



Schätzung der Intra-Klumpen-Korrelation eines dichotomen Merkmals - mit Anwendungen

Dissertation

zur

Erlangung des akademischen Grades

doctor rerum politicum (Dr. rer. pol.)

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Andreas Dartsch

aus Rostock

Rostock, Mai 2016

Gutachter:

1. Gutachter:

Prof. Dr. Rafael Weißbach,
Institut für Volkswirtschaftslehre, Universität Rostock

2. Gutachter:

Prof. Dr. Kathrin Duman,
Institut für Volkswirtschaftslehre, Universität Rostock

Datum der Einreichung: 25. Mai 2016

Datum der Verteidigung: 15. November 2016

Inhaltsverzeichnis

Abbildungsverzeichnis	V
Tabellenverzeichnis	VI
Bezeichnungen	VII
1 Einleitung	1
2 Modelle	7
2.1 Modell mit pauschaler Klumpenkorrelation	7
2.2 Beta-Binomial-Modell	16
3 Methoden	20
3.1 Auf dem CCM basierende Schätzer	21
3.1.1 Maximum-Likelihood-Schätzer	21
3.1.2 Quasi-Likelihood-Schätzer	21
3.1.3 Extended-Quasi-Likelihood-Schätzer	23
3.1.4 Pseudo-Likelihood-Schätzer	24
3.2 Auf dem Beta-Binomial-Modell basierende Schätzer	25
3.2.1 Direkter Beta-Binomial-Modell-Schätzer	26
3.2.2 Bayes-Schätzer	27
3.3 Modellunabhängige Schätzer	29
3.3.1 Kappa-Schätzer	30
3.3.2 Weighted-Empirical-Pairwise-Schätzer	32
3.3.3 Pairwise-Equal-Weights-Schätzer	32
3.3.4 ANOVA-Schätzer	33
3.3.5 Moment-Method-Based-Schätzer	33
3.4 Konfidenzintervall für die Intra-Klumpen-Korrelation	35
4 Simulationen	37
5 Anwendungen	39
5.1 Datenbeschreibung	39
5.1.1 Zahngesundheitsstudie	39
5.1.2 Insolvenzen in Mecklenburg-Vorpommern 2006	41

Inhaltsverzeichnis

5.2	ICC-Schätzung für die Datensätze ZAHN und INSOL	43
5.3	Stichprobenplanung (ZAHN)	48
5.4	Das Alter als Kovariable (ZAHN)	49
5.5	Konfidenzintervall für die Intra-Klumpen-Korrelation	50
6	Ausblick	53
Anhang		VIII
A.1	Daten	VIII
A.1.1	Zahngesundheitsstudie	VIII
A.1.2	Insolvenzen in Mecklenburg-Vorpommern 2006	X
A.2	Herleitungen	X
A.2.1	Momente einer CCM-verteilten Zufallsvariable	X
A.2.2	Zerlegung der Varianz nach der 1. Ordnung	XII
A.2.3	Momente der A-posteriori-Verteilung von Π	XIII
A.3	SAS Quellcode	XIII
A.3.1	Simulationen	XIII
A.3.2	Konfidenzintervall für die Intra-Klumpen-Korrelation	XLVI
Literaturverzeichnis		L
Eidesstattliche Versicherung		LV

Abbildungsverzeichnis

5.1	Daten mit Klumpen-Design im Zeitverlauf	42
5.2	Log-Quasi-Likelihood-Funktion (Datensatz ZAHN)	43
5.3	Extended-Quasi-Likelihood-Funktion (Datensatz ZAHN)	44
5.4	Extended-Quasi-Likelihood-Funktion (Datensatz INSOL)	45
5.5	Pseudo-Likelihood-Funktion (Datensatz INSOL)	45

Tabellenverzeichnis

2.1	VIF in Abhängigkeit von ICC und Klumpengröße	10
3.1	Schätzmethoden nach zu Grunde liegendem Modell	20
4.1	Simulierter Schätz-Mittelwert	38
5.1	Karies-Prävalenz nach Alter	40
5.2	Schätzwerte	47
5.3	Zahngesundheitsdaten (ZAHN) nach Alter	49
5.4	Konfidenzintervalle für ρ	51
A.1	Datensatz ZAHN	VIII
A.2	Datensatz INSOL	X

