, erreichten alle anderen Variablen als beste Anpassung eine FP1 Transformation.

Die Ergebnisse für das $BQIC_u$ unterschieden sich zu den vorherigen. Der Wert des $BQIC_u$ nach dem Modellbildungsschritt war 1796,75 (Tabellen 5.2 und A.1(b)). Keine Variable wurde FP2 transformiert. Die unabhängigen Variablen für Alter, den Körpermasseindex, das Geschlecht und die Bluthochdruckbehandlung hatten linearen Einfluss auf die Zielvariable im finalen Modell (Tabellen 5.2 und A.1(b)). Alle anderen Variablen wurden FP1 transformiert. Diejenigen Variablen, die sowohl unter der Verwendung des QIC, des QIC_u als auch des $BQIC_u$ FP1 transformiert waren, hatten in allen Modellen dieselben Transformationen (Tabelle 5.2 Spalten zwei und fünf).

5.6.2. Die Modelle nach der schrittweisen Prozedur unter Verwendung des kompletten Datensatzes

Tabelle 5.3: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des kompletten Datensatzes und QIC oder QIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; Konz.: Konzentration, Chol.: Cholesterin.

Variable	Parameter	QIC		Parameter	QIC_u	
		Schätzer	p-Wert		Schätzer	p-Wert
Regressionskonstante		4,36	0,01		14,91	$1,41 \cdot 10^{-5}$
Alter	age^{-1}	21,53	$< 10^{-15}$	$age^{-0,5}$	-19,80	$2,32 \cdot 10^{-15}$
	$age^{-0,5}$	-32,30		$age^{-0,5} \cdot \ln(age)$	-14,00	
Körpermasseindex	$bmi^{0,5}$	4,95	$< 10^{-15}$	$bmi^{0,5}$	4,94	$< 10^{-15}$
Serum-Totalcholesterin-Konz.						
Nüchtern-Serum-HDL-Chol.-Konz.	hdl^2	2,12	$4,81 \cdot 10^{-6}$	hdl^2	2,11	$5,67 \cdot 10^{-6}$
Familiengröße	fs^3	$-1,92 \cdot 10^{-3}$	$2,14 \cdot 10^{-6}$	fs^3	$-1,94 \cdot 10^{-3}$	$2,40 \cdot 10^{-6}$
Alkoholkonsum	$drkt^3$	1,07	$9,40 \cdot 10^{-4}$	$drkt^3$	1,08	$8,54 \cdot 10^{-4}$
	$drkt^3 \cdot \ln(drkt)$	-2,16		$drkt^3 \cdot \ln(drkt)$	-2,18	
Körperhöhe						
Nüchtern-Serum-Triglycerid-Konz.	$\ln(tgw)$	0,46	$9,02 \cdot 10^{-5}$	$\ln(tgw)$	0,45	$1,03 \cdot 10^{-4}$
Nüchtern-Serum-Glukose-Konz.	$glucw^{-1}$	-17,40	0,04	$glucw^{-0,5}$	-10,93	0,04
Zigaretten pro Tag						
Geschlecht	sex	-0,41	$5,93 \cdot 10^{-3}$	sex	-0,41	$5,61 \cdot 10^{-3}$
Bluthochdruckbehandlung						
		QIC=1709,12				QIC _u = 1709,64

Nach der schrittweisen Prozedur sank das QIC auf 1709,12 (Tabellen 5.3 und A.2(a)). Die unabhängigen Variablen für die Serum-Totalcholesterin-Konzentration, die Körper-

Tabelle 5.4: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des vollständigen Datensatzes und $BQIC_u$ als Kriterium der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall.

Variable	Parameter	Schätzer	p -Wert
Regressionskonstante		-2,63	0,14
Alter	age	0,64	$< 10^{-15}$
Körpermasseindex	bmi	1,47	$< 10^{-15}$
Serum-Totalcholesterin-Konzentration			
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	2,31	$4,01 \cdot 10^{-7}$
Familiengröße	fs ³	$-1,96 \cdot 10^{-3}$	$4,46 \cdot 10^{-7}$
Alkoholkonsum			
Körperhöhe			
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,48	$5,27 \cdot 10^{-5}$
Nüchtern-Serum-Glukose-Konzentration	glucw ^{-0,5}	-12,40	0,02
Zigaretten pro Tag			
Geschlecht	sex	-0,53	$2,08 \cdot 10^{-4}$
Bluthochdruckbehandlung			
	$BQIC_u = 1771,97$		

höhe, die Bluthochdruckbehandlung und das Rauchverhalten wurden aus dem Modell ausgeschlossen (siehe Tabellen 5.3 Spalte zwei oder A.2(a) Spalte zwei). Während der schrittweisen Prozedur änderte sich keine Transformation der verbleibenden Variablen im Vergleich zum vollen Modell (Tabellen 5.2 und 5.3 oder Tabellen A.1(a) und A.2(a)).

Die Ergebnisse unter Verwendung des QIC_u unterschieden sich nach der schrittweisen Prozedur zu denen unter der Verwendung des QIC (Tabelle 5.3 oder Tabellen A.2(a) und A.2(b)). Das QIC_u sank auf 1709,64 (Tabellen 5.3 oder A.2(b)). Dieselben Variablen wie unter der Verwendung des QIC wurden eliminiert (Tabelle 5.3 oder Tabellen A.2(a)–A.2(b)). Wiederum wurden die Variablen für Alter und Alkoholkonsum FP2 transformiert. Während sich die Transformation für Alter von $s = (-1; -0,5)$ im vollständigen Modell zu $s = (-0,5; -0,5)$ nach der schrittweisen Prozedur änderte, blieb sie für den Alkoholkonsum vor und nach der schrittweisen Prozedur identisch (Tabellen 5.2 und 5.3). Alle anderen unabhängigen Variablen, außer den binären, wurden FP1 transformiert. All diese Variablen, bis auf die Nüchtern-Serum-Glukose-Konzentration, behielten ihre Transformation wie im vollständigen Modell. Im vollständigen Modell zeigte die Nüchtern-Serum-Glukose-Konzentration einen inversen Einfluss. Nach der schrittweisen Prozedur änderte sich die Transformation zu -0,5.

Die Verwendung des $BQIC_u$ führte nach der schrittweisen Prozedur zu einem Wert von 1771,97 (Tabellen 5.4 oder A.2(c)). Zusätzlich zu den Variablen für die Serum-Totalcholesterin-Konzentration, die Körperhöhe, die Bluthochdruckbehandlung und das Rauchverhalten, die bereits in den Modellen unter Verwendung des QIC und des QIC_u

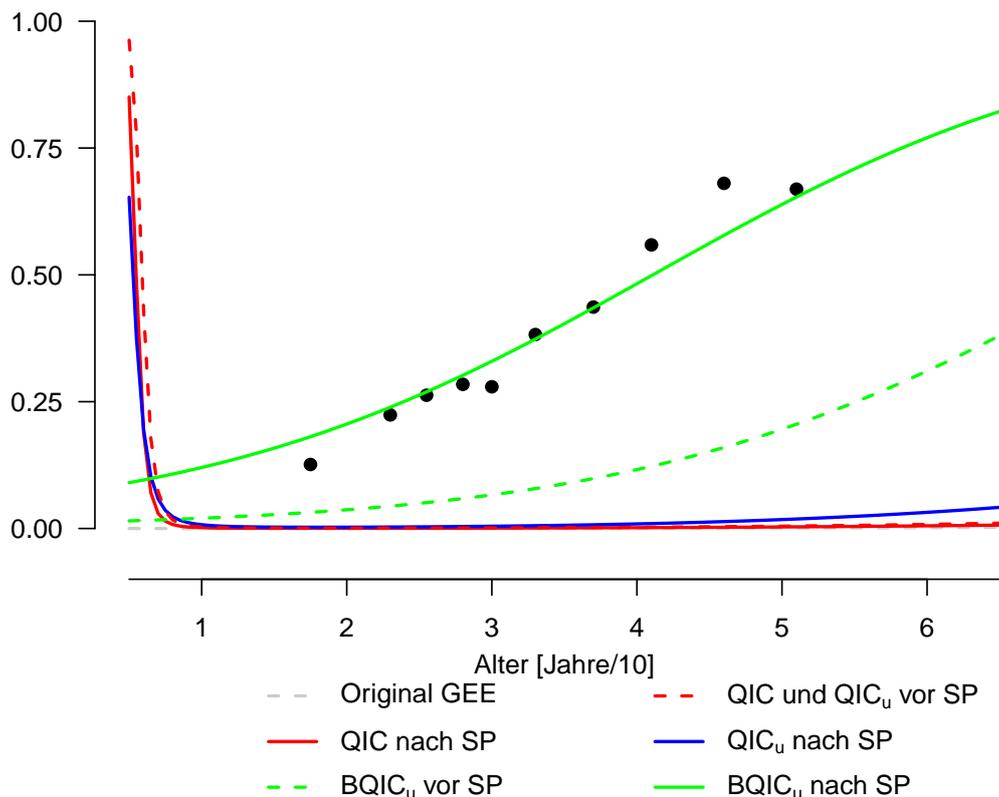


Abbildung 5.4: Vergleich der Anpassungen an das Alter unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün).

ausgeschlossen wurden, wurde auch die Variable für den Alkoholkonsum eliminiert (Tabellen 5.4 oder A.2(c)). Wie bereits vor der schrittweisen Prozedur war auch danach keine Variable FP2 transformiert. Lediglich für die Variable der Nüchtern-Serum-Glukose-Konzentration änderte sich der Einfluss von einem inversen Zusammenhang im vollständigen Modell hin zu -0,5 nach der schrittweisen Prozedur (Tabellen 5.2 und 5.4 oder A.1(b) und A.2(c)).

Um die Vorteile der Kombination aus GEE und MFP genauer zu beleuchten und den Nutzen des neu entwickelten Kriteriums BQIC_u zu verdeutlichen, soll im Folgenden die Anpassung der Modelle anhand der Einflussvariable für Alter betrachtet werden (Abbildung 5.4). Dazu wurden die Probanden zunächst anhand ihres Alters in Dezile eingeteilt. Im Anschluss daran wurde berechnet, wie viele Personen in jedem Dezil

an Bluthochdruck litten. Der Anteil derjenigen, die an Bluthochdruck erkrankt waren, wurde dann gegen das mittlere Alter in jedem Dezil aufgetragen (Abbildung 5.4). Eine gute Anpassung verlief sehr nah an diesen Punkten.

Die vorhergesagten Modelle unter der Verwendung von QIC oder QIC_u vor und nach der schrittweisen Prozedur waren jeweils sehr weit entfernt von einer guten Anpassung (Abbildung 5.4, rote und blaue Linien). All diese Modelle liefern die Information, dass es eine hohe Wahrscheinlichkeit gibt, in sehr jungem Alter an Bluthochdruck zu erkranken. Gleichzeitig zeigen die Vorhersagen, dass eine Wahrscheinlichkeit sehr nahe bei Null existiert, an Bluthochdruck ab einem Alter von zehn Jahren zu erkranken (Abbildung 5.4, rote und blaue Linien). Selbst das häufig genutzte Standard GEE, welches bei der Auswertung dieser Daten wohl zur Anwendung gekommen wäre, liefert als Vorhersage, dass es zu keinem Zeitpunkt eine Wahrscheinlichkeit gibt, an Bluthochdruck zu erkranken (Abbildung 5.4, graue gestrichelte Linie). Die Anpassung führte unter Verwendung von $BQIC_u$ zu einer deutlich besseren Vorhersage im vollen Modell vor der schrittweisen Prozedur (Abbildung 5.4, grüne gestrichelte Linie). Nach der schrittweisen Prozedur war die Anpassung zwischen Alter und der Wahrscheinlichkeit an Bluthochdruck zu erkranken nahezu perfekt (Abbildung 5.4, grüne durchgezogene Linie). Die Ergebnisse für die weiteren Variablen, welche nach der schrittweisen Prozedur in allen Modell blieben, waren ähnlich (Abbildungen A.1–A.6).

Im Weiteren wird das Modell nach der schrittweisen Prozedur unter Verwendung des $BQIC_u$ als Selektions- und Gütekriterium und dessen Interpretation betrachtet. Die positiven Schätzer für die unabhängigen Variablen für das Alter und den Körpermassenindex können wie gewohnt interpretiert werden. Auch die Schätzer für die Nüchtern-Serum-HDL-Konzentration und die Nüchtern-Serum-Triglycerid-Konzentration scheinen intuitiv richtig. Je höher der Wert einer dieser vier unabhängigen Variablen ist, desto höher ist die Wahrscheinlichkeit an Bluthochdruck zu erkranken (Abbildungen 5.4, A.1, A.2 und A.4). Der negative Schätzer für das Geschlecht ist ebenfalls sehr einfach zu interpretieren, da 1 für Frauen kodiert. Daher ist die Wahrscheinlichkeit für Studienteilnehmerinnen, an Bluthochdruck zu erkranken, niedriger als bei Studienteilnehmern (Abbildung A.5).

Der negative Schätzer für den Einfluss der Nüchtern-Serum-Glukose-Konzentration scheint auf den ersten Blick unpassend. Betrachtet man diesen aber gemeinsam mit der Transformation der zugrunde liegenden Variablen, welche $-0,5$ ist, erschließt sich ein passender Zusammenhang (Abbildung A.6). Eine entsprechende Kurve, die zum Exponenten $-0,5$ gehört, hätte zunächst sehr hohe Werte bei niedrigen Werten der un-

abhängigen Variablen. Durch das negative Vorzeichen des Schätzers kehrt sich dieses Verhalten allerdings um. Das bedeutet, dass kleine Werte der Nüchtern-Serum-Glukose-Konzentration auch niedrige Wahrscheinlichkeiten zur Folge haben, während hohe Konzentrationen auch hohe Wahrscheinlichkeiten nach sich ziehen. Dieses Beispiel verdeutlicht, dass, sollte eine Transformation von 1 verschieden sein, immer der Schätzer und die Transformation gleichzeitig betrachtet werden müssen, um sinnvolle Interpretationen herbeiführen zu können. Das negative Vorzeichen gemeinsam mit der kubischen Transformation des Einflusses der Familiengröße auf die Wahrscheinlichkeit an Bluthochdruck zu erkranken, spiegelt den Verlauf in den Daten der Framingham Heart Study wider (Abbildung A.3).

5.6.3. Die zehnfach Kreuzvalidierung

Die vollständigen Modelle der zehnfach Kreuzvalidierung erzielten häufiger FP2 Transformationen der Einflussvariablen, wenn QIC oder QIC_u als Selektions- und Gütekriterien benutzt wurden (Tabelle A.3 und Tabellen A.5–A.14). In keinem Fall wurden FP2 Transformationen als beste Anpassung für eine der Einflussvariablen gewählt, wenn $BQIC_u$ als Selektions- und Gütekriterium genutzt wurde (Tabelle A.3 und Tabellen A.5–A.14). Zusätzlich waren die gewählten Transformationen unter Verwendung von $BQIC_u$ stabiler über alle zehn Validierungen hinweg. Das bedeutet, dass für eine Einflussvariable durchschnittlich weniger verschiedene Transformationen zur Anwendung kamen, wenn $BQIC_u$ genutzt wurde (Tabelle A.3).

Nach der schrittweisen Prozedur wurden unter Verwendung von $BQIC_u$ als Selektions- und Gütekriterium sparsamere Modelle als beste Modelle gewählt (Tabellen A.4 und A.15–A.24). Sparsam bedeutet in diesem Zusammenhang, dass niedrig-dimensionalere Transformationen für die einzelnen Einflussvariablen, häufig ein linearer Zusammenhang, und insgesamt weniger Einflussvariablen pro Modell gewählt wurden. Wiederrum waren die gewählten Transformationen zwischen den Validierungen stabiler, wenn $BQIC_u$ zur Anwendung kam (Tabelle A.4).

Des Weiteren war der durchschnittliche Anteil korrekt vorhergesagter Zielvariablen in den Modellen nach der schrittweisen Prozedur unter Verwendung von $BQIC_u$ am höchsten (73,18%, empirisches 95%-Konfidenzintervall: 70,45% bis 77,53%). Im Vergleich dazu lag der Anteil bei QIC_u bei 72,71% (empirisches 95%-Konfidenzintervall: 68,65% bis 77,60%) und bei QIC bei 72,87% (empirisches 95%-Konfidenzintervall: 69,14% bis 77,01%) (Tabelle A.25). Obwohl der durchschnittliche Anteil korrekt vorhergesagter

Zielvariablen in den vollständigen Modellen unter Verwendung des QIC mit 73,44% im Vergleich zu der Verwendung des $BQIC_u$ mit 73,18% leicht höher lag, waren die empirischen 95%-Konfidenzintervalle für $BQIC_u$ am schmalsten und der minimale Anteil der korrekt vorhergesagten Zielvariablen war für $BQIC_u$ sowohl in den vollständigen Modellen als auch in denen nach der schrittweisen Prozedur am höchsten (Tabelle A.25).

5.7. Diskussion

Wie sollte der funktionelle Zusammenhang zwischen verschiedenen Risikofaktoren und einer abhängigen Variable bei wiederholten Messungen oder in Familienstudien modelliert werden? Hier wurde vorgeschlagen FP und GEE zu vereinen. FP werden dabei genutzt, um die funktionelle Form des eventuell nicht linearen Zusammenhangs festzulegen, und die GEE berücksichtigen die Korrelationsstruktur. Dieser Ansatz wurde bereits zuvor betrachtet (Cui et al., 2009) und seine theoretischen Eigenschaften von Thompson et al. (2003) studiert. Bisher fehlte es allerdings an einem einfachen, stabilen und rechenkompatiblen Algorithmus. Der vorgeschlagene Ansatz wählt die am besten passenden GEE unter einer großen Anzahl möglicher Modelle aus. Der Algorithmus kann mit einer beliebigen Anzahl unabhängiger Variablen umgehen und ist fähig, binäre, kategoriale und stetige Einflussvariablen gemeinsam in einem Rahmen zu analysieren. Zusätzlich ist eine schrittweise Prozedur implementiert, welche die einflussreichsten Variablen auf die Zielvariable identifiziert. In vorhergehenden Arbeiten wurde das QIC genutzt, um über die funktionelle Form des Zusammenhangs zu entscheiden (Cui et al., 2009). In dieser Arbeit wurde ein neues Kriterium, das $BQIC_u$, vorgeschlagen, um diese Auswahl zu treffen. Es wurde gezeigt, dass das neue Kriterium zu einfacheren, besser angepassten Modellen führt.

Um mögliche Anwendungen zu zeigen, wurden die Daten der Framingham Heart Study noch einmal analysiert. Dabei waren die vollständigen Modelle unter Verwendung des QIC oder QIC_u als Kriterium der Güte der Anpassung oder Selektion identisch. Die Verwendung von $BQIC_u$ resultierte in einem sparsamen vollständigem Modell, in dem keine Variable FP2 transformiert vorlag. Insbesondere wurde für das Alter ein linearer Einfluss gewählt als das $BQIC_u$ als Kriterium der Güte der Anpassung und der Selektion genutzt wurde, während es FP2 transformiert vorlag, als die QIC oder QIC_u Ansätze genutzt wurden.

Obwohl die $BQIC_u$ Werte größer als die des QIC oder QIC_u waren, beinhalteten diese

Modelle weniger Variablen und weniger Transformationen im Vergleich zu QIC und QIC_u . Der Unterschied in den Werten wird durch die unterschiedlichen Strafterme erzeugt. Die Werte der verschiedenen Gütekriterien sollten daher nicht miteinander verglichen werden, um zum besten Modell zu gelangen. Die Ergebnisse wurden zudem durch eine zehnfach Kreuzvalidierung bestätigt.

Die Interpretation der Regressionskoeffizienten eines GEE-MFP Modells unterscheidet sich von der Interpretation eines gewöhnlichen GEE Modells. Ist der Exponent der Einflussvariable von 1 verschieden, sollte die Parameterschätzung immer gemeinsam mit dem Exponenten der entsprechenden Variable betrachtet werden. Zum Beispiel war der Schätzer für die Variable der Nüchtern-Serum-Glukose-Konzentration negativ. In einem gewöhnlichen GEE wäre die Interpretation also, dass eine steigende Nüchtern-Serum-Glukose-Konzentration zu einer niedrigeren Wahrscheinlichkeit, an Bluthochdruck zu erkranken, führt. Da in diesem Modell der Exponent der Nüchtern-Serum-Glukose-Konzentration allerdings bei $-0,5$ lag, ergibt sich die Interpretation, dass je höher die Nüchtern-Serum-Glukose-Konzentration ist, desto höher ist auch die Wahrscheinlichkeit, an Bluthochdruck zu erkranken.