Bezeichnungen

μ	Erwartungswert (Parameter)
π	Erfolgswahrscheinlichkeit
ρ	Intra-Klumpen-Korrelationskoeffizient
ρ_0	wahrer Wert Intra-Klumpen-Korrelationskoeffizient
<i>CRS</i>	Cluster Randomized Sample
<i>def f</i>	Design Effect
<i>E</i>	Erwartungswert (Operator)
<i>e</i>	Präzision
<i>ICC</i>	Intraclass Correlation Coefficient (= ρ)
<i>k</i>	Anzahl der Klumpen
<i>Kov</i>	Kovarianz
<i>Kor</i>	Korrelation
<i>L</i>	Likelihood
<i>l</i>	Log-Likelihood
<i>MQA</i>	Mittlere Abweichungsquadratsumme zwischen den Klumpen
<i>MQR</i>	Mittlere Abweichungsquadratsumme innerhalb der Klumpen
<i>N</i>	Stichprobengröße
n_i	Anzahl der Elemente in Klumpen i
n_{SRS}	Fallzahl bei einfacher Zufallsstichprobe
n_{CRS}	Fallzahl bei Klumpenstichprobe
<i>Pr</i>	Wahrscheinlichkeitsmaß
<i>SRS</i>	Simple Random Sample
<i>SQA</i>	Summe der quadratischen Abweichungen zwischen den Klumpen
<i>SQR</i>	Summe der quadratischen Abweichungen innerhalb der Klumpen
$u_{\alpha/2}$	$(1 - \alpha/2)$ - Quantil der Standardnormalverteilung
<i>Var</i>	Varianz
<i>VIF</i>	Varianzinflationsfaktor
<i>VIF*</i>	Varianzinflationsfaktor dominiert durch gewichtete durchschnittliche Klumpengröße
X_{ij}	Binäre Zielgröße, Wert von Element j , $j = 1, \dots, n_i$ in Klumpen i
Y_i	Gesamtanzahl an Erfolgen in Klumpen i

1 Einleitung

Abhängigkeiten zwischen statistischen Einheiten sind ein wesentliches Hindernis bei der Datenanalyse. Ziel dieser Arbeit ist die Modellierung und Schätzung von Querschnittsabhängigkeiten in Klumpenstichproben um einen Beitrag zu einem besseren Verständnis von Datenstrukturen zu liefern.

Klumpenstichproben können bei der statistischen Analyse von Datenbeständen auftreten. Das reicht von inferenziellen oder explorativen Datenanalysen in der Umfrageforschung über Kreditrisiken bis hin zu klinischen Studien. Oftmals werden dort Methoden mit einfachen Stichproben an Stelle von Klumpenstichproben verwendet, was nicht immer den Kriterien einer profunden und zuverlässigen Datenanalyse genügt. Hier liegt Potential zur Vermeidung handwerklicher Fehler.

Daten liegen in Form einer Klumpenstichprobe vor, wenn die jeweiligen Merkmalsträger nicht individuell aus der Grundgesamtheit gezogen werden, sondern in Gruppen (auch Klumpen genannt). Die Zusammensetzung der Gruppen wird durch ein oder mehrere ausgewählte Merkmale bestimmt, die für alle Merkmalsträger einer Gruppe übereinstimmen (z. B. Schule, Branche, Bundesland). Innerhalb der Gruppen werden alle Merkmalsträger (auch: Elemente der Gruppe) vollständig erfasst, so dass es zu unterschiedlich großen – unbalancierten – bzw. gleich großen – balancierten – Klumpen kommen kann. Dieses Design führt erstens dazu, dass die Verteilung der Daten in einem Klumpen nicht mehr unbedingt der Verteilung der Grundgesamtheit entspricht. Zweitens ist die Verteilung der Klumpen-Teilgesamtheit im Allgemeinen nicht mehr zufällig. Beides führt zu einer Veränderung der nominellen Varianz der beobachteten Zielgröße (Kish (1965, Kapitel 5.4)).

Wenn die empirische Varianz die theoretische Varianz der angenommenen Verteilung überschreitet, liegt Überdispersion vor (Cox und Snell (1989, Seite 107)). Weitere Bezeichnungen für diese Besonderheit sind Varianzinflation oder auch Varianzaufblähung. Kish (1965, Kapitel 5) prägt schließlich den Begriff *design effect*, der sich in der englischsprachigen Literatur durchgesetzt hat.

Der Varianzinflationsfaktor¹ (VIF) (auch: Varianzaufblähungsfaktor) ist der Faktor, um den die Varianz unterschätzt wird (Ganninger *et al.* (2007)). Er gibt den Multiplikator an,

¹In Abgrenzung dazu bezeichnet der Varianzinflationsfaktor in der Regressionsanalyse ein Maß, welches auf Multikollinearität hinweist.

„mit dem bei gleichem Stichprobenumfang die Varianz bei einfacher Zufallsauswahl versehen werden muss, um die Varianz durch das Klumpenverfahren zu erhalten“ (Hartung *et al.* (2009, Seite 290)). Das bedeutet, dass beide Verfahren die gleiche Präzision liefern, wenn innerhalb der Klumpen keine Abhängigkeiten bestehen. Kish (1965, Kapitel 5) definiert den VIF folglich als Verhältnis der Varianz des Schätzers im Klumpen-Design (CRS) zur einfachen Stichprobe (SRS)

$$VIF = \frac{Var_{CRS}(\hat{\theta})}{Var_{SRS}(\hat{\theta})}.$$

Varianzinflation verschlechtert die Qualität von Schätzern weitergehender Betrachtungen (Fallzahlplanung, etc.) (Weißbach *et al.* (2015)). So geht z.B. Präzision in der Punktschätzung durch Varianzinflation verloren. Das bedeutet, dass Standardfehler zu klein sind, wenn die Varianzinflation ignoriert wird (Cox und Snell (1989, Seite 110)).

Für Klumpenstichproben gilt die Annahme der statistischen Unabhängigkeit der Elemente zwischen den Klumpen sowie der Abhängigkeit innerhalb der Klumpen. Die Abhängigkeit wird durch die Intra-Klumpen-Korrelation (auch Intra-Klassen-Korrelation oder *intraclass correlation*) beschrieben. Das zugehörige Maß ist der Intra-Klumpen-Korrelationskoeffizient (engl. *intraclass correlation coefficient* (ICC)), welcher analog zu anderen bekannten Zusammenhangsmaßen mathematisch mit ρ bezeichnet wird.

Der ICC gibt Auskunft über den Grad der Homogenität der Elemente eines Klumpens bezüglich des betrachteten Merkmals (Ganninger *et al.* (2007)). Hohe Werte implizieren einen hohen Grad an Einheitlichkeit innerhalb eines Klumpens sowie einen hohen Grad an Verschiedenheit zwischen Klumpen. Dabei bietet der ICC bei dichotomen Daten (Ereignis tritt ein oder tritt nicht ein) nur ein geringes Maß an Intuition für das Ausmaß, denn sein Wert ist stark von der Wahrscheinlichkeit des Eintretens des Ereignisses abhängig (Gulliford *et al.* (2005)).

Der Zusammenhang zwischen VIF und ICC wird im Allgemeinen durch die Formel

$$VIF = 1 + (\bar{b} - 1)\rho$$

angegeben. Dabei ist \bar{b} entweder die Klumpengröße oder die (gewichtete) durchschnittliche Klumpengröße (VIF für die gesamte Stichprobe mit Klumpen-Design). In der Praxis ist ρ unbekannt und muss geschätzt werden.

Querschnittsabhängigkeiten werden erst seit den 1990er Jahren eine größere Bedeutung bei der statistischen Analyse beigemessen. Seitdem findet sich eine größere Anzahl theoretischer und empirischer Arbeiten zu dem Thema, während es davor üblich war, von der Unabhängigkeit der Beobachtungen auszugehen (Ullah und Giles, David E. A. (2011, Kapitel 4.3)). Dabei gibt es zwei Gründe, die Unabhängigkeitsannahme zu hinterfragen. So kann die Ursache für Abhängigkeit entweder in einer zusätzlichen Quelle zufälliger Variabilität liegen oder in einer positiven Korrelation zwischen Elementen innerhalb eines

Klumpens. Beide Umstände können durch regionale oder anders geartete Stratifikation jeweils exklusiv auf die Elemente *eines* Klumpens wirken.

Bei dichotomen Merkmalen findet der ICC häufig Anwendung in klinischen Studien zur Kontrolle des Erfolges unterschiedlicher Versuchsgruppen (z. B. Dixon *et al.* (2002)). Aber auch in Finanzwirtschaft und Demografie sowie unlängst in Feldern der Systembiologie, Genetik und Physik wächst die Bedeutung des Konzeptes, welches dann häufig zur Planung optimaler Fallzahlen Einsatz findet.