Zusammenfassend empfehle ich, die Transformationen der fraktionellen Polynome gemeinsam mit den GEE in einer Analyse einer beliebigen Anzahl unabhängiger Variablen und einer korrelierten, möglicherweise nicht normalverteilten Zielvariablen zu nutzen. Dieser Ansatz sollte dann mit einer schrittweisen Prozedur unter Verwendung des neu vorgestellten Quasi Likelihood Informationskriteriums, welches auf dem BIC basiert, kombiniert werden.

6. Diskussion und Ausblick

Häufig werden in biometrischen, epidemiologischen oder auch klinischen Fragestellungen die zentralen Annahmen der Statistik verletzt. Hierbei handelt es sich insbesondere um die Normalverteilung der Zielvariablen als auch die Unabhängigkeit der Beobachtungen.

Die Normalverteilungsannahme ist verletzt, wenn Zählraten, wie die Anzahl epileptischer Anfälle, oder binäre Daten, beispielsweise das Vorliegen einer Erkrankung, von Interesse sind. Die Unabhängigkeit der Beobachtungen wird bei wiederholten Messungen an einem Patienten, gepaarten Daten, wie Messungen am linken und rechten Auge, oder auch in Familienstudien außer Kraft gesetzt.

Ist nur die Normalverteilungsannahme verletzt, sind GLM, beispielsweise das logistische Modell bei binären Zielvariablen, ein häufig genutzter Ansatz zur Regressionsanalyse.

Werden allerdings Abhängigkeiten ignoriert, kann es zu falschen Schlussfolgerungen kommen. In der Regel sind dies die Überschätzung der Genauigkeit und damit falsch-positive Ergebnisse.

Die GEE liefern eine Möglichkeit, mit diesen Abhängigkeiten und zeitgleich mit einer eventuellen Verletzung der Normalverteilungsannahme umzugehen. Beobachtungen werden in diesem Ansatz als Cluster angesehen. Zum Beispiel sind die Messungen an einem Patienten oder auch die Mitglieder einer Familie ein Cluster. Es wird angenommen, dass zwischen den Beobachtungen eines Clusters eine Verbindung besteht. Die Beobachtungen unterschiedlicher Cluster sind jedoch unabhängig. Bei den GEE ist besonders, dass im Gegensatz zu anderen Ansätzen keine korrekte Spezifikation der kompletten multivariaten Verteilung benötigt wird, sondern es genügt, die Mittelwertstruktur korrekt zu spezifizieren. Zusätzlich zu Mittelwert- und Varianzstruktur, die auch in den GLM gewählt werden müssen, muss bei den GEE noch eine Arbeitskorrelationsstruktur angegeben werden. Da die wahre Korrelationsmatrix in den meisten Fällen unbekannt ist, sollte diese Arbeitskorrelationsmatrix aufgrund biologisch plausibler und statistischer Gründe möglichst nah an der wahren Arbeitskorrelationsmatrix gewählt werden. Der Vorteil ist, dass, selbst wenn die Varianzen und Korrelationen fehlerhaft spezifiziert wurden, valide Schätzungen der Regressionskoeffizienten mithilfe der GEE möglich sind. Allerdings kann diese Robustheit einen Verlust in der Effizienz zur Folge haben. Die Wahl der Arbeitskorrelationsmatrix hat hier also einen besonderen Einfluss.

Die einfachste Wahl für die Arbeitskorrelationsmatrix ist die Einheitsmatrix, die die Unabhängigkeit zwischen den Beobachtungen innerhalb eines Clusters beschreibt. Dies führt zu den IEE. Die Effizienz der IEE im Vergleich mit den GEE hängt von unterschiedlichen Faktoren ab. Dazu werden unter anderen Faktoren die Verteilung der Einflussvariablen, die Größe der Cluster, die Regressionsparameter und auch die Korrelation zwischen den abhängigen Variablen gezählt. Es wurde gezeigt, dass die IEE genauso effizient sind wie die GEE mit einer nicht unabhängigen Arbeitskorrelationsstruktur, wenn die Größe der Cluster nicht stark schwanken und entweder alle Kovariaten Cluster konstant, Mittelwert balanciert oder eine Mischung aus diesen beiden Möglichkeiten sind. Da bei Beobachtungsstudien die Clustergrößen häufig stark schwanken, kann es in einem solchen Fall zu einem enormen Verlust der Effizienz kommen, sollten die IEE zur Anwendung kommen. Allgemein sollte nur eine angemessen begründete nicht unabhängige Arbeitskorrelationsstruktur gewählt werden. Allerdings sollte beachtet werden, dass eine intensive Modellierung der Arbeitskorrelationsmatrix nur vernachlässigbare Gewinne in der Effizienz liefern könnte.

Die statistischen Eigenschaften des Parameterschätzers sind von der Unverzerrtheit der GEE abhängig, welche für die IEE immer gegeben ist. Allerdings können die Parameterschätzer im Falle der GEE mit einer anderen Arbeitskorrelationsmatrix als der Einheitsmatrix verzerrt sein. In einem solchen Fall ist eine hinreichende Bedingung für die Konsistenz, dass die abhängige Variable unter Verwendung aller Kovariaten des gesamten Clusters korrekt spezifiziert ist.

Ein weiterer Vorteil der IEE ist es, dass sie meistens, sollte die Fallzahl nicht allzu klein sein, konvergieren. Bei einer Hinzunahme zusätzlicher Korrelationsparameter in das Modell konvergiert der Algorithmus seltener.

Die bisherige Darstellung liefert nur Situationen, in denen die IEE den GEE mit einer anderen Arbeitskorrelationsstruktur als der Einheitsmatrix vorzuziehen sind. Es wurden daher statistische Maßzahlen vorgestellt, um eine spezifische Arbeitskorrelationsstruktur zu wählen. Hierzu gehören insbesondere die Quasi Likelihood Informationskriterien (engl. quasi likelihood information criterion, QIC), welche sich aus der Addition eines Terms, der die Quasi Likelihood unter der Unabhängigkeitsannahme berechnet, und einem Term, der den Unterschied zwischen der geschätzten Modell basierten und der geschätzten robusten Kovarianzmatrix abbildet. Ist das Unabhängigkeitsmodell korrekt, kann eine vereinfachte Version des QIC, das QIC_u , zum Einsatz kommen. Während das QIC sowohl zur Selektion von Variablen der Mittelwert- als auch der Assoziationsstruktur geeignet ist, kann das QIC_u nur zur Selektion in der Mittelwertstruk-

tur eingesetzt werden. Es wird jeweils das Modell gewählt, welches das gewählte Kriterium minimiert.

Doch kurz und knapp: Wann sollten IEE zum Einsatz kommen, wann GEE?

Da in den meisten klinisch-epidemiologischen Studien die Clustergrößen nicht stark variieren und die unabhängigen Variablen zumeist Cluster konstant oder Mittelwert balanciert sind, würden die GEE lediglich einen geringen Zugewinn in der statistischen Macht liefern. Außerdem kommt es in einem solchen Fall seltener zur Konvergenz, was in einer klinisch-epidemiologischen Studie zur Empfehlung zugunsten der IEE führt.

Möglicherweise möchte man aber doch die optimale Arbeitskorrelationsmatrix für seine Daten auswählen. Hierbei bleibt es selbstverständlich möglich, dass die zugrundeliegende wahre Korrelationsstruktur sehr viel komplizierter ist, als die angenommene Arbeitskorrelationsstruktur es abbildet. Hier können Sensitivitätsanalysen herangezogen werden, um den Effekt verschiedener Arbeitskorrelationsstrukturen auf das Modell zu untersuchen.

Biologisch plausibel oder vielmehr intuitiv ist es, für Cluster, wie beispielsweise Haushalte oder Familien, die austauschbare Arbeitskorrelationsstruktur zu wählen. Für longitudinale Daten sollte die AR(1) Arbeitskorrelationsmatrix gegenüber einer Bandmatrix, beispielsweise der Dreibandmatrix, vorgezogen werden. Des Weiteren scheinen m -Abhängige Korrelationsstrukturen biologisch wenig plausibel. Liegt eine schwach abhängige dichotome Zielvariable vor, könnten die IEE das Modell der Wahl sein. In den meisten räumlichen Studien erscheint die AR(1) Struktur als besonders sinnvoll. Dennoch können hier auch periodische Korrelationsstrukturen möglich sein.

Eine weitere Möglichkeit zur optimalen Arbeitskorrelationsstruktur zu kommen, spiegelt sich einem zweistufigen Verfahren wieder. Im ersten Schritt wird mit den vorgestellten statistischen Ansätzen die Arbeitskorrelationsmatrix als die Beste gewählt, welche das Gütekriterium minimiert. Im zweiten Schritt werden eventuelle Parameterrestriktionen untersucht. Sollte das optimale Modell im Sinne des ersten Schritts diese Restriktionen verletzen, wird das zweitbeste Modell genutzt, sofern es die Restriktionen einhält. Dieses Verfahren wird fortgeführt bis man zu einem Modell kommt, das die Parameterrestriktionen einhält.

Die letztgenannte Möglichkeit fand Anwendung in der erneuten Auswertung einer doppelblinden, Placebo kontrollierten, randomisierten, multizentrischen, longitudinalen

len Studie mittels GEE. In dieser Studie wurde der Ödem protektive Effekt des vasoaktiven Medikaments SB-LOT bei Patienten, die an einer chronischen Veneninsuffizienz nach Kongestionsverminderung leiden, untersucht.

Die GEE fanden bisher selten Anwendung in kontrollierten klinischen Studien. Sie stellen aber durchaus eine Alternative zur Standardanalyse, die nur die letzte Nachuntersuchung betrachtet, dar. In der Analyse sollten die GEE bestmöglich angepasst werden. Dies geschah nicht nur mittels des Auswahlverfahrens der besten Arbeitskorrelationsstruktur, sondern es kamen auch noch Techniken der Regressionsdiagnostik zum Einsatz, um die Auswertung als Sensitivitätsanalyse zu begleiten.

Um die optimale Arbeitskorrelationsstruktur zu finden, sollten die Unabhängigkeitsstruktur, die AR(1) Struktur, die austauschbare Struktur sowie die 1-Abhängiger Struktur als Dreibandmatrix und der Ansatz keine Struktur, also kein Vorwissen einzubringen, verglichen werden. Als Kriterium wurde das QIC genutzt. Dieses wurde unter Verwendung der AR(1) Struktur minimal, was den theoretischen Ergebnissen zur Wahl der Arbeitskorrelationsstruktur entspricht. Es sollte erwähnt werden, dass die IEE den GEE wie erwartet nur knapp unterlegen waren.

In diese Analyse können alle Zeitpunkte und daher mehr Informationen über den Verlauf der Studie einbezogen werden. Es konnte ein Schätzer für die Differenz des Unterschenkelvolumens zwischen den Behandlungsgruppen (SB-LOT vs. Placebo) auf einer wöchentlichen Basis geliefert werden. Die Standardanalyse, eine ANCOVA, kann eine solche Interpretation nicht liefern. Auch in der GEE Analyse wurde die Überlegenheit des Medikaments über dem Placebo nachgewiesen. Der Effekt des Medikaments ist ein Unterschied im Unterschenkelvolumen von durchschnittlich 2,64 ml pro Woche (95%-Konfidenzintervall: 0,27–4,99 ml pro Woche). Am Ende der Studie bedeutete dies eine Differenz von durchschnittlich 31,68 ml (95%-Konfidenzintervall: 3,24–59,88 ml) im Unterschenkelvolumen bei Patienten unterschiedlicher Behandlungsgruppen.

Zur weiteren Verbesserung in der Anpassung der GEE kann als Sensitivitätsanalyse die Deletionsdiagnostik zum Einsatz kommen. Neben der Verbesserung der Anpassung können durch regressionsdiagnostische Techniken einflussreiche Beobachtungen identifiziert werden. Ändern sich die Ergebnisse nach der Entfernung von Hebelpunkten oder einflussreichen Clustern nicht oder nur leicht, können die Studienergebnisse als robust interpretiert werden. In dieser Arbeit wurden allgemeine deletionsdiagnostische Verfahren, beispielweise Residualabbildungen und Cluster-Cook-Statistiken, verwendet und erstmalig mit rechenzeitintensiven Halbnormalabbildungen mit simulier-

ten Einhüllenden kombiniert. Speziell wurde außerdem der Leave-one-out-Ansatz bei Halbnormalabbildungen erweitert. Dazu war es notwendig, alle Abbildungen von zwei unabhängigen Gutachtern auswerten zu lassen, um Patienten mit einem starken Einfluss visuell zu identifizieren. Die von den Gutachtern ausgeschlossenen Patienten hatten ganz spezifische Charakteristiken, wie beispielsweise die größte Veränderung in der Zielvariable Unterschenkelvolumen. Ein Nachteil dieses neuen Ansatzes ist es allerdings, dass alle Halbnormalabbildungen einzeln visuell bewertet werden mussten. Eine Automatisierung zur Identifizierung einflussreicher Patienten auf die Abbildung oder die simulierten Einhüllenden würde die Arbeitsbelastung deutlich senken. Die Ergebnisse der regressionsdiagnostischen Ansätze in der SB-LOT-Studie haben gezeigt, dass sich die Hauptaussage auch nach Ausschluss einflussreicher Patienten von der Analyse nicht verändert. Dadurch wird die Validität des Schlusses, dass ein signifikanter Ödem protektiver Effekt von SB-LOT vorliegt, unterstrichen.

Sowohl die optimierte Wahl der Arbeitskorrelationsstruktur als auch die regressionsdiagnostischen Techniken können nur zu einer verbesserten Anpassung der GEE führen, sollte ein linearer Zusammenhang zwischen Einflussvariablen und Zielvariable bestehen. Diese Maßnahmen nutzen wenig, wenn der Zusammenhang nicht linear ist. Ein solcher Zusammenhang konnte von den GEE bisher nicht modelliert werden. FP hingegen sind bei unabhängigen Beobachtungen in der Lage, solche nicht linearen Kurvenverläufe anzupassen. Daher wurde in dieser Arbeit ein Algorithmus entwickelt, der die Vorteile der GEE, korrelierte Beobachtungen modellieren zu können, mit denen der FP verbindet. Der vorgestellte Algorithmus wählt das optimale Modell unter einer großen Anzahl möglicher Modelle aus. Dabei kann er mit einer beliebigen Anzahl unabhängiger Variablen umgehen und binäre, kategoriale und stetige Einflussvariablen gemeinsam analysieren. Hierbei kann vom Nutzer gewählt werden, ob eine unabhängige Variable linearen Einfluss haben soll oder ob ein Kurvenverlauf eines FP1 oder FP2 möglich sein soll. Implementiert ist eine Funktionsselektionsverfahren, der Modellbildungsprozess, das wahlweise um eine schrittweise Prozedur ergänzt werden kann, um die einflussreichsten Variablen im Modell zu halten. In vorhergehenden Arbeiten wurde das QIC als Maß der Güte und Selektion genutzt, um über die funktionelle Form des Zusammenhangs zwischen unabhängigen Variablen und Zielvariable zu entscheiden. In dieser Arbeit wurde ein alternatives Kriterium, das $BQIC_u$, vorgeschlagen, um diese Auswahl zu treffen. Dieses neue Kriterium basiert nicht, wie das QIC, auf Akaike's, sondern auf dem Bayesianischen Informationskriterium.

Zur Veranschaulichung des neuen Algorithmus in Kombination mit den verschiedenen Maßzahlen wurden die Daten der Framingham Heart Study noch einmal analysiert.

Zur Validierung der Ergebnisse und des Algorithmus wurde mit diesen Daten zusätzlich eine zehnfach Kreuzvalidierung durchgeführt.

Unter Verwendung des vollständigen Datensatzes ergab sich mithilfe des neuen Kriteriums zur Güte der Anpassung und Selektion ein niedrigdimensionaleres Modell als unter Verwendung der aus der Literatur bekannten Kriterien. Das bedeutet, dass sowohl unter Verwendung des QIC als auch des QIC_u Variablen FP2 transformiert wurden, die unter Verwendung des $BQIC_u$ entweder FP1 oder sogar nur linear in das Modell genommen wurden. Keine Variable wurde in dem durch das $BQIC_u$ als optimal identifizierte Modell mit einem FP2 Einfluss transformiert. Hinzu kam, dass nach der schrittweisen Prozedur und mit $BQIC_u$ als Kriterium der Güte der Anpassung und Selektion die wenigsten Variablen im Modell blieben und diese zudem die niedrigdimensionalsten Transformationen aufwiesen.

Abbildungen zeigten, dass, während die Modelle unter Verwendung von QIC und QIC_u weit von den zu erwartenden Kurvenverläufen entfernt waren, die mit $BQIC_u$ identifizierten Modelle vor und nach der schrittweisen Prozedur die erwünschten Verläufe sehr gut abbildeten.

Die zehnfach Kreuzvalidierung bestätigte den Nutzen des neuen Kriteriums. Auch hier zeigte sich, dass unter seiner Verwendung die sparsamsten Modelle vor und nach der schrittweisen Prozedur als die optimalen identifiziert wurden. Das bedeutet, dass der Algorithmus in Verbindung mit dem $BQIC_u$ vor und nach der schrittweisen Prozedur die niedrig-dimensionalsten Modelle als optimal identifizierte. Außerdem wurden durch diese Kombination in der schrittweisen Prozedur die meisten Einflussvariablen entfernt. In keinem Fall wurde durch das neue Kriterium ein FP2 für eine Einflussvariable als beste Transformation gewählt. Die durch den Algorithmus und das neue Kriterium identifizierten Modelle waren stabiler in dem Sinn, dass am wenigsten verschiedene Transformation für eine Variable gewählt wurden. Zudem zeigten sich unter Verwendung des $BQIC_u$ die größten Anteile korrekt vorhergesagter Einträge in der Zielvariablen. Die dazugehörigen empirischen Konfidenzintervalle waren am schmalsten.

Daher empfehle ich die Transformationen der fraktionellen Polynome gemeinsam mit den GEE und dem neuen Kriterium $BQIC_u$ für die optimale Anpassung in der Mittelwertstruktur zu nutzen. Dieser Algorithmus kann noch erweitert werden, sodass es auch möglich sein sollte, die Anpassung in der Assoziationsstruktur zu optimieren.

Zusammenfassend wurde die Vorhersage von abhängigen Zielvariablen in dieser Arbeit durch drei verschiedene Ansätze unter Verwendung der GEE verbessert. Nach einem Überblick über die Wahl der besten Arbeitskorrelationsstruktur wurden die theoretischen Erkenntnisse an einer longitudinalen klinischen Studie, die bisher nur selten mithilfe der GEE ausgewertet werden, veranschaulicht. Hier zeigte sich, dass ein Erkenntnisgewinn im Vergleich zu der Standardanalyse einer solchen Studie hervorgerufen werden konnte, da mehr Informationen aus den Daten genutzt wurden. Des Weiteren wurde anhand dieser Studie gezeigt, dass regressionsdiagnostische Verfahren als weitere Sensitivitätsanalyse die GEE in der Auswertung begleiten sollten. Insbesondere wurde erstmalig Halbnormalabbildungen mit simulierten Einhüllenden in Verbindung mit klassischen deletionsdiagnostischen Verfahren angewendet. Zudem wurden die Halbnormalabbildungen in einer Leave-one-out-Strategie als deletionsdiagnostisches Verfahren zur Verbesserung der Anpassung vorgeschlagen. Um zukünftig auch nicht lineare Zusammenhänge zwischen Einflussvariablen abbilden zu können, wurden die GEE mit den FP in einem Algorithmus vereint. Hier wurde außerdem ein neues Kriterium zur Güte der Anpassung und Selektion eingeführt, welches zu einem deutlichen Gewinn in der Vorhersagekraft führte. Sowohl der Algorithmus als auch der Nutzen des neuen Kriteriums wurden anhand einer epidemiologischen Studie gezeigt. Diese Arbeit zeigt daher deutlich, dass die GEE in verschiedenen Anwendungsgebieten und Studientypen zum Einsatz kommen und mit den vorgeschlagenen Methoden die Vorhersage von korrelierten Zielvariablen deutlich verbessern können.

7. Zusammenfassung

Fragestellung In zahlreichen biometrischen, epidemiologischen und klinischen Studien wird die Vorherhersage einer Zielvariablen, beispielsweise das Unterschenkelvolumen bei Patienten mit venöser Insuffizienz oder ob Bluthochdruck vorliegt, anhand verschiedener Einflussvariablen gewünscht. Dies können bei Patienten in einer longitudinalen Studie zur venösen Insuffizienz die Therapieform und der Beobachtungszeitpunkt oder in einer epidemiologischen Familienstudie zum Bluthochdruck eine Vielzahl bekannter Prädiktoren sein. Oftmals werden dabei die zentralen Annahmen der Statistik verletzt. Dies sind insbesondere die Normalverteilung der Daten bei der Frage, ob eine Erkrankung vorliegt, und die Unabhängigkeit der Daten durch wiederholte Messungen an einer Person oder die Verwandtschaftsverhältnisse der Personen einer Familie. Hier liefern die Standardmethoden wie das lineare oder das logistische Modell nur verzerrte Schätzungen. Eine Möglichkeit, die Problematik in den Griff zu bekommen, sind die verallgemeinerten Schätzgleichungen. Hier werden abhängige Beobachtungen, wie wiederholte Messungen oder Mitglieder einer Familie, als ein Cluster angesehen. Zwischen den Beobachtungen eines Clusters besteht eine Verbindung, unterschiedliche Cluster sind unabhängig. Um zu robusten Schätzungen der Regressionskoeffizienten, also der Effekte, zu kommen, muss eine Arbeitskorrelationsstruktur, die die Verbindungen innerhalb eines Clusters möglichst gut widerspiegelt, angegeben werden. Varianzen und Korrelationen dürfen dabei fehlspezifiziert werden. Die Robustheit kann aber einen Verlust in der Effizienz zur Folge haben. Wie wählt man eine Arbeitskorrelationsstruktur so, dass die Schätzungen effizient sind? Wie kann die Vorhersage durch verallgemeinerte Schätzgleichungen noch verbessert werden?

Material und Methoden Zur Wahl der Arbeitskorrelationsmatrix wird eine Literaturübersicht angefertigt. Anschließend wird eine longitudinale Studie zur venösen Insuffizienz erneut analysiert. Hier werden die verallgemeinerten Schätzgleichungen durch klassische Deletionsdiagnostik, die Ausreißer oder einflussreiche Punkte in einem Datensatz findet, und einer neuen Leave-one-out-Strategie bei Halbnormalabbildungen mit simulierten Einhüllenden zur Verbesserung der Anpassung begleitet. Ein Algorithmus, der die Vorteile der verallgemeinerten Schätzgleichungen mit denen der fraktionellen Polynome, nicht lineare Kurvenverläufe modellieren zu können, verbindet, wurde entworfen und an einer epidemiologischen Familienstudie zum Vorliegen von Bluthochdruck getestet. Ein neues Kriterium der Güte der Anpassung, das nicht wie die bisherigen Quasi Likelihood Informationskriterien auf Akaike's, sondern auf dem Bayesianischen Informationskriterium beruht, wurde entwickelt, um die Anpassung zu verbessern.

Ergebnisse Die Einheitsmatrix beschreibt die Unabhängigkeit der Beobachtungen eines Clusters und ist die einfachste Wahl, robuste und unverzerrte Schätzer zu erhalten. Zur Verbesserung der Effizienz kann in Familienstudien die austauschbare, für longi-

tudinale Daten die autoregressive Struktur erster Ordnung gewählt werden. Letztere wurde auch in der Beispielstudie als optimal identifiziert. Der neue deletionsdiagnostische Ansatz führte zu einem Ausschluss von Patienten, was das Ergebnis der Studie, die Überlegenheit der neuen Therapie, nicht beeinflusste. Der Schätzer für den Effekt ist daher als robust anzusehen. Verallgemeinerte Schätzgleichungen verbunden mit fraktionellen Polynomen lieferten mit dem neuen Informationskriterium in der zweiten Beispielstudie niedrig-dimensionalere, sparsamere und besser angepasste Modelle als mit den bekannten Kriterien. Durch eine zehnfach Kreuzvalidierung wurde gezeigt, dass auch der Anteil korrekt vorhergesagter Einträge in der Zielvariablen vergrößert wurde.

Diskussion Verallgemeinerte Schätzgleichungen sind eine sinnvolle Alternative zur Vorhersage korrelierter eventuell nicht normalverteilter Zielvariablen. In dieser Arbeit wurden neue Ansätze bei verallgemeinerten Schätzgleichungen vorgestellt. Ihre Anwendung führt zu einem höheren Informationsgewinn und einer verbesserten Vorhersage im Vergleich zu den Standardmethoden. In longitudinalen Studien liefern sie nicht nur eine Entscheidung zur Überlegenheit einer Therapie, sondern zusätzlich die Veränderung der Zielgröße pro Zeiteinheit. Die Vereinigung von klassischer Deletionsdiagnostik mit der neuen Leave-one-out-Strategie bei Halbnormalabbildungen ist eine sinnvolle Ergänzung, um Ausreißer oder einflussreiche Punkte in den Daten zu identifizieren. Verallgemeinerte Schätzgleichungen verbunden mit fraktionellen Polynomen verbesserten, insbesondere bei nicht linearen Zusammenhängen, für eine beliebige Anzahl verschieden skaliertes Einflussvariablen mithilfe eines neuen Bayesianischen Quasi Likelihood Informationskriteriums die Vorhersage deutlich.

A. Ergebnisse des GEE-MFP-Algorithmus

A.1. Ergebnisse unter Verwendung des vollständigen Datensatzes

Tabelle A.1: Ergebnisse der vollständigen Modelle unter Verwendung des vollständigen Datensatzes und QIC , QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1711,64$ or $QIC_u=1711,98$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		5,46	1,29; 9,64	0,01	
Alter	age^{-1}	23,45	14,95; 31,96	$6,55 \cdot 10^{-8}$	$2,16 \cdot 10^{-15}$
	$age^{-0,5}$	-34,37	-45,02; -23,72	$2,52 \cdot 10^{-10}$	
Körpermasseindex	$bmi^{0,5}$	5,07	3,99; 6,14	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,02	-1,11; 1,14	0,98	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,23	1,25; 3,22	$9,35 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,80 \cdot 10^{-3}$	$-2,62 \cdot 10^{-3}$; $-9,79 \cdot 10^{-4}$	$1,72 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^3$	0,94	0,02; 1,86	0,05	$4,69 \cdot 10^{-3}$
	$drkt^3 \cdot \ln(drkt)$	-1,93	-3,88; 0,01	0,05	
Körperhöhe	hgt^{-2}	-33,46	-98,58; 31,66	0,31	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,43	0,17; 0,68	$9,35 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-16,46	-32,66; -0,26	0,05	
Zigaretten pro Tag	cpd_k^{-1}	-0,31	-0,69; 0,06	0,10	
Geschlecht	sex	-0,30	-0,67; 0,08	0,12	
Bluthochdruckbehandlung	hrx	0,66	-0,33; 1,64	0,19	

(b) $BQIC_u=1786,86$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-4,49	-6,96; -2,03	$3,48 \cdot 10^{-4}$	
Alter	age	0,62	0,48; 0,75	$< 10^{-15}$	
Körpermasseindex	bmi	1,49	1,16; 1,82	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,07	-1,08; 1,21	0,91	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,18	1,22; 3,15	$9,06 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,80 \cdot 10^{-3}$	$-2,58 \cdot 10^{-3}$; $-1,03 \cdot 10^{-3}$	$5,15 \cdot 10^{-6}$	
Alkoholkonsum	drkt	0,53	-0,17; 1,23	0,14	
Körperhöhe	hgt^{-2}	-9,09	-75,08; 56,91	0,79	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,43	0,17; 0,68	$9,69 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-17,93	-34,04; -1,83	0,03	
Zigaretten pro Tag	cpd_k^{-1}	-0,26	-0,64; 0,11	0,16	
Geschlecht	sex	-0,40	-0,80; -0,01	0,04	
Bluthochdruckbehandlung	hrx	0,69	-0,28; 1,65	0,16	

Tabelle A.2: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung des vollständigen Datensatzes und QIC , QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1709,12$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		4,36	0,87; 7,84	0,01	
Alter	age^{-1}	21,53	14,09; 28,97	$1,41 \cdot 10^{-8}$	$< 10^{-15}$
	$age^{-0,5}$	-32,30	-41,75; -22,86	$2,02 \cdot 10^{-11}$	
Körpermasseindex	$bmi^{0,5}$	4,95	3,91; 6,00	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,12	1,21; 3,02	$4,81 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,92 \cdot 10^{-3}$	$-2,71 \cdot 10^{-3}$; $-1,13 \cdot 10^{-3}$	$2,14 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^3$	1,07	0,17; 1,96	0,02	$9,40 \cdot 10^{-4}$
	$drkt^3 \cdot \ln(drkt)$	-2,16	-4,04; -0,28	0,02	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,23; 0,69	$9,02 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-17,40	-33,57; -1,22	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,41	-0,70; -0,12	$5,93 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(b) $QIC_u = 1709,64$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		14,91	8,18; 21,64	$1,41 \cdot 10^{-5}$	
Alter	$age^{-0,5}$	-19,80	-25,20; -14,40	$6,48 \cdot 10^{-13}$	$2,32 \cdot 10^{-15}$
	$age^{-0,5} \cdot \ln(age)$	-14,00	-19,09; -8,92	$6,81 \cdot 10^{-8}$	
Körpermasseindex	$bmi^{0,5}$	4,94	3,90; 5,99	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,11	1,20; 3,02	$5,67 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,94 \cdot 10^{-3}$	$-2,74 \cdot 10^{-3}$; $-1,13 \cdot 10^{-3}$	$2,40 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^3$	1,08	0,18; 1,97	0,02	$8,54 \cdot 10^{-4}$
	$drkt^3 \cdot \ln(drkt)$	-2,18	-4,06; -0,30	0,02	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,45	0,23; 0,68	$1,03 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-10,93	-21,17; -0,69	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,41	-0,70; -0,12	$5,61 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(c) $BQIC_u = 1771,97$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-2,63	-6,15; 0,90	0,14	
Alter	age	0,64	0,51; 0,78	$< 10^{-15}$	
Körpermasseindex	bmi	1,47	1,15; 1,80	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,31	1,42; 3,20	$4,01 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,96 \cdot 10^{-3}$	$-2,72 \cdot 10^{-3}$; $-1,20 \cdot 10^{-3}$	$4,46 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,48	0,25; 0,71	$5,27 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-12,40	-22,54; -2,26	0,02	
Zigaretten pro Tag					
Geschlecht	sex	-0,53	-0,81; -0,25	$2,08 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

A.2. Ergebnisse der zehnfach Kreuzvalidierung

Tabelle A.5: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–9 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1518,14$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		18,49	10,54; 26,43	$5,07 \cdot 10^{-6}$	
Alter	$age^{-0,5}$	-20,96	-27,21; -14,71	$4,95 \cdot 10^{-11}$	$1,91 \cdot 10^{-13}$
	$age^{-0,5} \cdot \ln(age)$	-15,20	-21,18; -9,23	$6,17 \cdot 10^{-7}$	
Körpermasseindex	bmi^3	0,27	0,18; 0,37	$3,35 \cdot 10^{-8}$	$1,73 \cdot 10^{-11}$
	$bmi^3 \cdot \ln(bmi)$	-0,15	-0,21; -0,08	$5,86 \cdot 10^{-6}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,20	-1,44; 1,03	0,75	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,81	0,80; 2,82	$4,32 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,95 \cdot 10^{-3}$	$-2,77 \cdot 10^{-3}$; $-1,13 \cdot 10^{-3}$	$3,27 \cdot 10^{-6}$	$7,68 \cdot 10^{-4}$
	$drkt^3$	1,17	0,20; 2,14	0,02	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,42	-4,47; -0,37	0,02	$7,68 \cdot 10^{-4}$
	hgt^{-2}	-30,48	-99,61; 38,65	0,39	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,42	0,14; 0,70	$3,50 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^2$	$8,29 \cdot 10^{-3}$	$-5,69 \cdot 10^{-6}$; 0,02	0,05	
Zigaretten pro Tag	cpd_k^{-2}	-0,25	-0,56; 0,07	0,12	
Geschlecht	sex	-0,22	-0,63; 0,19	0,29	
Bluthochdruckbehandlung	hrx	0,46	-0,56; 1,48	0,38	

(b) $QIC_u=1518,27$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		16,67	8,83; 24,51	$3,07 \cdot 10^{-5}$	
Alter	$age^{-0,5}$	-21,27	-27,54; -14,99	$3,18 \cdot 10^{-11}$	$1,21 \cdot 10^{-13}$
	$age^{-0,5} \cdot \ln(age)$	-15,47	-21,50; -9,44	$4,92 \cdot 10^{-7}$	
Körpermasseindex	bmi	1,64	1,27; 2,00	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,15	-1,39; 1,09	0,81	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,15	0,95; 3,35	$4,47 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,97 \cdot 10^{-3}$	$-2,79 \cdot 10^{-3}$; $-1,16 \cdot 10^{-3}$	$1,82 \cdot 10^{-6}$	$1,11 \cdot 10^{-3}$
	$drkt^3$	1,13	0,16; 2,10	0,02	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,34	-4,38; -0,30	0,02	$1,11 \cdot 10^{-3}$
	hgt^{-2}	-30,81	-100,50; 38,85	0,39	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,45	0,16; 0,74	$2,11 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^3$	$5,13 \cdot 10^{-4}$	$-2,72 \cdot 10^{-5}$; $1,05 \cdot 10^{-3}$	0,06	
Zigaretten pro Tag	cpd_k^{-2}	-0,25	-0,56; 0,07	0,12	
Geschlecht	sex	-0,25	-0,66; 0,16	0,23	
Bluthochdruckbehandlung	hrx	0,50	-0,53; 1,53	0,35	

(c) $BQIC_u=1591,45$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-6,26	-8,15; -4,36	$9,45 \cdot 10^{-11}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	bmi	1,59	1,22; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,18	-1,43; 1,06	0,77	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,78	0,79; 2,78	$4,44 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,99 \cdot 10^{-3}$	$-2,81 \cdot 10^{-3}$; $-1,17 \cdot 10^{-3}$	$1,86 \cdot 10^{-6}$	$1,86 \cdot 10^{-6}$
	$drkt$	0,69	-0,06; 1,44	0,07	
Alkoholkonsum	hgt^{-2}	-13,54	-83,13; 56,05	0,70	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,42	0,13; 0,70	$3,74 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^2$	$8,70 \cdot 10^{-3}$	$4,73 \cdot 10^{-4}$; 0,02	0,04	
Zigaretten pro Tag	cpd_k^{-2}	-0,23	-0,54; 0,07	0,13	
Geschlecht	sex	-0,30	-0,72; 0,12	0,16	
Bluthochdruckbehandlung	hrx	0,48	-0,52; 1,48	0,35	

Tabelle A.6: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–8 und 10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1474,76$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		19,00	10,53; 27,46	$1,09 \cdot 10^{-5}$	
Alter	$age^{-0,5}$	-21,42	-27,84; -15,00	$6,31 \cdot 10^{-11}$	$1,50 \cdot 10^{-13}$
	$age^{-0,5} \cdot \ln(age)$	-15,71	-21,85; -9,56	$5,47 \cdot 10^{-7}$	
Körpermasseindex	$bmi^{0,5}$	4,76	3,60; 5,93	$1,33 \cdot 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,15	-1,38; 1,07	0,80	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,24	1,14; 3,34	$6,20 \cdot 10^{-5}$	
Familiengröße	fs^3	$-5,92 \cdot 10^{-3}$	$-0,01; 3,06 \cdot 10^{-3}$	0,20	
	$drkt^3$	0,81	-0,17; 1,80	0,11	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-1,79	-3,66; 0,07	0,06	0,01
Körperhöhe	hgt^{-2}	-43,13	-115,60; 29,30	0,24	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,39	0,11; 0,66	$5,72 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-13,18	-23,94; -2,42	0,02	
Zigaretten pro Tag	cpd_k^{-2}	-0,29	-0,62; 0,04	0,08	
Geschlecht	sex	-0,30	-0,73; 0,12	0,17	
Bluthochdruckbehandlung	hrx	0,54	-0,47; 1,56	0,30	

(b) $QIC_u=1472,05$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		3,45	-2,47; 9,38	0,25	
Alter	age^{-1}	25,45	16,23; 34,67	$6,31 \cdot 10^{-8}$	$2,01 \cdot 10^{-15}$
	$age^{-0,5}$	-36,99	-48,51; -25,47	$3,12 \cdot 10^{-10}$	
Körpermasseindex	$\ln(bmi)$	3,86	2,88; 4,83	$9,99 \cdot 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,07	-1,16; 1,30	0,91	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,90	1,63; 4,16	$7,21 \cdot 10^{-6}$	
Familiengröße	fs^{-2}	-0,19	-0,13; 0,50	0,25	
	$drkt^{-2}$	$1,18 \cdot 10^{-3}$	$-6,98 \cdot 10^{-4}; 3,05 \cdot 10^{-3}$	0,22	
Alkoholkonsum	$drkt^{-2} \cdot \ln(drkt)$	$-2,59 \cdot 10^{-4}$	$-1,45 \cdot 10^{-4}; 6,64 \cdot 10^{-4}$	0,21	0,08
Körperhöhe	hgt^{-2}	-46,51	-119,60; 26,59	0,21	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,48	0,20; 0,76	$9,29 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(glucw)$	2,23	0,54; 3,91	$9,45 \cdot 10^{-3}$	
Zigaretten pro Tag	cpd_k^{-1}	-0,45	-0,88; -0,02	0,04	
Geschlecht	sex	-0,39	-0,83; 0,05	0,09	
Bluthochdruckbehandlung	hrx	0,65	-0,39; 1,70	0,22	

(c) $BQIC_u=1544,01$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-7,68	-16,08; 0,73	0,07	
Alter	age	0,63	0,48; 0,77	$< 10^{-15}$	
	$bmi^{0,5}$	4,58	3,41; 5,75	$1,47 \cdot 10^{-14}$	
Serum-Totalcholesterin-Konzentration	$chol^3$	$9,65 \cdot 10^{-3}$	-0,02; 0,04	0,52	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,17	1,13; 3,21	$4,52 \cdot 10^{-5}$	
Familiengröße	fs^{-2}	0,18	-0,13; 0,49	0,26	
Alkoholkonsum	$drkt$	0,45	-0,30; 1,19	0,24	
Körperhöhe	$hgt^{0,5}$	0,74	-1,88; 3,36	0,58	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,38	0,10; 0,65	$7,16 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-14,28	-24,94; -3,62	$8,63 \cdot 10^{-3}$	
Zigaretten pro Tag	cpd_k^{-2}	-0,24	-0,57; 0,08	0,13	
Geschlecht	sex	-0,37	-0,82; 0,07	0,10	
Bluthochdruckbehandlung	hrx	0,56	-0,44; 1,57	0,27	

Tabelle A.7: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–7 und 9–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1489,96$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		22,38	14,97; 29,78	$3,13 \cdot 10^{-9}$	
Alter	$age^{-0,5}$	-22,43	-28,69; -16,16	$2,30 \cdot 10^{-12}$	$< 10^{-15}$
	$age^{-0,5} \cdot \ln(age)$	-17,02	-23,10; -10,95	$3,97 \cdot 10^{-8}$	
Körpermasseindex	$\ln(bmi)$	4,03	3,12; 4,93	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,03	-1,18; 1,24	0,96	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,59	1,55; 3,62	$9,06 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,84 \cdot 10^{-3}$	$-2,76 \cdot 10^{-3}$; $-9,20 \cdot 10^{-4}$	$8,79 \cdot 10^{-5}$	$1,68 \cdot 10^{-3}$
	$drkt^3$	1,16	0,13; 2,18	0,03	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,32	-4,43; -0,21	0,03	
Körperhöhe	hgt^3	$1,62 \cdot 10^{-3}$	$-1,90 \cdot 10^{-3}$; $5,13 \cdot 10^{-3}$	0,37	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,55	0,28; 0,82	$5,36 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-11,56	-22,94; -0,18	0,05	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,59	-1,22; 0,04	0,07	
Geschlecht	sex	-0,26	-0,67; 0,15	0,21	
Bluthochdruckbehandlung	hrx	0,99	-0,22; 2,20	0,11	

(b) $QIC_u=1490,39$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		14,54	5,76; 23,32	$1,17 \cdot 10^{-3}$	
Alter	$age^{-0,5}$	-22,42	-28,69; -16,16	$2,31 \cdot 10^{-12}$	$< 10^{-15}$
	$age^{-0,5} \cdot \ln(age)$	-17,02	-23,09; -10,94	$4,00 \cdot 10^{-8}$	
Körpermasseindex	$\ln(bmi)$	4,02	3,12; 4,93	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,02	-1,19; 1,23	0,97	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,58	1,55; 3,62	$9,26 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,84 \cdot 10^{-3}$	$-2,76 \cdot 10^{-3}$; $-9,19 \cdot 10^{-4}$	$8,86 \cdot 10^{-5}$	$1,69 \cdot 10^{-3}$
	$drkt^3$	1,15	0,13; 2,18	0,03	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,32	-4,43; -0,20	0,03	
Körperhöhe	hgt^3	$1,61 \cdot 10^{-3}$	$-1,91 \cdot 10^{-3}$; $5,12 \cdot 10^{-3}$	0,37	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,55	0,28; 0,82	$5,53 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(glucw)$	1,82	0,01; 3,62	0,05	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,59	-1,22; 0,04	0,07	
Geschlecht	sex	-0,26	-0,67; 0,15	0,21	
Bluthochdruckbehandlung	hrx	0,99	-0,22; 2,20	0,11	

(c) $BQIC_u=1564,55$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-10,39	-14,68; -6,09	$2,14 \cdot 10^{-6}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	$\ln(bmi)$	3,89	2,97; 4,81	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,06	-1,18; 1,30	0,92	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,56	1,55; 3,58	$7,63 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,87 \cdot 10^{-3}$	$-2,76 \cdot 10^{-3}$; $-9,76 \cdot 10^{-4}$	$4,04 \cdot 10^{-5}$	$1,69 \cdot 10^{-3}$
	$drkt$	0,59	-0,19; 1,36	0,14	
Körperhöhe	hgt^3	$7,54 \cdot 10^{-4}$	$-2,81 \cdot 10^{-3}$; $4,31 \cdot 10^{-3}$	0,68	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,53	0,27; 0,81	$8,77 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(glucw)$	1,92	0,13; 3,70	0,04	
Zigaretten pro Tag	$\ln(cpd_k)$	0,20	-0,02; 0,43	0,08	
Geschlecht	sex	-0,34	-0,76; 0,09	0,12	
Bluthochdruckbehandlung	hrx	1,00	-0,20; 2,21	0,10	

Tabelle A.8: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–6 und 8–10 und QIC , QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1564,37$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		10,01	5,31; 14,71	$3,00 \cdot 10^{-5}$	
Alter	age^{-1}	26,69	16,92; 36,45	$8,47 \cdot 10^{-8}$	$3,64 \cdot 10^{-15}$
	$age^{-0,5}$	-38,27	-50,38; -26,17	$5,83 \cdot 10^{-10}$	
Körpermasseindex	bmi	1,61	1,26; 1,96	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,09	-1,24; 1,06	0,88	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,30	1,28; 3,32	$9,85 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,95 \cdot 10^{-3}$	$-2,76 \cdot 10^{-3}$; $-1,14 \cdot 10^{-3}$	$2,54 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^{-2}$	$-3,31 \cdot 10^{-5}$	$-7,26 \cdot 10^{-5}$; $6,46 \cdot 10^{-6}$	0,10	
Körperhöhe	hgt^{-2}	-38,26	-107,00; 30,46	0,28	
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,79	0,25; 1,33	$4,09 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-15,69	-32,86; 1,48	0,07	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,37	-0,97; 0,23	0,23	
Geschlecht	sex	-0,33	-0,70; 0,05	0,09	
Bluthochdruckbehandlung	hrx	0,88	-0,22; 1,98	0,12	

(b) $QIC_u=1563,69$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		9,89	5,22; 14,55	$3,25 \cdot 10^{-5}$	
Alter	age^{-1}	26,53	16,78; 36,28	$9,71 \cdot 10^{-8}$	$4,78 \cdot 10^{-15}$
	$age^{-0,5}$	-38,01	-50,08; -25,94	$6,72 \cdot 10^{-10}$	
Körpermasseindex	bmi	1,61	1,26; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^3$	$8,70 \cdot 10^{-3}$	-0,02; 0,04	0,55	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,26	1,26; 3,26	$9,10 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,96 \cdot 10^{-3}$	$-2,77 \cdot 10^{-3}$; $-1,14 \cdot 10^{-3}$	$2,78 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^{-2}$	$-3,39 \cdot 10^{-5}$	$-7,36 \cdot 10^{-5}$; $5,68 \cdot 10^{-6}$	0,09	
Körperhöhe	hgt^{-2}	-39,31	-108,40; 29,75	0,26	
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,75	0,20; 1,29	$7,08 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-15,68	-32,84; 1,49	0,07	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,36	-0,96; 0,24	0,24	
Geschlecht	sex	-0,32	-0,70; 0,06	0,09	
Bluthochdruckbehandlung	hrx	0,86	-0,24; 1,96	0,12	

(c) $BQIC_u=1638,12$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-4,20	-6,75; -1,64	$1,31 \cdot 10^{-3}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	bmi	1,55	1,20; 1,90	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^3$	$8,11 \cdot 10^{-3}$	-0,02; 0,04	0,58	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,23	1,24; 3,22	$1,06 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,97 \cdot 10^{-3}$	$-2,80 \cdot 10^{-3}$; $-1,14 \cdot 10^{-3}$	$3,37 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^{-2}$	$-2,50 \cdot 10^{-5}$	$-6,34 \cdot 10^{-5}$; $1,33 \cdot 10^{-5}$	0,20	
Körperhöhe	hgt^{-2}	-17,73	-87,20; 51,75	0,62	
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,76	0,22; 1,30	$6,05 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-16,82	-33,92; 0,29	0,05	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,34	-0,92; 0,25	0,26	
Geschlecht	sex	-0,41	-0,79; -0,02	0,04	
Bluthochdruckbehandlung	hrx	0,82	-0,26; 1,91	0,14	

Tabelle A.9: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–5 und 7–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1557,90

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-5,21	-8,60; -1,82	$2,56 \cdot 10^{-3}$	
Alter	age^{-1}	-5,45	-6,73; -4,16	$< 10^{-15}$	$< 10^{-15}$
	$\text{age}^{-1} \cdot \ln(\text{age})$	-12,78	-16,90; -8,66	$1,18 \cdot 10^{-9}$	
Körpermasseindex	$\text{bmi}^{0,5}$	5,05	3,89; 6,21	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^{-2}	0,37	-0,80; 1,54	0,54	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,79	1,73; 3,85	$2,61 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,99 \cdot 10^{-3}$	$-2,73 \cdot 10^{-3}$; $-1,25 \cdot 10^{-3}$	$1,47 \cdot 10^{-7}$	
	drkt^{-2}	$1,36 \cdot 10^{-3}$	$-3,80 \cdot 10^{-4}$; $3,10 \cdot 10^{-3}$	0,13	
Alkoholkonsum	$\text{drkt}^{-2} \cdot \ln(\text{drkt})$	$3,01 \cdot 10^{-4}$	$-7,38 \cdot 10^{-5}$; $6,75 \cdot 10^{-4}$	0,12	0,03
	hgt^{-2}	1,25	-64,93; 67,44	0,97	
Körperhöhe	hgt^{-2}	1,25	-64,93; 67,44	0,97	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,47	0,20; 0,74	$5,25 \cdot 10^{-4}$	
	glucw^3	0,01	$3,37 \cdot 10^{-3}$; 0,02	$8,09 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$\text{glucw}^3 \cdot \ln(\text{glucw})$	$-4,62 \cdot 10^{-3}$	$-8,11 \cdot 10^{-3}$; $-1,12 \cdot 10^{-3}$	$9,64 \cdot 10^{-3}$	$1,57 \cdot 10^{-4}$
	$\ln(\text{cpd}_k)$	0,20	-0,01; 0,42	0,07	
Zigaretten pro Tag	$\ln(\text{cpd}_k)$	0,20	-0,01; 0,42	0,07	
Geschlecht	sex	-0,55	-0,95; -0,15	$6,77 \cdot 10^{-3}$	
Bluthochdruckbehandlung	hrx	0,87	-0,17; 1,91	0,10	

(b) QIC_u=1557,95

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-5,21	-8,60; -1,82	$2,56 \cdot 10^{-3}$	
Alter	age^{-1}	-5,45	-6,73; -4,16	$< 10^{-15}$	$< 10^{-15}$
	$\text{age}^{-1} \cdot \ln(\text{age})$	-12,78	-16,90; -8,66	$1,18 \cdot 10^{-9}$	
Körpermasseindex	$\text{bmi}^{0,5}$	5,05	3,89; 6,21	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^{-2}	0,37	-0,80; 1,54	0,54	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,79	1,73; 3,85	$2,61 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,99 \cdot 10^{-3}$	$-2,73 \cdot 10^{-3}$; $-1,25 \cdot 10^{-3}$	$1,47 \cdot 10^{-7}$	
	drkt^{-2}	$1,36 \cdot 10^{-3}$	$-3,80 \cdot 10^{-4}$; $3,10 \cdot 10^{-3}$	0,13	
Alkoholkonsum	$\text{drkt}^{-2} \cdot \ln(\text{drkt})$	$3,01 \cdot 10^{-4}$	$-7,38 \cdot 10^{-5}$; $6,75 \cdot 10^{-4}$	0,12	0,03
	hgt^{-2}	1,25	-164,93; 67,44	0,97	
Körperhöhe	hgt^{-2}	1,25	-164,93; 67,44	0,97	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,47	0,20; 0,74	$5,25 \cdot 10^{-4}$	
	glucw^3	0,01	$3,37 \cdot 10^{-3}$; 0,02	$8,09 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$\text{glucw}^3 \cdot \ln(\text{glucw})$	$-4,62 \cdot 10^{-3}$	$-8,11 \cdot 10^{-3}$; $-1,12 \cdot 10^{-3}$	$9,64 \cdot 10^{-3}$	$1,57 \cdot 10^{-4}$
	$\ln(\text{cpd}_k)$	0,20	-0,01; 0,42	0,06	
Zigaretten pro Tag	$\ln(\text{cpd}_k)$	0,20	-0,01; 0,42	0,06	
Geschlecht	sex	-0,55	-0,95; -0,15	$6,77 \cdot 10^{-3}$	
Bluthochdruckbehandlung	hrx	0,87	-0,17; 1,91	0,10	

(c) BQIC_u=1633,00

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-6,47	-8,43; -4,51	$8,95 \cdot 10^{-11}$	
Alter	age	0,63	0,48; 0,77	$< 10^{-15}$	$< 10^{-15}$
	bmi	1,50	1,15; 1,85	$< 10^{-15}$	
Körpermasseindex	bmi	1,50	1,15; 1,85	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^3	$-6,80 \cdot 10^{-3}$	-0,04; 0,02	0,64	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,53	1,53; 3,53	$7,55 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,00 \cdot 10^{-3}$	$-2,74 \cdot 10^{-3}$; $-1,25 \cdot 10^{-3}$	$1,72 \cdot 10^{-7}$	
	drkt^{-2}	$-2,87 \cdot 10^{-5}$	$-6,82 \cdot 10^{-5}$; $1,07 \cdot 10^{-5}$	0,15	
Alkoholkonsum	drkt^{-2}	$-2,87 \cdot 10^{-5}$	$-6,82 \cdot 10^{-5}$; $1,07 \cdot 10^{-5}$	0,15	
	hgt^{-2}	24,17	-41,98; 90,33	0,47	
Körperhöhe	hgt^{-2}	24,17	-41,98; 90,33	0,47	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,43	0,17; 0,69	$1,41 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^{-2}	-87,53	-170,90; -4,18	0,04	
	$\text{cpd}_k^{0,5}$	0,19	-0,09; 0,48	0,18	
Zigaretten pro Tag	$\text{cpd}_k^{0,5}$	0,19	-0,09; 0,48	0,18	
Geschlecht	sex	-0,60	-0,99; -0,21	$2,33 \cdot 10^{-3}$	
Bluthochdruckbehandlung	hrx	0,81	-0,22; 1,85	0,12	

Tabelle A.10: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–4 und 6–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1536,73

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		5,57	0,84; -10,29	0,02	
Alter	age^{-1}	24,14	15,58; 32,69	$3,23 \cdot 10^{-8}$	$< 10^{-15}$
	$age^{-0,5}$	-35,38	-46,15; -24,61	$1,20 \cdot 10^{-10}$	
Körpermasseindex	$bmi^{0,5}$	5,22	4,09; 6,35	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,19	-0,96; 1,33	0,75	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,45	1,19; 3,72	$1,47 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,70 \cdot 10^{-3}$	$-2,46 \cdot 10^{-3}$; $-9,33 \cdot 10^{-4}$	$1,31 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^2$	0,87	-0,01; 1,75	0,05	
Körperhöhe	hgt^3	$1,19 \cdot 10^{-3}$	$-2,40 \cdot 10^{-3}$; $4,78 \cdot 10^{-3}$	0,51	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,40	0,12; 0,68	$4,45 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-10,47	-21,38; 0,45	0,06	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,42	-1,03; 0,19	0,17	
Geschlecht	sex	-0,34	-0,76; 0,07	0,11	
Bluthochdruckbehandlung	hrx	0,64	-0,41; 1,68	0,23	

(b) QIC_u=1536,67

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		5,57	0,84; 10,29	0,02	
Alter	age^{-1}	24,14	15,58; 32,69	$3,23 \cdot 10^{-8}$	$< 10^{-15}$
	$age^{-0,5}$	-35,38	-46,15; -24,61	$1,20 \cdot 10^{-10}$	
Körpermasseindex	$bmi^{0,5}$	5,22	4,09; 6,35	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,19	-0,96; 1,33	0,75	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,45	1,19; 3,72	$1,47 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,70 \cdot 10^{-3}$	$-2,46 \cdot 10^{-3}$; $-9,33 \cdot 10^{-4}$	$1,31 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^2$	0,87	-0,01; 1,75	0,05	
Körperhöhe	hgt^3	$1,19 \cdot 10^{-3}$	$-2,40 \cdot 10^{-3}$; $4,78 \cdot 10^{-3}$	0,51	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,40	0,12; 0,68	$4,45 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-10,47	-21,38; 0,45	0,06	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,42	-1,03; 0,19	0,17	
Geschlecht	sex	-0,34	-0,76; 0,07	0,11	
Bluthochdruckbehandlung	hrx	0,64	-0,41; 1,68	0,23	

(c) BQIC_u=1609,88

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-3,60	-7,71; 0,52	0,09	
Alter	age	0,63	0,49; 0,77	$< 10^{-15}$	
Körpermasseindex	bmi	1,54	1,19; 1,89	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,20	-0,96; 1,37	0,73	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,34	1,10; 3,59	$2,30 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,75 \cdot 10^{-3}$	$-2,49 \cdot 10^{-3}$; $-1,02 \cdot 10^{-3}$	$2,69 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^2$	0,85	$-2,68 \cdot 10^{-3}$; 1,70	0,51	
Körperhöhe	hgt^{-2}	2,30	-68,68; 73,28	0,95	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,39	0,11; 0,67	$5,64 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-11,46	-22,35; -0,58	0,04	
Zigaretten pro Tag	$cpd_k^{-0,5}$	-0,33	-0,93; 0,26	0,27	
Geschlecht	sex	-0,44	-0,86; -0,03	0,04	
Bluthochdruckbehandlung	hrx	0,64	-0,38; 1,66	0,22	

Tabelle A.11: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–3 und 5–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1561,68

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		7,98	3,95; 12,00	$1,02 \cdot 10^{-4}$	
Alter	age ⁻¹	22,61	13,54; 31,68	$1,02 \cdot 10^{-6}$	
	age ^{-0,5}	-33,13	-44,48; -21,78	$1,06 \cdot 10^{-8}$	$5,33 \cdot 10^{-13}$
Körpermasseindex	ln(bmi)	4,38	3,44; 5,31	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol ⁻²	0,13	-1,05; 1,31	0,83	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	2,19	1,18; 3,21	$2,39 \cdot 10^{-5}$	
Familiengröße	fs ³	$8,96 \cdot 10^{-3}$	$-8,86 \cdot 10^{-3}$; 0,03	0,32	
	fs ³ · ln(fs)	$-5,12 \cdot 10^{-3}$	-0,01; $3,23 \cdot 10^{-3}$	0,23	0,14
Alkoholkonsum	drkt ³	0,86	-0,11; 1,83	0,08	
	drkt ³ · ln(drkt)	-1,77	-3,81; 0,27	0,09	0,01
Körperhöhe	hgt ⁻²	-38,79	-105,00; 27,40	0,25	
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,41	0,15; 0,67	$2,31 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ⁻²	-55,03	-140,00; 29,90	0,20	
Zigaretten pro Tag	cpd_k ⁻²	-0,30	-0,60; $-2,08 \cdot 10^{-3}$	0,05	
Geschlecht	sex	-0,28	-0,67; 0,12	0,17	
Bluthochdruckbehandlung	hrx	0,54	-0,46; 1,54	0,29	

(b) QIC_u=1561,42

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		5,44	0,65; 10,23	0,03	
Alter	age ⁻¹	22,98	13,80; 32,16	$9,27 \cdot 10^{-7}$	
	age ^{-0,5}	-33,49	-44,92; -22,07	$9,19 \cdot 10^{-9}$	$4,38 \cdot 10^{-13}$
Körpermasseindex	ln(bmi)	4,40	3,46; 5,34	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol ⁻²	0,28	-0,87; 1,44	0,63	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ⁻²	-0,06	-0,12; $2,79 \cdot 10^{-3}$	0,06	
	hdl ³	1,68	0,54; 2,82	$4,00 \cdot 10^{-3}$	$2,15 \cdot 10^{-3}$
Familiengröße	fs ³	$-1,74 \cdot 10^{-3}$	$-2,60 \cdot 10^{-3}$; $-8,73 \cdot 10^{-4}$	$8,26 \cdot 10^{-5}$	
Alkoholkonsum	drkt ⁻²	$-2,52 \cdot 10^{-5}$	$-6,57 \cdot 10^{-5}$; $1,53 \cdot 10^{-5}$	0,22	
Körperhöhe	hgt ⁻²	-36,70	-103,50; 30,08	0,28	
Nüchtern-Serum-Triglycerid-Konzentration	tgw ^{0,5}	1,02	0,45; 1,60	$4,32 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ³	0,01	$-2,61 \cdot 10^{-3}$; 0,02	0,12	
	glucw ³ · ln(glucw)	$-3,67 \cdot 10^{-3}$	$-8,43 \cdot 10^{-3}$; $1,09 \cdot 10^{-3}$	0,13	0,03
Zigaretten pro Tag	cpd_k ⁻¹	-0,44	-0,83; -0,05	0,03	
Geschlecht	sex	-0,30	-0,70; 0,09	0,13	
Bluthochdruckbehandlung	hrx	0,63	-0,38; 1,64	0,22	

(c) BQIC_u=1633,93

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-4,92	-7,52; -2,32	$2,05 \cdot 10^{-4}$	
Alter	age	0,59	0,45; 0,73	$< 10^{-15}$	
Körpermasseindex	ln(bmi)	4,22	3,29; 5,15	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol ⁻²	0,18	-1,03; 1,39	0,77	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	2,13	1,14; 3,12	$2,37 \cdot 10^{-5}$	
Familiengröße	fs ³	$-1,68 \cdot 10^{-3}$	$-2,55 \cdot 10^{-3}$; $-8,10 \cdot 10^{-4}$	$1,56 \cdot 10^{-4}$	
Alkoholkonsum	drkt	0,50	-0,23; 1,23	0,18	
Körperhöhe	hgt ⁻²	-16,57	-83,43; 50,29	0,63	
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,41	0,14; 0,67	$2,62 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ⁻¹	-12,86	-29,93; 4,21	0,14	
Zigaretten pro Tag	cpd_k ⁻²	-0,26	-0,56; 0,04	0,09	
Geschlecht	sex	-0,35	-0,77; 0,06	0,09	
Bluthochdruckbehandlung	hrx	0,56	-0,41; 1,54	0,26	

Tabelle A.12: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1–2 und 4–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1596,09$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		4,37	0,30; 8,44	0,04	
Alter	age^{-1}	23,04	14,42; 31,66	$1,62 \cdot 10^{-7}$	$1,33 \cdot 10^{-14}$
	$age^{-0,5}$	-33,52	-44,33; -22,71	$1,22 \cdot 10^{-9}$	
Körpermasseindex	$bmi^{0,5}$	4,97	3,88; 6,06	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,20	-0,95; 1,34	0,74	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,23	1,24; 3,23	$1,16 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,73 \cdot 10^{-3}$	$-2,58 \cdot 10^{-3}$; $-8,75 \cdot 10^{-4}$	$6,98 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^3$	0,94	-0,02; 1,89	0,06	$6,20 \cdot 10^{-3}$
	$drkt^3 \cdot \ln(drkt)$	-1,95	-3,95; 0,05	0,06	
Körperhöhe	hgt^{-2}	-35,23	-102,70; 32,28	0,31	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,47	0,21; 0,73	$4,07 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-2}$	-70,48	-154,00; 13,07	0,10	
Zigaretten pro Tag	cpd_k^{-1}	-0,34	-0,73; 0,04	0,08	
Geschlecht	sex	-0,35	-0,74; 0,04	0,08	
Bluthochdruckbehandlung	hrx	0,70	-0,29; 1,69	0,16	

(b) $QIC_u=1596,43$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		14,48	6,97; 21,99	$1,57 \cdot 10^{-4}$	
Alter	$age^{-0,5}$	-20,15	-26,19; -14,11	$6,09 \cdot 10^{-11}$	$7,61 \cdot 10^{-14}$
	$age^{-0,5} \cdot \ln(age)$	-15,02	-20,83; -9,22	$3,90 \cdot 10^{-7}$	
Körpermasseindex	$bmi^{0,5}$	4,96	3,87; 6,05	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,18	-0,96; 1,33	0,75	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,19	1,19; 3,20	$1,81 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,75 \cdot 10^{-3}$	$-2,60 \cdot 10^{-3}$; $-8,95 \cdot 10^{-4}$	$5,93 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^2$	2,12	-0,12; 4,36	0,06	0,01
	$drkt^3$	-1,30	-2,81; 0,22	0,09	
Körperhöhe	hgt^{-2}	-32,86	-100,30; 34,60	0,34	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,20; 0,72	$5,10 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-14,04	-30,89; 2,82	0,10	
Zigaretten pro Tag	cpd_k^{-1}	-0,34	-0,72; 0,05	0,09	
Geschlecht	sex	-0,35	-0,74; 0,04	0,08	
Bluthochdruckbehandlung	hrx	0,68	-0,31; 1,68	0,18	

(c) $BQIC_u=1669,43$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-7,43	-10,39; -4,47	$8,81 \cdot 10^{-7}$	
Alter	age^2	0,08	0,06; 0,10	$< 10^{-15}$	
Körpermasseindex	$bmi^{0,5}$	4,80	3,71; 5,90	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	0,20	-0,97; 1,36	0,74	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,16	1,17; 3,14	$1,84 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,76 \cdot 10^{-3}$	$-2,61 \cdot 10^{-3}$; $-9,10 \cdot 10^{-4}$	$4,91 \cdot 10^{-5}$	
Alkoholkonsum	drkt	0,60	-0,14; 1,35	0,11	
Körperhöhe	hgt^{-2}	-13,29	-80,30; 53,72	0,70	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,20; 0,72	$5,34 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-15,33	-32,07; 1,41	0,07	
Zigaretten pro Tag	cpd_k^{-1}	-0,32	-0,70; 0,06	0,10	
Geschlecht	sex	-0,43	-0,83; -0,02	0,04	
Bluthochdruckbehandlung	hrx	0,69	-0,30; 1,67	0,17	

Tabelle A.13: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 1 und 3–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1561,06$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		3,08	0,22; 5,94	0,03	
Alter	age^{-1}	-5,00	-6,21; -3,79	$< 10^{-15}$	$1,07 \cdot 10^{-15}$
	$age^{-1} \cdot \ln(age)$	-11,54	-15,68; -7,40	$4,62 \cdot 10^{-8}$	
Körpermasseindex	bmi	1,46	1,12; 1,80	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-2}$	-0,05	-1,18; 1,08	0,93	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,79	1,62; 3,96	$3,17 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,69 \cdot 10^{-3}$	$-2,53 \cdot 10^{-3}$; $-8,45 \cdot 10^{-4}$	$8,52 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^3$	0,93	-0,02; 1,88	0,06	$5,12 \cdot 10^{-3}$
	$drkt^3 \cdot \ln(drkt)$	-2,01	-4,00; -0,03	0,05	
Körperhöhe	hgt^{-2}	-38,16	-106,90; 30,55	0,28	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,53	0,27; 0,79	$8,28 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-20,03	-37,01; -3,05	0,02	
Zigaretten pro Tag	cpd_k	0,06	-0,03; 0,16	0,18	
Geschlecht	sex	-0,29	-0,69; 0,12	0,16	
Bluthochdruckbehandlung	hrx	0,64	-0,38; 1,66	0,22	

(b) $QIC_u=1561,65$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		9,63	4,68; 14,57	$1,38 \cdot 10^{-4}$	
Alter	age^{-1}	20,36	11,27; 29,45	$1,13 \cdot 10^{-5}$	$6,52 \cdot 10^{-11}$
	$age^{-0,5}$	-30,34	-41,67; -19,01	$1,53 \cdot 10^{-7}$	
Körpermasseindex	bmi	1,46	1,12; 1,80	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-1}$	-0,06	-1,41; 1,30	0,93	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,75	1,57; 3,93	$5,26 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,70 \cdot 10^{-3}$	$-2,55 \cdot 10^{-3}$; $-8,53 \cdot 10^{-4}$	$8,35 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^2$	2,22	-0,01; 4,44	0,05	$7,17 \cdot 10^{-3}$
	$drkt^3$	-1,39	-2,90; 0,11	0,07	
Körperhöhe	hgt^{-2}	-36,37	-105,20; 32,49	0,30	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,52	0,25; 0,79	$1,42 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-12,40	-23,15; -1,65	0,02	
Zigaretten pro Tag	$cpd_k^{0,5}$	0,20	-0,10; 0,49	0,19	
Geschlecht	sex	-0,28	-0,69; 0,13	0,18	
Bluthochdruckbehandlung	hrx	0,63	-0,39; 1,65	0,23	

(c) $BQIC_u=1633,77$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Intercept	Intercept	-2,66	-6,75; 1,44	0,20	
Alter	age	0,57	0,43; 0,71	$< 10^{-15}$	
Körpermasseindex	bmi	1,41	1,07; 1,75	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	$chol^{-1}$	-0,02	-1,37; 1,32	0,97	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,73	1,57; 3,90	$4,31 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,70 \cdot 10^{-3}$	$-2,51 \cdot 10^{-3}$; $-8,83 \cdot 10^{-4}$	$4,41 \cdot 10^{-5}$	
Alkoholkonsum	drkt	0,55	-0,20; 1,29	0,15	
Körperhöhe	hgt^{-2}	-15,94	-84,69; 52,81	0,65	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,52	0,25; 0,79	$1,38 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-0,5}$	-13,27	-23,95; -2,58	0,01	
Zigaretten pro Tag	cpd_k	0,06	-0,03; 0,15	0,20	
Geschlecht	sex	-0,37	-0,79; 0,05	0,08	
Bluthochdruckbehandlung	hrx	0,65	-0,35; 1,65	0,20	

Tabelle A.14: Ergebnisse der vollständigen Modelle unter Verwendung der Teildatensätze 2–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1547,34

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		10,54	6,50; 14,58	$3,14 \cdot 10^{-7}$	
Alter	age^{-1}	25,46	16,58; 34,35	$1,93 \cdot 10^{-8}$	$< 10^{-15}$
	$\text{age}^{-0,5}$	-36,80	-47,97; -25,63	$1,08 \cdot 10^{-10}$	
Körpermasseindex	bmi	1,59	1,22; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^{-2}	0,02	-1,20; 1,24	0,98	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,09	1,06; 3,11	$6,58 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,76 \cdot 10^{-3}$	$-2,58 \cdot 10^{-3}$; $-9,47 \cdot 10^{-4}$	$2,33 \cdot 10^{-5}$	
	drkt^3	0,89	-0,07; 1,85	0,07	0,01
Alkoholkonsum	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-1,86	-3,89; 0,18	0,07	
Körperhöhe	hgt^{-2}	-52,80	-121,30; 15,70	0,13	
Nüchtern-Serum-Triglycerid-Konzentration	$\text{tgw}^{-0,5}$	-0,69	-1,15; -0,23	$3,22 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^{-2}	-88,12	-167,60; -8,69	0,03	
Zigaretten pro Tag	cpd_k^{-2}	-0,28	-0,57; 0,02	0,07	
Geschlecht	sex	-0,23	-0,63; 0,17	0,25	
Bluthochdruckbehandlung	hrx	0,57	-0,45; 1,59	0,27	

(b) QIC_u=1547,83

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		19,94	12,22; 27,66	$4,17 \cdot 10^{-7}$	
Alter	$\text{age}^{-0,5}$	-21,78	-28,06; -15,50	$1,06 \cdot 10^{-11}$	$4,53 \cdot 10^{-15}$
	$\text{age}^{-0,5} \cdot \ln(\text{age})$	-16,31	-22,31; -10,32	$9,52 \cdot 10^{-8}$	
Körpermasseindex	bmi	1,58	1,21; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^{-2}	0,04	-1,18; 1,26	0,95	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,09	1,06; 3,13	$7,26 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,80 \cdot 10^{-3}$	$-2,64 \cdot 10^{-3}$; $-9,62 \cdot 10^{-4}$	$2,51 \cdot 10^{-5}$	
	drkt^3	0,88	-0,08; 1,85	0,07	0,01
Alkoholkonsum	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-1,83	-3,86; 0,21	0,08	
Körperhöhe	hgt^{-2}	-50,67	-119,20; 17,84	0,15	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,41	0,14; 0,68	$2,71 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^{-2}	-87,84	-167,20; -8,44	0,03	
Zigaretten pro Tag	cpd_k^{-2}	-0,28	-0,58; 0,02	0,07	
Geschlecht	sex	-0,24	-0,64; 0,16	0,24	
Bluthochdruckbehandlung	hrx	0,55	-0,47; 1,57	0,29	

(c) BQIC_u=1622,15

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-5,32	-7,77; -2,87	$2,02 \cdot 10^{-5}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	bmi	1,53	1,16; 1,89	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration	chol^{-2}	0,06	-1,19; 1,31	0,93	
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,06	1,04; 3,08	$7,40 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,83 \cdot 10^{-3}$	$-2,66 \cdot 10^{-3}$; $-9,99 \cdot 10^{-4}$	$1,63 \cdot 10^{-5}$	
	drkt^3	0,47	-0,26; 1,20	0,21	0,41
Körperhöhe	hgt^2	0,01	-0,02; 0,05	0,41	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,41	0,14; 0,68	$3,04 \cdot 10^{-3}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^{-2}	-94,67	-173,60; -15,71	0,02	
Zigaretten pro Tag	cpd_k^{-2}	-0,27	-0,56; 0,03	0,08	
Geschlecht	sex	-0,32	-0,74; 0,10	0,13	
Bluthochdruckbehandlung	hrx	0,57	-0,44; 1,59	0,27	

Tabelle A.15: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–9 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1514,65$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		16,47	9,82; 23,11	$1,21 \cdot 10^{-6}$	
Alter	$age^{-0,5}$	-20,09	-25,82; -14,36	$6,45 \cdot 10^{-12}$	$3,51 \cdot 10^{-14}$
	$age^{-0,5} \cdot \ln(age)$	-14,00	-19,33; -8,67	$2,65 \cdot 10^{-7}$	
Körpermasseindex	bmi^3	0,27	0,17; 0,36	$4,91 \cdot 10^{-8}$	$3,33 \cdot 10^{-11}$
	$bmi^3 \cdot \ln(bmi)$	-0,14	-0,21; -0,08	$8,11 \cdot 10^{-6}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,76	0,83; 2,69	$2,17 \cdot 10^{-4}$	
Familiengröße	fs^3	$-2,05 \cdot 10^{-3}$	$-2,85 \cdot 10^{-3}$; $-1,25 \cdot 10^{-3}$	$4,88 \cdot 10^{-7}$	$1,57 \cdot 10^{-4}$
	$drkt^3$	1,29	0,33; 2,24	$8,06 \cdot 10^{-3}$	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,63	-4,62; -0,64	$9,68 \cdot 10^{-3}$	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,21; 0,72	$3,44 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^2$	$8,63 \cdot 10^{-3}$	$3,82 \cdot 10^{-4}$; 0,02	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,33	-0,65; -0,01	0,04	
Bluthochdruckbehandlung					

(b) $QIC_u=1514,75$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		15,12	8,54; 21,70	$6,69 \cdot 10^{-6}$	
Alter	$age^{-0,5}$	-20,23	-25,98; -14,49	$5,12 \cdot 10^{-12}$	$2,86 \cdot 10^{-14}$
	$age^{-0,5} \cdot \ln(age)$	-14,13	-19,49; -8,77	$2,39 \cdot 10^{-7}$	
Körpermasseindex	bmi	1,59	1,24; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,75	0,82; 2,68	$2,28 \cdot 10^{-4}$	
Familiengröße	fs^3	$-2,06 \cdot 10^{-3}$	$-2,85 \cdot 10^{-3}$; $-1,26 \cdot 10^{-3}$	$4,39 \cdot 10^{-7}$	$1,66 \cdot 10^{-4}$
	$drkt^3$	1,28	0,33; 2,23	$8,11 \cdot 10^{-3}$	
Alkoholkonsum	$drkt^3 \cdot \ln(drkt)$	-2,60	-4,59; -0,62	0,01	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,21; 0,72	$3,42 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^3$	$5,35 \cdot 10^{-4}$	$-1,13 \cdot 10^{-6}$; $1,07 \cdot 10^{-3}$	0,05	
Zigaretten pro Tag					
Geschlecht	sex	-0,35	-0,66; -0,41	0,03	
Bluthochdruckbehandlung					

(c) $BQIC_u=1566,39$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-6,37	-7,50; -5,23	$< 10^{-15}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	bmi	1,57	1,22; 1,93	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,73	0,80; 2,66	$2,58 \cdot 10^{-4}$	
Familiengröße	fs^3	$-2,09 \cdot 10^{-3}$	$-2,91 \cdot 10^{-3}$; $-1,28 \cdot 10^{-3}$	$4,68 \cdot 10^{-7}$	$0,03$
	$drkt$	0,81	0,07; 1,55		
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,20; 0,71	$4,61 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^3$	$5,47 \cdot 10^{-4}$	$1,96 \cdot 10^{-5}$; $1,07 \cdot 10^{-3}$	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,35	-0,66; -0,04	0,03	
Bluthochdruckbehandlung					

Tabelle A.16: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–8 und 10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1471,42

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		16,32	9,05; 23,59	$1,09 \cdot 10^{-5}$	
Alter	$\text{age}^{-0,5}$	-20,02	-25,78; -14,26	$9,82 \cdot 10^{-12}$	$5,22 \cdot 10^{-14}$
	$\text{age}^{-0,5} \cdot \ln(\text{age})$	-13,99	-19,36; -8,63	$3,23 \cdot 10^{-7}$	
Körpermasseindex	$\text{bmi}^{0,5}$	4,64	3,51; 5,77	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,13	1,14; 3,13	$2,78 \cdot 10^{-5}$	
Familiengröße					
Alkoholkonsum	drkt^3	0,94	-0,03; 1,91	0,06	$3,62 \cdot 10^{-3}$
	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-2,03	-3,85; -0,21	0,03	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,43	0,19; 0,68	$5,46 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$\text{glucw}^{-0,5}$	-13,49	-24,19; -2,79	0,01	
Zigaretten pro Tag					
Geschlecht	sex	-0,45	-0,76; -0,13	$5,67 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(b) QIC_u=1481,30

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		6,77	-1,39; 14,93	0,10	
Alter	$\text{age}^{-0,5}$	-20,72	-26,67; -14,77	$8,90 \cdot 10^{-12}$	$2,77 \cdot 10^{-14}$
	$\text{age}^{-0,5} \cdot \ln(\text{age})$	-14,79	-20,40; -9,18	$2,36 \cdot 10^{-7}$	
Körpermasseindex	$\text{bmi}^{0,5}$	4,79	3,65; 5,93	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,85	1,67; 4,03	$2,13 \cdot 10^{-6}$	
Familiengröße					
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,47	0,22; 0,73	$3,04 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(\text{glucw})$	2,33	0,65; 4,00	$6,45 \cdot 10^{-3}$	
Zigaretten pro Tag	cpd_k^{-2}	-0,34	-0,66; -0,02	0,04	
Geschlecht	sex	-0,54	-0,84; -0,23	$5,46 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

(c) BQIC_u=1521,45

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-15,44	-19,49; -11,40	$6,88 \cdot 10^{-14}$	
Alter	age	0,66	0,51; 0,80	$< 10^{-15}$	$7,11 \cdot 10^{-15}$
	$\text{bmi}^{0,5}$	4,51	3,38; 5,65	$7,11 \cdot 10^{-15}$	
Körpermasseindex					
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,33	1,34; 3,32	$4,08 \cdot 10^{-6}$	
Familiengröße					
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,46	0,21; 0,71	$2,55 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(\text{glucw})$	2,34	0,69; 4,00	$5,61 \cdot 10^{-3}$	
Zigaretten pro Tag					
Geschlecht	sex	-0,53	-0,83; -0,23	$5,53 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

Tabelle A.17: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–7 und 9–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1489,38

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		13,62	5,21; 22,03	$1,51 \cdot 10^{-3}$	
Alter	$\text{age}^{-0,5}$	-21,76	-27,56; -15,97	$1,80 \cdot 10^{-13}$	$< 10^{-15}$
	$\text{age}^{-0,5} \cdot \ln(\text{age})$	-16,04	-21,59; -10,48	$1,51 \cdot 10^{-8}$	
Körpermasseindex	$\ln(\text{bmi})$	3,91	3,03; 4,79	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,44	1,48; 3,41	$7,27 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,98 \cdot 10^{-3}$	$-2,85 \cdot 10^{-3}$; $-1,11 \cdot 10^{-3}$	$8,82 \cdot 10^{-6}$	
Alkoholkonsum	drkt^3	1,33	0,33; 2,32	$8,96 \cdot 10^{-3}$	$2,30 \cdot 10^{-4}$
	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-2,61	-4,66; -0,56	0,01	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,59	0,35; 0,83	$1,90 \cdot 10^{-6}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(\text{glucw})$	1,90	0,11; 3,69	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,37	-0,68; -0,06	0,02	
Bluthochdruckbehandlung					

(b) QIC_u=1490,14

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		14,24	6,16; 22,33	$5,58 \cdot 10^{-4}$	
Alter	$\text{age}^{-0,5}$	-21,77	-27,56; -15,97	$1,79 \cdot 10^{-13}$	$< 10^{-15}$
	$\text{age}^{-0,5} \cdot \ln(\text{age})$	-16,04	-21,59; -10,49	$1,51 \cdot 10^{-8}$	
Körpermasseindex	$\ln(\text{bmi})$	3,91	3,03; 4,79	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,44	1,48; 3,41	$7,39 \cdot 10^{-7}$	
Familiengröße	fs^3	$-1,98 \cdot 10^{-3}$	$-2,85 \cdot 10^{-3}$; $-1,11 \cdot 10^{-3}$	$8,97 \cdot 10^{-6}$	
Alkoholkonsum	drkt^3	1,32	0,33; 2,32	$9,03 \cdot 10^{-3}$	$2,34 \cdot 10^{-4}$
	$\text{drkt}^3 \cdot \ln(\text{drkt})$	-2,61	-4,66; -0,56	0,01	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,59	0,35; 0,83	$1,98 \cdot 10^{-6}$	
Nüchtern-Serum-Glukose-Konzentration	$\text{glucw}0,5$	1,19	0,06; 2,32	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,37	-0,69; -0,06	0,02	
Bluthochdruckbehandlung					

(c) BQIC_u=1555,21

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-14,53	-18,80; -10,26	$2,48 \cdot 10^{-11}$	
Alter	age^2	0,09	0,07; 0,11	$< 10^{-15}$	
Körpermasseindex	$\text{bmi}^{0,5}$	4,79	3,68; 5,90	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,77	1,83; 3,71	$7,80 \cdot 10^{-9}$	
Familiengröße	fs^3	$-1,96 \cdot 10^{-3}$	$-2,80 \cdot 10^{-3}$; $-1,12 \cdot 10^{-3}$	$4,69 \cdot 10^{-6}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,58	0,33; 0,83	$4,79 \cdot 10^{-6}$	
Nüchtern-Serum-Glukose-Konzentration	$\ln(\text{glucw})$	2,11	0,32; 3,90	0,02	
Zigaretten pro Tag	$\ln(\text{cpd}_k)$	0,23	0,01; 0,45	0,04	
Geschlecht	sex	-0,50	-0,80; -0,20	$1,21 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

Tabelle A.18: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–6 und 8–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1577,09$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		7,34	3,69; 10,99	$8,06 \cdot 10^{-5}$	
Alter	age^{-1}	23,15	14,83; 31,47	$4,92 \cdot 10^{-8}$	$1,22 \cdot 10^{-15}$
	$age^{-0,5}$	-34,61	-45,12; -24,09	$1,11 \cdot 10^{-10}$	
Körpermasseindex	bmi	1,58	1,24; 1,92	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,75	1,66; 3,85	$8,60 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,05 \cdot 10^{-3}$	$-2,84 \cdot 10^{-3}$; $-1,26 \cdot 10^{-3}$	$3,61 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,97	0,46; 1,47	$1,70 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-17,88	-34,90; -0,86	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,52	-0,82; -0,23	$4,61 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

(b) $QIC_u=1576,77$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		16,61	9,68; 23,54	$2,62 \cdot 10^{-6}$	
Alter	$age^{-0,5}$	-20,99	-26,91; -15,07	$3,75 \cdot 10^{-12}$	$2,52 \cdot 10^{-14}$
	$age^{-0,5} \cdot \ln(age)$	-14,87	-20,49; -9,24	$2,25 \cdot 10^{-7}$	
Körpermasseindex	bmi	1,58	1,24; 1,91	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,75	1,65; 3,85	$9,47 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,07 \cdot 10^{-3}$	$-2,87 \cdot 10^{-3}$; $-1,27 \cdot 10^{-3}$	$3,89 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,97	0,46; 1,48	$1,70 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-17,76	-34,78; -0,75	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,53	-0,82; -0,23	$4,36 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

(c) $BQIC_u=1622,46$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-6,61	-8,92; -4,29	$2,20 \cdot 10^{-8}$	
Alter	age	0,66	0,52; 0,80	$< 10^{-15}$	
Körpermasseindex	bmi	1,53	1,19; 1,87	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,71	1,62; 3,79	$1,11 \cdot 10^{-6}$	
Familiengröße	fs^3	$-2,00 \cdot 10^{-3}$	$-2,77 \cdot 10^{-3}$; $-1,23 \cdot 10^{-3}$	$3,62 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$tgw^{0,5}$	0,96	0,46; 1,46	$1,79 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-1}$	-18,85	-35,80; -1,90	0,03	
Zigaretten pro Tag					
Geschlecht	sex	-0,52	-0,81; -0,23	$4,82 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

Tabelle A.19: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–5 und 7–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1557,75

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-0,53	-3,00; 1,94	0,67	
Alter	age^{-1}	-5,63	-6,93; -4,32	$< 10^{-15}$	$< 10^{-15}$
	$\text{age}^{-1} \cdot \ln(\text{age})$	-12,73	-16,51; -8,95	$4,23 \cdot 10^{-11}$	
Körpermasseindex	bmi	1,56	1,21; 1,90	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,49	1,53; 3,46	$4,34 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,07 \cdot 10^{-3}$	$-2,82 \cdot 10^{-3}$; $-1,32 \cdot 10^{-3}$	$5,70 \cdot 10^{-8}$	
Alkoholkonsum	drkt^{-2}	$-3,98 \cdot 10^{-5}$	$-7,84 \cdot 10^{-5}$; $2,17 \cdot 10^{-7}$	0,05	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,46	0,23; 0,70	$1,34 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^3	0,01	$2,94 \cdot 10^{-3}$; 0,02	0,01	$2,39 \cdot 10^{-4}$
	$\text{glucw}^3 \cdot \ln(\text{glucw})$	$-4,38 \cdot 10^{-3}$	$-7,80 \cdot 10^{-3}$; $-9,70 \cdot 10^{-4}$	0,01	
Zigaretten pro Tag					
Geschlecht	sex	-0,48	-0,78; -0,19	$1,88 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(b) QIC_u=1557,78

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-0,53	-3,00; 1,94	0,67	
Alter	age^{-1}	-5,63	-6,93; -4,32	$< 10^{-15}$	$< 10^{-15}$
	$\text{age}^{-1} \cdot \ln(\text{age})$	-12,73	-16,51; -8,95	$9,23 \cdot 10^{-11}$	
Körpermasseindex	$\text{bmi}^{0,5}$	1,56	1,21; 1,90	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,49	1,53; 3,46	$4,34 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,07 \cdot 10^{-3}$	$-2,82 \cdot 10^{-3}$; $-1,32 \cdot 10^{-3}$	$5,70 \cdot 10^{-8}$	
Alkoholkonsum	drkt^{-2}	$-3,98 \cdot 10^{-5}$	$-7,94 \cdot 10^{-5}$; $-2,17 \cdot 10^{-7}$	0,05	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,46	0,23; 0,70	$1,34 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^3	0,01	$2,94 \cdot 10^{-3}$; 0,02	0,01	$2,39 \cdot 10^{-4}$
	$\text{glucw}^3 \cdot \ln(\text{glucw})$	$-4,38 \cdot 10^{-3}$	$-7,80 \cdot 10^{-3}$; $-9,70 \cdot 10^{-4}$	0,01	
Zigaretten pro Tag					
Geschlecht	sex	-0,48	-0,78; -0,19	$1,18 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(c) BQIC_u=1619,22

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-5,86	-7,35; -4,37	$1,20 \cdot 10^{-14}$	
Alter	age	0,66	0,52; 0,80	$< 10^{-15}$	
Körpermasseindex	bmi	1,51	1,17; 1,85	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,58	1,61; 3,55	$1,94 \cdot 10^{-7}$	
Familiengröße	fs^3	$-2,05 \cdot 10^{-3}$	$-2,82 \cdot 10^{-3}$; $-1,29 \cdot 10^{-3}$	$1,35 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(\text{tgw})$	0,44	0,20; 0,68	$2,97 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	glucw^{-2}	-87,85	-170,50; -5,19	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,56	-0,85; -0,28	$1,25 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

Tabelle A.20: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–4 und 6–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1532,73					
Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		5,80	1,21; 10,39	0,01	
Alter	age ⁻¹	22,96	15,26; 30,65	5,01 · 10 ⁻⁹	< 10 ⁻¹⁵
	age ^{-0,5}	-34,13	-43,89; -24,37	7,31 · 10 ⁻¹²	
Körpermasseindex	bmi ^{0,5}	5,12	4,02; 6,22	< 10 ⁻¹⁵	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,28	1,12; 3,43	1,11 · 10 ⁻⁴	
Familiengröße	fs ³	-1,79 · 10 ⁻³	-2,53 · 10 ⁻³ ; -1,04 · 10 ⁻³	2,45 · 10 ⁻⁶	
Alkoholkonsum	drkt ²	0,99	0,11; 1,86	0,03	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,41	0,16; 0,67	1,39 · 10 ⁻³	
Nüchtern-Serum-Glukose-Konzentration	glucw ^{-0,5}	-11,00	-21,86; -0,15	0,05	
Zigaretten pro Tag					
Geschlecht	sex	-0,42	-0,73; -0,11	8,53 · 10 ⁻³	
Bluthochdruckbehandlung					

(b) QIC _u =1572,10					
Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		2,74	-0,57; 6,05	0,10	
Alter	age ⁻¹	23,80	16,22; 31,38	7,43 · 10 ⁻¹⁰	< 10 ⁻¹⁵
	age ^{-0,5}	-35,17	-44,77; -25,57	6,88 · 10 ⁻¹³	
Körpermasseindex	bmi ^{0,5}	5,41	4,34; 6,48	< 10 ⁻¹⁵	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	1,76	0,74; 2,77	6,96 · 10 ⁻⁴	
Familiengröße	fs ³	-1,95 · 10 ⁻³	-2,71 · 10 ⁻³ ; -1,19 · 10 ⁻³	4,85 · 10 ⁻⁷	
Alkoholkonsum	drkt	0,88	0,14; 1,62	0,02	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,37	0,13; 0,62	2,43 · 10 ⁻³	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag					
Geschlecht	sex	-0,40	-0,71; -0,10	9,09 · 10 ⁻³	
Bluthochdruckbehandlung					

(c) BQIC _u =1584,55					
Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-10,89	-14,85; -6,93	6,92 · 10 ⁻⁸	
Alter	age	0,64	0,50; 0,78	< 10 ⁻¹⁵	< 10 ⁻¹⁵
Körpermasseindex	bmi	1,52	1,18; 1,86	< 10 ⁻¹⁵	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	1,93	0,93; 2,92	1,44 · 10 ⁻⁴	
Familiengröße	fs ³	-1,78 · 10 ⁻³	-2,52 · 10 ⁻³ ; -1,04 · 10 ⁻³	2,27 · 10 ⁻⁶	
Alkoholkonsum	drkt ²	0,97	0,12; 1,82	0,03	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,37	0,12; 0,62	3,31 · 10 ⁻³	
Nüchtern-Serum-Glukose-Konzentration	ln(glucw)	1,82	0,11; 3,53	0,04	
Zigaretten pro Tag					
Geschlecht	sex	-0,42	-0,73; -0,12	6,96 · 10 ⁻³	
Bluthochdruckbehandlung					

Tabelle A.21: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–3 und 5–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1611,88

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		6,52	3,33; 9,72	$6,39 \cdot 10^{-5}$	
Alter	age ⁻¹	21,41	13,20; 29,62	$3,20 \cdot 10^{-7}$	$5,15 \cdot 10^{-14}$
	age ^{-0,5}	-31,98	-42,25; -21,71	$1,04 \cdot 10^{-9}$	
Körpermasseindex	ln(bmi)	4,51	3,64; 5,38	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	2,16	1,23; 3,08	$5,52 \cdot 10^{-6}$	
Familiengröße	fs ³	$-1,88 \cdot 10^{-3}$	$-2,69 \cdot 10^{-3}$; $-1,06 \cdot 10^{-3}$	$6,73 \cdot 10^{-6}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,43	0,18; 0,67	$5,83 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag	cpd_k ⁻¹	-0,41	-0,80; -0,02	0,04	
Geschlecht	sex	-0,51	-0,79; -0,23	$3,22 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

(b) QIC_u=1572,30

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		12,53	5,06; 20,01	$1,02 \cdot 10^{-3}$	
Alter	age ^{-0,5}	-18,85	-24,84; -12,86	$6,83 \cdot 10^{-10}$	$1,28 \cdot 10^{-11}$
	age ^{-0,5} · ln(age)	-13,36	-19,10; -7,62	$5,05 \cdot 10^{-6}$	
Körpermasseindex	ln(bmi)	4,31	3,41; 5,22	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ⁻²	-0,06	-0,12; $2,41 \cdot 10^{-3}$	0,06	$1,98 \cdot 10^{-3}$
	hdl ³	1,66	0,56; 2,76	$3,14 \cdot 10^{-3}$	
Familiengröße	fs ³	$-1,78 \cdot 10^{-3}$	$-2,64 \cdot 10^{-3}$; $-9,29 \cdot 10^{-4}$	$4,32 \cdot 10^{-5}$	
Alkoholkonsum					
Nüchtern-Serum-Triglycerid-Konzentration	tgw ^{0,5}	1,01	0,47; 1,54	$2,47 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ³	$9,97 \cdot 10^{-3}$	$-3,03 \cdot 10^{-3}$; 0,02	0,13	0,04
	glucw ³ · ln(glucw)	$-3,54 \cdot 10^{-3}$	$-8,34 \cdot 10^{-3}$; $1,26 \cdot 10^{-3}$	0,15	
Zigaretten pro Tag	cpd_k ⁻¹	-0,45	-0,84; -0,07	0,02	
Geschlecht	sex	-0,47	-0,77; -0,18	$1,66 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(c) BQIC_u=1653,27

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-11,25	-13,14; -9,35	$< 10^{-15}$	
Alter	age	0,63	0,49; 0,77	$< 10^{-15}$	
Körpermasseindex	bmi ^{0,5}	5,38	4,31; 6,46	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl ²	2,06	1,15; 2,98	$1,01 \cdot 10^{-5}$	
Familiengröße	fs ³	$-1,92 \cdot 10^{-3}$	$-2,72 \cdot 10^{-3}$; $-1,12 \cdot 10^{-3}$	$2,48 \cdot 10^{-6}$	
Alkoholkonsum					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,46	0,22; 0,70	$2,04 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag					
Geschlecht	sex	-0,55	-0,83; -0,27	$1,31 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

Tabelle A.22: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1–2 und 4–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1626,60

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		2,32	-0,90; 5,55	0,16	
Alter	age^{-1}	21,83	14,42; 29,24	$7,90 \cdot 10^{-9}$	$< 10^{-15}$
	$age^{-0,5}$	-32,35	-41,76; -22,94	$1,57 \cdot 10^{-11}$	
Körpermasseindex	$bmi^{0,5}$	5,12	4,09; 6,15	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,94	1,01; 2,88	$4,57 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,91 \cdot 10^{-3}$	$-2,70 \cdot 10^{-3}$; $-1,12 \cdot 10^{-3}$	$2,28 \cdot 10^{-6}$	$2,54 \cdot 10^{-3}$
	$drkt^2$	2,44	0,30; 4,58	0,03	
Alkoholkonsum	$drkt^3$	-1,48	-2,93; -0,02	0,05	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,48	0,25; 0,71	$4,86 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag					
Geschlecht	sex	-0,49	-0,78; -0,19	$1,25 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(b) QIC_u=1627,31

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		11,47	5,24; 17,69	$3,07 \cdot 10^{-4}$	
Alter	$age^{-0,5}$	-19,95	-25,32; -14,58	$3,33 \cdot 10^{-13}$	$< 10^{-15}$
	$age^{-0,5} \cdot \ln(age)$	-14,57	-19,65; -9,49	$1,93 \cdot 10^{-8}$	
Körpermasseindex	$bmi^{0,5}$	5,14	4,10; 6,17	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,92	0,99; 2,85	$5,19 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,92 \cdot 10^{-3}$	$-2,72 \cdot 10^{-3}$; $-1,13 \cdot 10^{-3}$	$1,99 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^{0,5}$	0,78	0,10; 1,47	0,03	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,48	0,25; 0,71	$5,12 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag					
Geschlecht	sex	-0,47	-0,78; -0,17	$2,00 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

(c) BQIC_u=1674,03

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-9,65	-11,43; -7,87	$< 10^{-15}$	
Alter	age^2	0,08	0,06; 0,10	$< 10^{-15}$	
Körpermasseindex	$bmi^{0,5}$	5,02	3,98; 6,06	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^3	1,97	0,98; 2,96	$1,02 \cdot 10^{-4}$	
Familiengröße	fs^3	$-1,89 \cdot 10^{-3}$	$-2,71 \cdot 10^{-3}$; $-1,07 \cdot 10^{-3}$	$6,01 \cdot 10^{-6}$	
Alkoholkonsum	$drkt^{0,5}$	0,77	0,09; 1,44	0,03	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,45	0,22; 0,68	$1,27 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration					
Zigaretten pro Tag					
Geschlecht	sex	-0,46	-0,76; -0,16	$2,69 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

Tabelle A.23: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 1 und 3–10 und QIC, QIC_u und BQIC_u als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) QIC=1558,22

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		9,00	4,65; 13,35	$4,97 \cdot 10^{-5}$	
Alter	age ⁻¹	18,53	10,52; 26,55	$5,88 \cdot 10^{-6}$	$1,75 \cdot 10^{-11}$
	age ^{-0,5}	-28,38	-38,50; -18,26	$3,84 \cdot 10^{-8}$	
Körpermasseindex	bmi	1,44	1,11; 1,76	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,62	1,55; 3,70	$1,76 \cdot 10^{-6}$	
Familiengröße	fs ³	$-1,80 \cdot 10^{-3}$	$-2,61 \cdot 10^{-3}$; $-9,85 \cdot 10^{-4}$	$1,46 \cdot 10^{-5}$	
Alkoholkonsum	drkt ²	2,45	0,32; 4,58	0,02	$2,06 \cdot 10^{-3}$
	drkt ³	-1,52	-2,96; -0,07	0,04	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,55	0,31; 0,79	$8,69 \cdot 10^{-6}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ^{-0,5}	-13,18	-23,91; -2,46	0,02	
Zigaretten pro Tag					
Geschlecht	sex	-0,41	-0,72; -0,09	0,01	
Bluthochdruckbehandlung					

(b) QIC_u=1558,46

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-2,00	-5,63; 1,64	0,28	
Alter	age ³	0,08	0,03; 0,13	$6,59 \cdot 10^{-4}$	$1,34 \cdot 10^{-5}$
	age ³ · ln(age)	-0,04	-0,07; -0,01	$4,51 \cdot 10^{-3}$	
Körpermasseindex	bmi	1,41	1,08; 1,74	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,59	1,52; 3,67	$2,23 \cdot 10^{-6}$	
Familiengröße	fs ³	$-1,79 \cdot 10^{-3}$	$-2,58 \cdot 10^{-3}$; $-1,01 \cdot 10^{-3}$	$8,10 \cdot 10^{-6}$	
Alkoholkonsum	drkt ²	2,38	0,26; 4,50	0,03	$2,62 \cdot 10^{-3}$
	drkt ³	-1,47	-2,91; -0,03	0,04	
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,54	0,30; 0,79	$1,43 \cdot 10^{-5}$	
Nüchtern-Serum-Glukose-Konzentration	glucw ^{-0,5}	-13,62	-24,28; -2,97	0,01	
Zigaretten pro Tag					
Geschlecht	sex	-0,41	-0,73; -0,90	0,01	
Bluthochdruckbehandlung					

(c) BQIC_u=1617,43

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-12,35	-16,28; -8,42	$7,25 \cdot 10^{-10}$	
Alter	age	0,60	0,46; 0,73	$< 10^{-15}$	
Körpermasseindex	bmi	1,40	1,07; 1,73	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl	2,93	1,88; 3,98	$4,32 \cdot 10^{-8}$	
Familiengröße	fs ³	$-1,81 \cdot 10^{-3}$	$-2,60 \cdot 10^{-3}$; $-1,02 \cdot 10^{-3}$	$7,16 \cdot 10^{-6}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	ln(tgw)	0,59	0,34; 0,84	$3,00 \cdot 10^{-6}$	
Nüchtern-Serum-Glukose-Konzentration	ln(glucw)	2,30	0,62; 3,98	$7,27 \cdot 10^{-3}$	
Zigaretten pro Tag					
Geschlecht	sex	-0,53	-0,82; -0,23	$4,46 \cdot 10^{-4}$	
Bluthochdruckbehandlung					

Tabelle A.24: Ergebnisse der Modelle nach der schrittweisen Prozedur unter Verwendung der Teildatensätze 2–10 und QIC, QIC_u und $BQIC_u$ als Kriterien der Güte der Anpassung und der Selektion. Gezeigt sind die Transformationen der Variablen und Informationen zu den zugehörigen Schätzern für die vollständigen Modelle. KI: Konfidenzintervall; p_k -Wert: kombinierter p -Wert.

(a) $QIC=1545,42$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		10,32	6,51; 14,14	$1,14 \cdot 10^{-7}$	
Alter	age^{-1}	25,31	16,80; 33,83	$5,57 \cdot 10^{-9}$	
	$age^{-0,5}$	-36,82	-47,56; -26,08	$1,80 \cdot 10^{-11}$	
Körpermasseindex	bmi	1,60	1,24; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,90	0,98; 2,83	$5,69 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,84 \cdot 10^{-3}$	$-2,66 \cdot 10^{-3}$; $-1,01 \cdot 10^{-3}$	$1,20 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^3$	1,07	0,15; 2,00	0,02	$1,18 \cdot 10^{-3}$
	$drkt^3 \cdot \ln(drkt)$	-2,22	-4,17; -0,27	0,03	
Körperhöhe	hgt^{-2}	-77,73	-132,50; -22,98	$5,40 \cdot 10^{-3}$	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,45	0,21; 0,69	$2,30 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-2}$	-94,01	-171,40; -16,60	0,02	
Zigaretten pro Tag					
Geschlecht					
Bluthochdruckbehandlung					

(b) $QIC_u=1545,95$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		10,32	6,51; 14,14	$1,14 \cdot 10^{-7}$	
Alter	age^{-1}	25,31	16,80; 33,83	$5,57 \cdot 10^{-9}$	$< 10^{-15}$
	$age^{-0,5}$	-36,82	-47,56; -26,08	$1,80 \cdot 10^{-11}$	
Körpermasseindex	bmi	1,60	1,24; 1,95	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	1,90	0,98; 2,83	$5,69 \cdot 10^{-5}$	
Familiengröße	fs^3	$-1,84 \cdot 10^{-3}$	$-2,66 \cdot 10^{-3}$; $-1,01 \cdot 10^{-3}$	$1,20 \cdot 10^{-5}$	
Alkoholkonsum	$drkt^3$	1,07	0,15; 2,00	0,02	$1,18 \cdot 10^{-3}$
	$drkt^3 \cdot \ln(drkt)$	-2,22	-4,17; -0,27	0,03	
Körperhöhe	hgt^{-2}	-77,73	-132,50; -22,98	$5,40 \cdot 10^{-3}$	
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,45	0,21; 0,69	$2,30 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-2}$	-94,01	-117,40; -16,60	0,02	
Zigaretten pro Tag					
Geschlecht					
Bluthochdruckbehandlung					

(c) $BQIC_u=1604,54$

Variable	Parameter	Schätzer	95%-KI	p -Wert	p_k -Wert
Regressionskonstante		-5,45	-6,91; -3,98	$3,03 \cdot 10^{-13}$	
Alter	age	0,64	0,50; 0,79	$< 10^{-15}$	
Körpermasseindex	bmi	1,47	1,11; 1,82	$< 10^{-15}$	
Serum-Totalcholesterin-Konzentration					
Nüchtern-Serum-HDL-Cholesterin-Konzentration	hdl^2	2,13	1,21; 3,05	$6,36 \cdot 10^{-6}$	
Familiengröße	fs^3	$-1,95 \cdot 10^{-3}$	$-2,70 \cdot 10^{-3}$; $-1,20 \cdot 10^{-3}$	$3,66 \cdot 10^{-7}$	
Alkoholkonsum					
Körperhöhe					
Nüchtern-Serum-Triglycerid-Konzentration	$\ln(tgw)$	0,46	0,21; 0,70	$2,57 \cdot 10^{-4}$	
Nüchtern-Serum-Glukose-Konzentration	$glucw^{-2}$	-105,30	-184,10; -26,59	$8,75 \cdot 10^{-3}$	
Zigaretten pro Tag					
Geschlecht	sex	-0,50	-0,80; -0,20	$1,02 \cdot 10^{-3}$	
Bluthochdruckbehandlung					

Tabelle A.25: Wahrscheinlichkeiten korrekt vorhergesagter Ergebnisse in der Zielvariablen in den vollständigen Modellen und den Modellen nach der schrittweisen Prozedur in der zehnfach Kreuzvalidierung unter Verwendung von QIC, QIC_u und $BQIC_u$ als Selektions- und Gütekriterien. SP: schrittweise Prozedur.

	QIC		QIC_u		$BQIC_u$	
	vor SP	nach SP	vor SP	nach SP	vor SP	nach SP
Minimum	0,6889	0,6889	0,6889	0,6858	0,6944	0,7037
2,5%-Quantil	0,6923	0,6914	0,6926	0,6865	0,6949	0,7045
1. Quartil	0,7084	0,7041	0,7063	0,7041	0,7209	0,7131
Median	0,7294	0,7238	0,7255	0,7261	0,7230	0,7215
Mittelwert	0,7344	0,7287	0,7296	0,7271	0,7315	0,7318
3. Quartil	0,7574	0,7554	0,7431	0,7472	0,7447	0,7538
97,5%-Quantil	0,7861	0,7701	0,7861	0,7760	0,7845	0,7753
Maximum	0,7920	0,7710	0,7920	0,7786	0,7920	0,7786

A.3. Abbildungen zum Vergleich der Modelle unter Verwendung der verschiedenen Gütekriterien

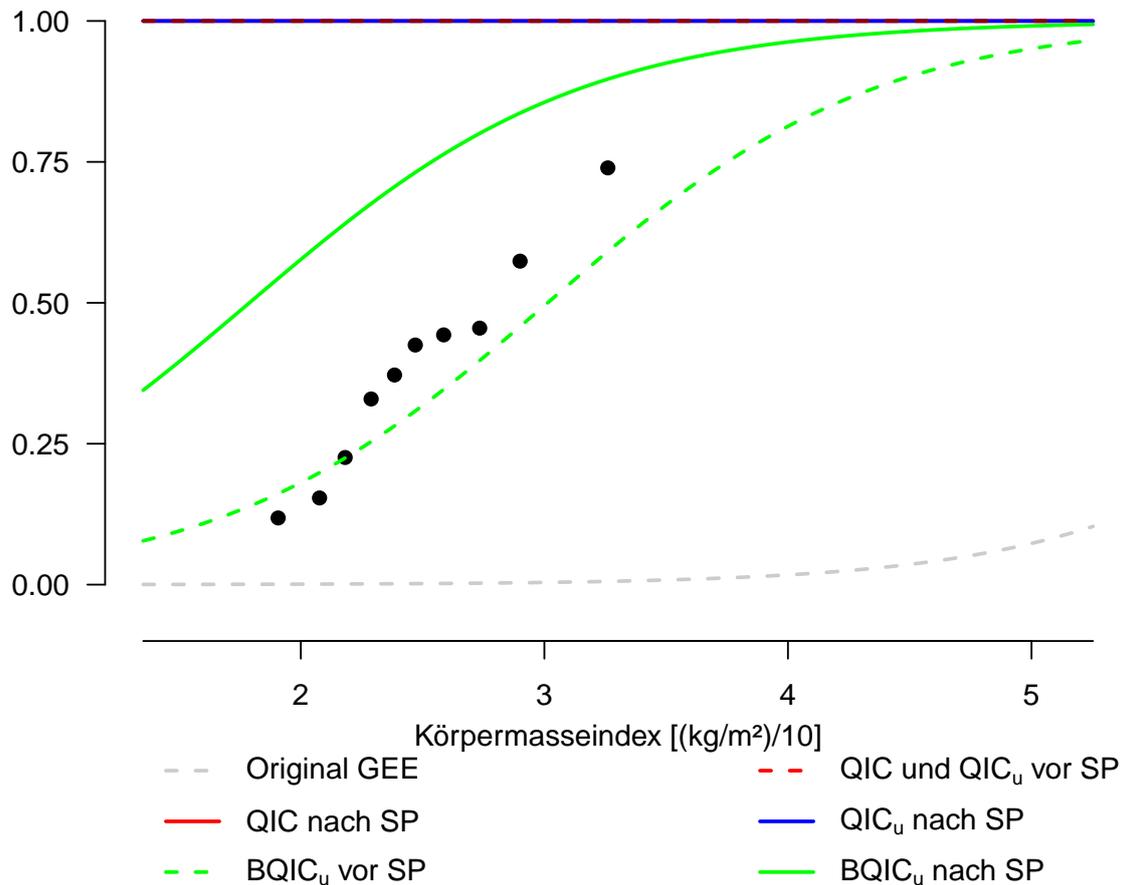


Abbildung A.1: Vergleich der Anpassungen an den Körpermasseindex unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün). SP: schrittweise Prozedur.

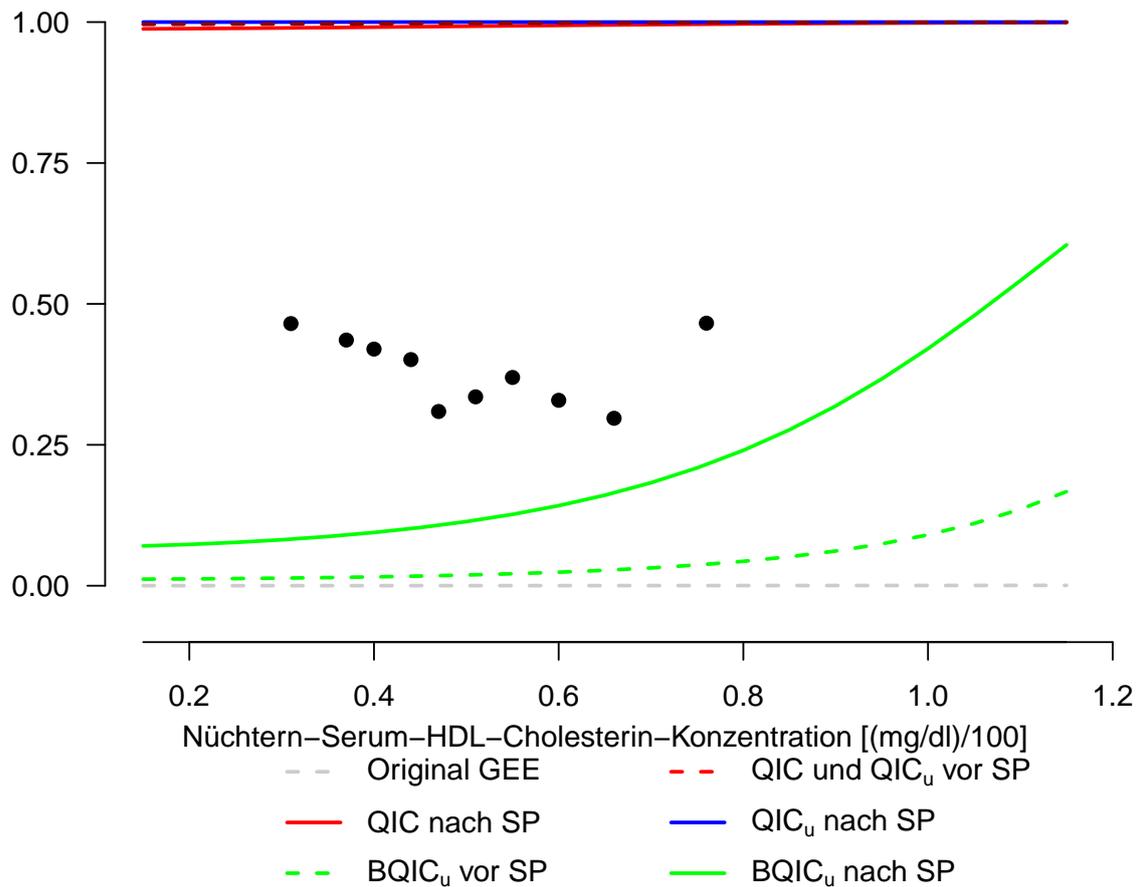


Abbildung A.2: Vergleich der Anpassungen an die Nüchtern-Serum-HDL-Cholesterin-Konzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün). SP: schrittweise Prozedur.

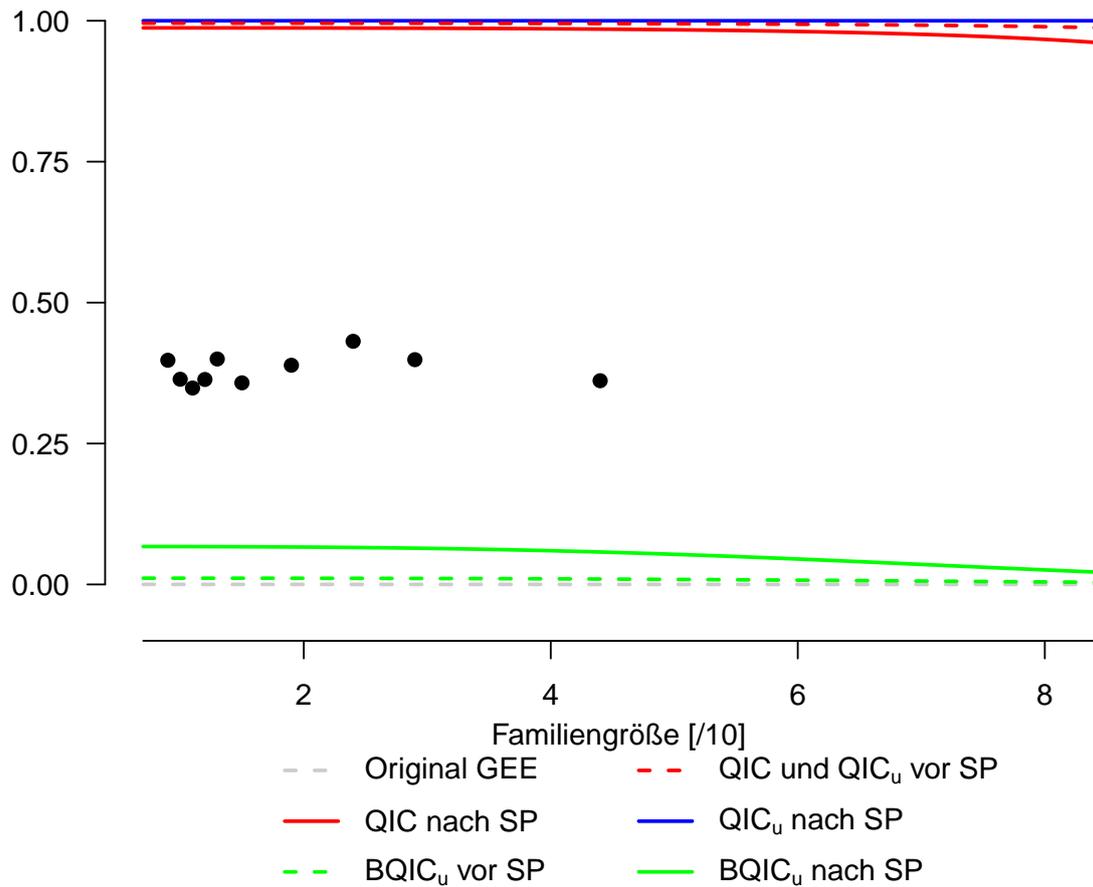


Abbildung A.3: Vergleich der Anpassungen an die Familiengröße unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün). SP: schrittweise Prozedur.

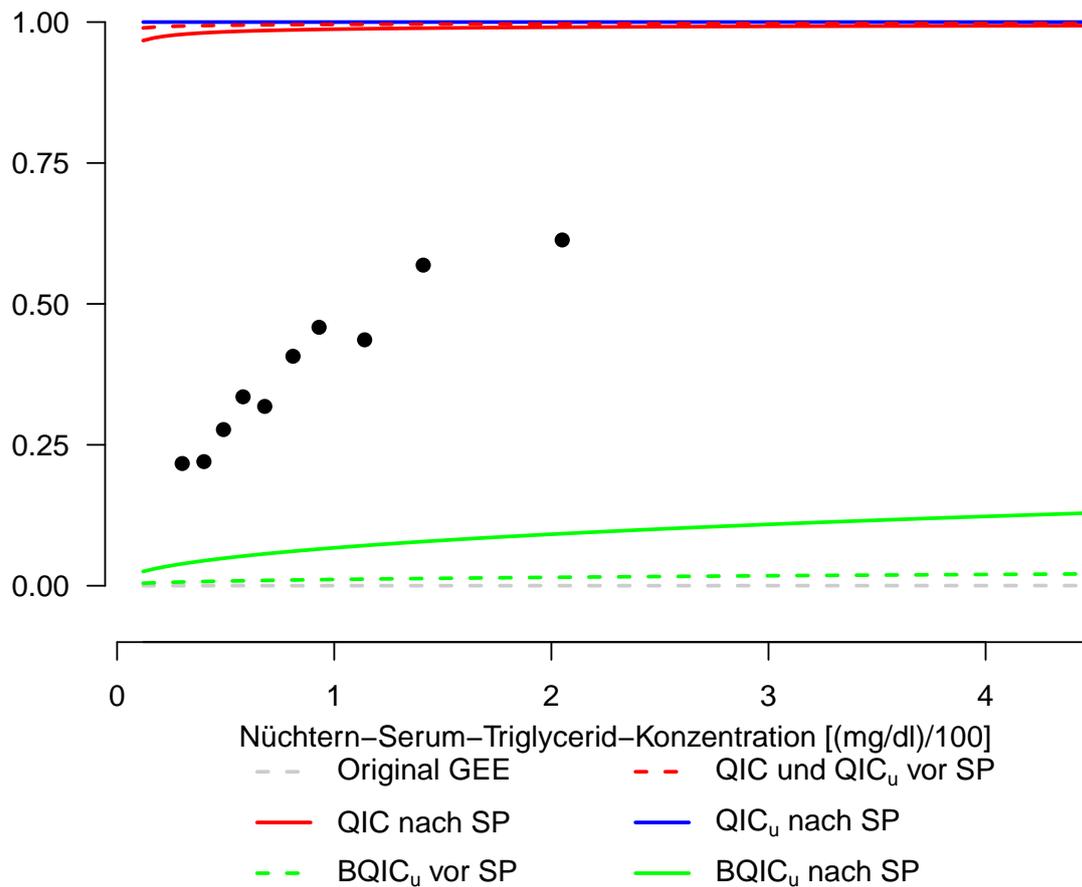


Abbildung A.4: Vergleich der Anpassungen an die Nüchtern-Serum-Triglycerid-Konzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün). SP: schrittweise Prozedur.

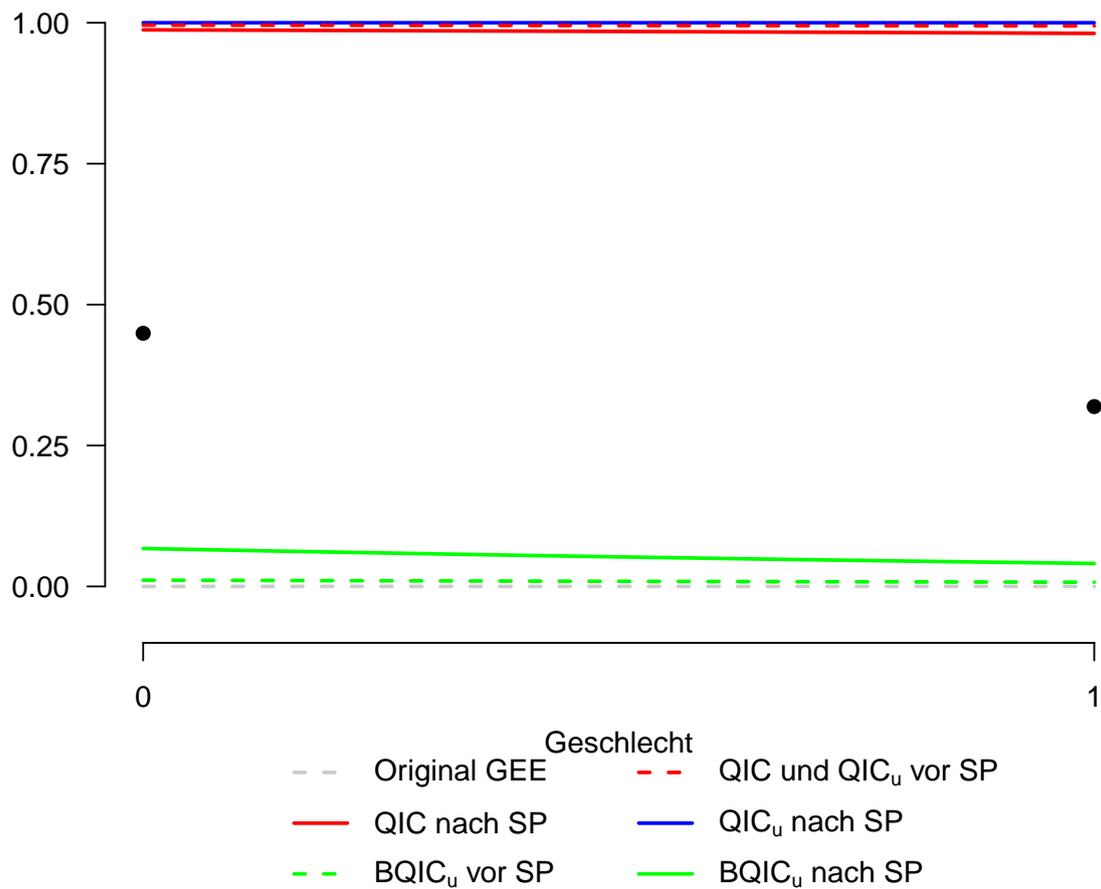


Abbildung A.5: Vergleich der Anpassungen an das Geschlecht unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und BQIC_u (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und BQIC_u (grün). SP: schrittweise Prozedur.

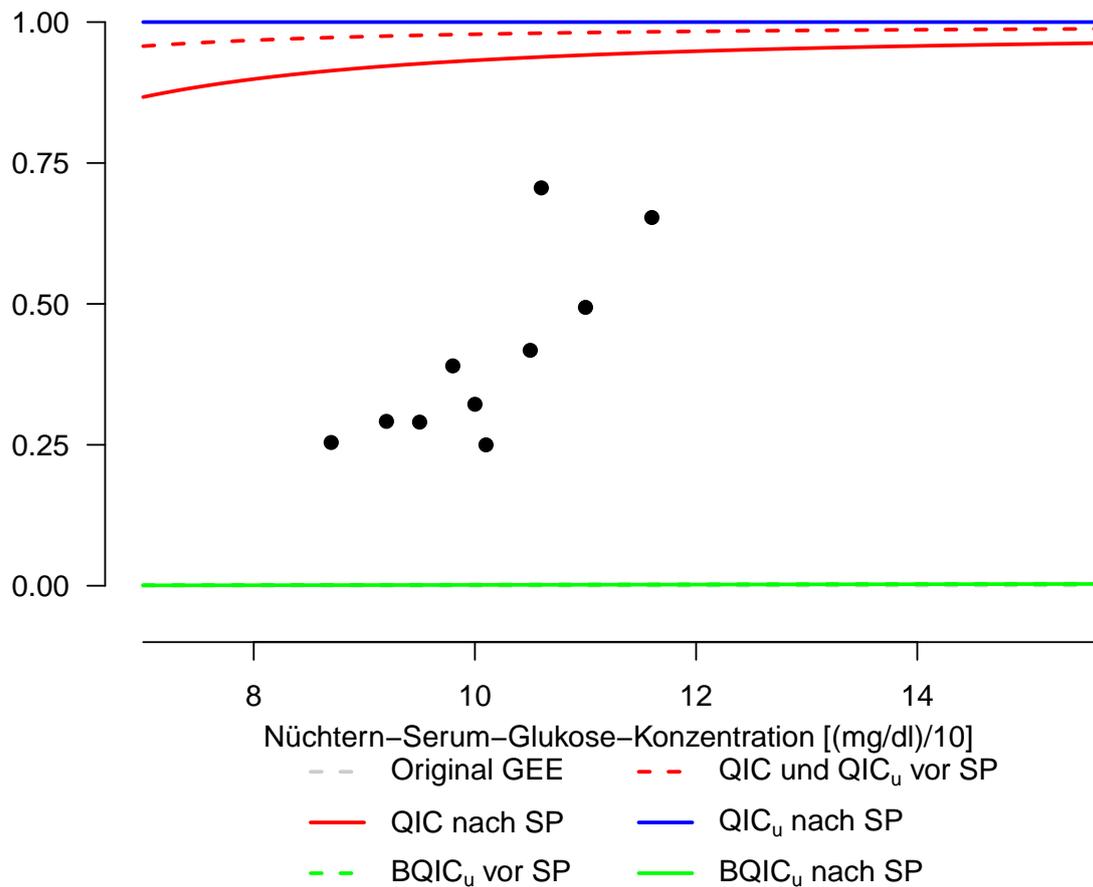


Abbildung A.6: Vergleich der Anpassungen an die Nüchtern-Serum-Glukose-Konzentration unter Verwendung der verschiedenen Selektions- und Anpassungskriterien vor und nach der schrittweisen Prozedur. Dargestellt sind das vollständige Modell, in dem alle Variablen linearen Einfluss haben (gestrichelt grau), die vollständigen Modelle unter Verwendung von QIC und QIC_u (gestrichelt rot) und $BQIC_u$ (gestrichelt grün), die Modelle nach den schrittweisen Prozeduren unter Verwendung des QIC (rot), QIC_u (blau) und $BQIC_u$ (grün). SP: schrittweise Prozedur.

Literaturverzeichnis

- Atkinson AC: Two graphical displays for outlying and influential observations in regression. *Biometrika* 68, 13–20 (1981)
- Baradat P, Adams W, Muller-Stack G: Population genetics and genetic conservation of forest trees. Backhuys Publishers, Amsterdam (1996)
- Belsley DA, Kuh E, Welsch RE: Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York (2004)
- Chaganty NR, Joe H: Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, 851–860 (2004)
- Cook RD, Weisberg S: Residuals and influence in regression. Chapman and Hall, New York (1982)
- Crowder M: On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* 82, 407–410 (1995)
- Cui J, De Klerk N, Abramson M, Del Monaco A, Benke G, Dennekamp M, Musk AW, Sim M: Fractional polynomials and model selection in generalized estimating equations analysis, with an application to a longitudinal epidemiologic study in Australia. *American Journal of Epidemiology* 169, 113–121 (2009)
- Cui J, Qian G: Selection of working correlation structure and best model in GEE analyses of longitudinal data. *Communications in Statistics - Simulation and Computation* 36, 987–996 (2007)
- Cupples LA, Yang Q, Demissie S, Copenhafer D, Levy D: Description of the Framing-

-
- ham Heart Study data for Genetic Analysis Workshop 13. *BMC Genetics* 4 Suppl 1, S2 (2003)
- Dahmen G, Ziegler A: Generalized estimating equations in controlled clinical trials: hypotheses testing. *Biometrical Journal* 46, 214–232 (2004)
- Dahmen G, Ziegler A: Independence estimating equations for controlled clinical trials with small sample sizes – interval estimation. *Methods of Information in Medicine* 45, 430–434 (2006)
- Emrich LJ, Piedmonte MR: On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* 41, 19–29 (1992)
- Evans S, Li L: A comparison of goodness of fit tests for the logistic GEE model. *Statistics in Medicine* 24, 1245–1261 (2005)
- Gourieroux C, Monfort A: *Statistics and econometric models*, Bd. 1. Cambridge University Press, Cambridge (1995)
- Hammill BG, Preisser JS: A SAS/IML software program for GEE and regression diagnostics. *Computational Statistics & Data Analysis* 51, 1197–1212 (2006)
- Hanley JA, Negassa A, Edwardes MD: GEE analysis of negatively correlated binary responses: a caution. *Statistics in Medicine* 19, 715–722 (2000)
- Hilbe JM, Hardin JW: *Generalized estimating equations*. Chapman & Hall (2002)
- Hin LY, Carey VJ, Wang YG: Criteria for working-correlation-structure selection in GEE. *The American Statistician* 61, 360–364 (2007)
- Hin LY, Wang YG: Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 28, 642–658 (2009)
- Jung K: Local influence in generalized estimating equations. *Scandinavian Journal of Statistics* 35, 286–294 (2008)
- Khan TH, Manzoor U: The relationship of family income, family size, age and cir-

- cumferences with blood pressure in the female students of the Bahauddin Zakariya University, Multan, Pakistan. *Anthropologischer Anzeiger; Bericht ueber die biologisch-anthropologische Literatur* 60, 293–298 (2002)
- Liang K, Zeger SL: Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22 (1986)
- Liu A, Boyett JM, Xiong X: Sample size calculation for planning group sequential longitudinal trials. *Statistics in Medicine* 19, 205–220 (2000)
- Mancl LA, Leroux BG: Efficiency of regression estimates for clustered data. *Biometrics* 52, 500–511 (1996)
- McDonald BW: Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 55, 391–397 (1993)
- Nelder JA, Wedderburn RWM: Generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 135, 370–384 (1972)
- Oh S, Carriere K, Park T: Model diagnostic plots for repeated measures data using the generalized estimating equations approach. *Computational Statistics & Data Analysis* 53, 222–232 (2008)
- Pan W: Akaike's information criterion in generalized estimating equations. *Biometrics* 57, 120–125 (2001a)
- Pan W: Model selection in estimating equations. *Biometrics* 57, 529–534 (2001b)
- Pan W, Connett JE: Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica* 12, 475–490 (2002)
- Pan W, Louis TA, Connett JE: A note on marginal linear regression with correlated response data. *The American Statistician* 54, 191–195 (2000)
- Pepe MS, Anderson GL: A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 23, 939 (1994)

- Preisser JS, Garcia DI: Alternative computational formulae for generalized linear model diagnostics: identifying influential observations with SAS software. *Computational Statistics & Data Analysis* 48, 755–764 (2005)
- Preisser JS, Perin J: Deletion diagnostics for marginal mean and correlation model parameters in estimating equations. *Statistics and Computing* 17, 381–393 (2007)
- Preisser JS, Qaqish BF: Deletion diagnostics for generalised estimating equations. *Biometrika* 83, 551–562 (1996)
- Preisser JS, Qaqish BF, Perin J: A note on deletion diagnostics for estimating equations. *Biometrika* 95, 509–513 (2008)
- Prentice RL: Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44, 1033–1048 (1988)
- Royston P, Altman DG: Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43, 429–467 (1994)
- Royston P, Sauerbrei W: *Multivariable model-building*. John Wiley & Sons, New York (2008)
- Sauerbrei W, Royston P: Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162, 71–94 (1999)
- Sauerbrei W, Royston P, Binder H: Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 26, 5512–5528 (2007)
- Schwarz G: Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464 (1978)
- Sherman M, Cessie SI: A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation* 26, 901–925 (1997)

- Shults J, Chaganty NR: Analysis of serially correlated data using quasi-least squares. *Biometrics* 54, 1622–1630 (1998)
- Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T: A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine* 28, 2338–2355 (2009)
- Spiess M, Hamerle A: On the properties of GEE estimators in the presence of invariant covariates. *Biometrical Journal* 38, 931–940 (1996)
- Thomas W, Cook RD: Assessing influence on regression coefficients in generalized linear models. *Biometrika* 76, 741–749 (1989)
- Thompson WK, Xie M, White HR: Transformations of covariates for longitudinal data. *Biostatistics* 4, 353–364 (2003)
- Vanscheidt W, Rabe E, Naser-Hijazi B, Ramelet AA, Partsch H, Diehm C, Schultze-Ehrenburg U, Spengel F, Wirsching M, Götz V, Schnitker J, Henneicke-von Zepelin HH: The efficacy and safety of a coumarin-/troxerutin-combination (SB-LOT) in patients with chronic venous insufficiency: a double blind placebo-controlled randomised study. *VASA. Journal for Vascular Diseases* 31, 185–190 (2002)
- Venezuela MK, Botter DA, Sandoval MC: Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation* 77, 879–888 (2007)
- Venezuela MK, Sandoval MC, Botter DA: Local influence in estimating equations. *Computational Statistics & Data Analysis* 55, 1867–1883 (2011)
- Vens M, Ziegler A: Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: a case study. *Computational Statistics & Data Analysis* 56, 1232–1242 (2012)
- Wang Y, Carey V: Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika* 90, 29–41 (2003)

- Wei WH, Fung WK: The mean-shift outlier model in general weighted regression and its applications. *Computational Statistics & Data Analysis* 30, 429–441 (1999)
- Zeger SL, Liang K: Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121–130 (1986)
- Ziegler A: Generalized estimating equations. In: Ahrens W, Pigeot I (Hrsg.) Handbook of epidemiology, im Druck: Springer, Heidelberg, 2. Aufl. (2013)
- Ziegler A, Arminger G: Parameter estimation and regression diagnostics using generalized estimating equations. In: *SoftStat'95 Advances in Statistical Software*, 5, 229–237: Lucius & Lucius, Heidelberg (1996)
- Ziegler A, Arminger G, Bachleitner R: Pseudo Maximum Likelihood Schätzung und Regressionsdiagnostik für Zähldaten: Suizid durch Tourismus? *GUTE FRAGE* 79, 170 – 195 (1995)
- Ziegler A, Kastner C, Blettner M: The generalised estimating equations: an annotated bibliography. *Biometrical Journal* 40, 115–139 (1998)
- Ziegler A, Kastner C, Chang-Claude J: Analysis of pregnancy and other factors on detection of human papilloma virus (HPV) infection using weighted estimating equations for follow-up data. *Statistics in Medicine* 22, 2217–2233 (2003)
- Ziegler A, Kastner C, Grömping U, Blettner M: Generalized Estimating Equations: Herleitung und Anwendung. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 27, 67–91 (1996)
- Ziegler A, Vens M: Generalized estimating equations: notes on the choice of the working correlation matrix. *Methods of Information in Medicine* 49, 421–425; Diskussion 426–432 (2010)

Danksagung

Mein Dank richtet sich besonders an Herrn Univ.-Prof. Dr. rer. nat. Andreas Ziegler für die Bereitstellung des Themas und die umfassende wissenschaftliche Betreuung.

Herrn Dr. Hans-Heinrich Henneicke-von Zepelin danke ich für die Überlassung der SB-LOT-Daten.

Ich bedanke mich weiterhin dafür, dass ich die Daten der Framingham Heart Study in einer unabhängigen Analyse (Studiennummer 1072) benutzen durfte.

Ein großes Dankeschön geht an alle Mitarbeiter_innen, auch ehemalige, des Instituts für Medizinische Biometrie und Statistik für die gute Zusammenarbeit und die zahlreichen Hilfestellungen bei allen Problemen. Hier möchte ich mich besonders für die Unterstützung von Jördis Stolpmann bei der Programmierung meiner Algorithmen, bei Frank Sandig für die Hilfe mit Abbildungen aller Art, bei Univ.-Prof. Dr. rer. hum. biol. Inke R. König und Dr. rer. hum. biol. Jochen Kruppa für die Erklärungen zur Kreuzvalidierung und deren Auswertung, bei Dipl.-Math. Christina Loley und Dipl.-Math. Janja Nahrstaedt für die Begutachtung zahlreicher Halbnormalabbildungen, bei Dr. rer. hum. biol. Arne Schillert für die Unterstützung bei technischen Problemen, bei B.Sc. Nicole Hessler, bei Dipl.-Math. Theresa Holste für die vielen Mittagessen und Diskussionen, bei Gabriele Schatton, bei Dr. rer. hum. biol. Claudia Hemmelmann für zahlreiche Diskussionen, Übersetzungshilfen und für die tolle Zeit im gemeinsamen Büro bedanken.

Dr. rer. nat. Silke Szymczak danke ich für die zahlreichen Diskussionen, für das Korrekturlesen und für alles, was sie sonst noch für mich getan hat.

Danke auch an Norma Möller, Leonie Kerkhoff und Andrea Lüth für die Korrekturen.

Vielen ganz besonders herzlichen Dank möchte ich meinen Eltern Brigitte und Willi, meinen Geschwistern Carsten und Andrea und der erweiterten Familie sagen.

Habt mich still und geduldig verwöhnt und beschenkt.

Das war völlig selbstlos,
als wäre es normal.

Geboren, erzogen, beschützt und geliebt.

Ja, Ihr ward immer da.

Habt mir Wege geebnet und Schatten verscheucht,
euch mit mir gefreut und deshalb ist das hier für euch.

Pur

Lebenslauf

Persönliche Daten

Name Maren Andrea Vens
Alter 31 Jahre

Studium

04/2008 – heute Promotionsstudium
Universität zu Lübeck
10/2005 – 01/2008 Computational Life Science
Abschluss: Master of Science (M.Sc.)
Universität zu Lübeck
10/2002 – 09/2005 Computational Life Science
Abschluss: Bachelor of Science (B.Sc.)
Universität zu Lübeck

Berufliche Tätigkeiten

01/2008 – heute Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Wissenschaftliche Mitarbeiterin
04/2007 – 11/2007 Institut für Biochemie und Institut für Theoretische Informatik
ETH Zürich
Wissenschaftliche Assistentin
10/2006 – 02/2007 Institut für Mathematik
Universität zu Lübeck
Studentische Hilfskraft
09/2005 – 07/2006 Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Studentische Hilfskraft
10/2004 – 02/2005 Institut für Mathematik
Universität zu Lübeck
Studentische Hilfskraft

Eigene Publikationen

Originalarbeiten

Hemmelmann C, Brose S, **Vens M**, Hebebrand J, Ziegler A: Perzentilen des Body-Mass-Index auch für 18- bis 80-Jährige? Daten der Nationalen Verzehrsstudie II. *DMW - Deutsche Medizinische Wochenschrift* 135, 848–852 (2010)

Schillert A, Schwarz DF, **Vens M**, Szymczak S, König IR, Ziegler A: ACPA: automated cluster plot analysis of genotype data. *BMC Proceedings* 3, S58 (2009)

Vens M, Schillert A, König IR, Ziegler A: Look who is calling: a comparison of genotype calling algorithms. *BMC Proceedings* 3, S59 (2009)

Vens M, Ziegler A: Estimating disequilibrium coefficients. *Methods in Molecular Biology* 850, 103–117 (2012a)

Vens M, Ziegler A: Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: a case study. *Computational Statistics & Data Analysis* 56, 1232–1242 (2012b)

Timmann C, Thye T, **Vens M**, Evans J, May J, Ehmen C, Sievertsen J, Muntau B, Ruge G, Loag W, Ansong D, Antwi S, Asafo-Adjei E, Nguah SB, Kwakye KO, Akoto AO, Sylverken J, Brendel M, Schuldt K, Loley C, Franke A, Meyer CG, Agbenyega T, Ziegler A, Horstmann RD: Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* 489, 443–446 (2012)

Timmann C, van der Kamp E, Kleensang A, König IR, Thye T, Bättner DW, Hamelmann C, Marfo Y, **Vens M**, Brattig N, Ziegler A, Horstmann RD: Human genetic resistance to *Onchocerca volvulus*: evidence for linkage to chromosome 2p from an autosome-wide scan. *The Journal of Infectious Diseases* 198, 427–433 (2008)

Ziegler A, **Vens M**: Generalized estimating equations: notes on the choice of the working correlation matrix. *Methods of Information in Medicine* 49, 421–425; Diskussion 426–432 (2010)

Buchbeiträge

Vens M, Wellek S, Ziegler A: Genotyp-basierte Maße für Kopplungsungleichgewicht. In: Foraita R, Ziegler A, Hemmelmann C (Hrsg.) Biometrische Aspekte der Genomanalyse IV. Schwerpunkt: Epigenetik, 102–109: Shaker, Aachen (2009)

Vens M, Ziegler A: Estimating disequilibrium coefficients. In: Elston RC, Satagopan JM, Sun S (Hrsg.) Statistical human genetics: methods and protocols, 103–117: Humana Press (2012)

Vorträge

Hemmelmann C, Brose S, **Vens M**, Ziegler A: BMI-Perzentilkurven deutscher Erwachsener im internationalen Vergleich (2010). 55. GMDS Jahrestagung, Mannheim; Hrsg. von P Schmücker, K-H Ellsäcker und S Hayna: *55. GMDS Jahrestagung: Effiziente und wirtschaftliche Gesundheitsversorgung von heute und morgen - nur mit Medizinischer Dokumentation, Medizinischer Informatik, Medizinischer Biometrie und Epidemiologie - Abstractband*, ISBN: 978-3-932971-11-2

König IR, Hemmelmann C, **Vens M**, Ziegler A: Das SMART-Board im Unterricht „Medizinische Biometrie für Mediziner“: Ein Erfahrungsbericht (2010). 55. GMDS Jahrestagung, Mannheim; Hrsg. von P Schmücker, K-H Ellsäcker und S Hayna: *55. GMDS Jahrestagung: Effiziente und wirtschaftliche Gesundheitsversorgung von heute und morgen - nur mit Medizinischer Dokumentation, Medizinischer Informatik, Medizinischer Biometrie und Epidemiologie - Abstractband*, ISBN: 978-3-932971-11-2

Vens M, Stolpmann J, Hemmelmann C, Ziegler A: Modeling of continuous covariates in the mean structure of generalized estimating equations (2011). CEN 2011: 2nd conference of the Central European Network, 27. ROeS Seminar, 57. Biometrisches Kolloquium, Zürich, Schweiz; Hrsg. von N Neumann, H U Burger, K Ickstadt, S Mejza und D Heinzmann: *CEN 2011 Bridging Theory and Application - Abstract booklet* - , ISBN: 978-3-033-03083-1

Vens M, Stolpmann J, Hemmelmann C, Ziegler A: Multivariable fraktionelle Polynome für korrelierte abhängige Variablen unter Verwendung von verallgemeinerten Schätzgleichungen (2012). 58. Biometrisches Kolloquium, Berlin, Deutschland

Vens M, Veligodskiy A, Sbalzarini IF, Kroschewski R: A model to predict maturity and morphological changes during epithelial cyst development (2007). *All-SystemsX.ch-Day*, September 2007, Lausanne, Schweiz, ausgewählt aus den studentischen Beiträgen, zusätzlich zum Poster

Ziegler A, Schillert A, **Vens M**, Zeller T, Blankenberg S: On the quality of genotype calling algorithms (2010). 2. *Joint Statistical Meeting Deutsche Arbeitsgemeinschaft Statistik „Statistics under one umbrella“* und 56. *Biometrisches Kolloquium*, Dortmund

Poster

Vens M, Stolpmann J, Hemmelmann C, Ziegler A: Iss damit Du groß und stark wirst!? Grenzen eines statistischen Modells und seine Verbesserungen (2013). *Uni im Dialog - Lübecker Doktorandentage*, Juni 2013, Lübeck, Deutschland

Vens M, Veligodskiy A, Sbalzarini IF, Kroschewski R: A model to predict maturity and morphological changes during epithelial cyst development (2007). *All-SystemsX.ch-Day*, September 2007, Lausanne, Schweiz