Pagel *et al.* (2011) studieren bspw. den ICC im Rahmen von Fallzahlplanungen für die Erfolgsbewertung perinataler Medizin auf kommunaler Ebene. In Familienstudien wird er genutzt, um die Ähnlichkeit der Familienmitglieder hinsichtlich bestimmter Charakteristiken wie Blutdruck, Gewicht und Größe zu beurteilen. Die Einflüsse von Ernährung und Aktivität auf das Gewicht von Vorschulkindern untersuchen Boles *et al.* (2013) und nutzen dabei den ICC, um die Abhängigkeit von sitzenden Tätigkeiten zu erfassen. Thompson *et al.* (2012) finden Zusammenhänge bzgl. des Alters, des Raucherstatus und der Ernährungsgewohnheiten unter Patienten, die die gleiche Arztpraxis besuchen. Die Autoren berechnen Werte für die Intra-Klumpen-Korrelation zwischen 0,118 und 0,265. Eine Untersuchung an schwedischen Geschwisterpaaren nutzt den ICC um den Einfluss der Wohngegend auf das Herzinfarkttrisiko abzuschätzen (Merlo *et al.* (2013)). Lester *et al.* (2013) untersuchen in ihrer Familien-Studie die psychische Gesundheit von Kindern, deren Vater für das US-Militär im Auslandseinsatz gedient hat. Sie gehen dabei von einer intrafamiliären Korrelation des Stresslevels bei den Kindern in Höhe von ca. 0,2 aus.

Ein weiteres demografisches Erklärungsmodell in dem Zusammenhang findet sich in den Arbeiten von Weißbach und Herzog (z. B. Weißbach und Herzog (2009)), die das Auftreten von Karies bei Intra-Klumpen-korrelierten Stichproben untersuchen. Dort liegt das Hauptaugenmerk auf der Stichprobenplanung unter Beachtung von Klumpenkorrelation. Dass es einen Zusammenhang zwischen Klumpenkorrelation und Kariesprävalenz gibt, zeigen Gulliford *et al.* (2005). Denn eine höhere Wahrscheinlichkeit für das Auftreten von Karies hat eine höhere Varianz zur Folge (vgl. Bernoulli-Verteilung für Wahrscheinlichkeiten $Pr(\text{„Erfolg“}) < 0,5$).

Arbeiten aus der Biologie nutzen den ICC bspw. zum Nachweis einer mono- oder dizygoten Schwangerschaft, da die genetischen Übereinstimmungen bei monozygoten Zwillingen höher sind (z. B. Jin *et al.* (2014)). Fisher (1970, Seite 178 ff) widmet ein komplettes Kapitel der empirischen ICC-Berechnung bei (Zwillings-)Paaren. Sehr hohe Werte für den ICC (0,95) erreichen Blokland *et al.* (2014) in ihrer Untersuchung genetischer Faktoren, die auf die Aktivität von Kleinhirnen bei Menschen wirken.

Weitere Anwendung findet der ICC in entsprechenden Studien als Index für die Übereinstimmung zwischen mehreren Beurteilern bei der Begutachtung der betrachteten Elemente (Inter-Rater-Reliabilität) (z. B. Fleiss und Cuzick (1979) oder Chmura Kraemer *et al.* (2002)). Jablonski-Momeni *et al.* (2013) verwenden Werte für den ICC zwischen

0,82 und 0,98 für die Übereinstimmung zweier Gutachter, die eine neue Fluoreszenz-Kamera für die Karieserkennung testeten.

Populäre Ansätze zur Modellierung von Intra-Klumpen-Korrelationen in der Finanzwirtschaft sind Unternehmenswertmodelle wie *Credit Metrics* oder *CreditRisk+*, Intensitätsmodelle mit exogenen Zusammenhängen oder die Modellierung über Copulas (Duffie und Singleton (2003, Seite 230)). In *CreditRisk+* führt das Vorhandensein von Klumpenkorrelation zu erhöhten Kreditausfällen, indem alternativ zur direkten Endogenisierung der Korrelation die Volatilität der Ausfallsrate beachtet wird (Credit Suisse First Boston (1997, Seite 15)). Die Berechnung der Korrelation ist schwierig und geschieht häufig über Umwege und unter Hinzunahme weiterer Annahmen (z.B. über die vorliegende Verteilung).

Bei Insolvenzdaten wird davon ausgegangen, dass die beobachtete Korrelation zwischen Kreditausfällen kausal mit einem „*background factor*“ – der Konjunktur – zusammenhängt (Credit Suisse First Boston (1997, Seite 14)). In einer Rezession steigt beispielsweise die Wahrscheinlichkeit von gemeinsamen Ausfällen (Crouhy *et al.* (2009, Seite 332)). Es kann daher davon ausgegangen werden, dass der ICC über die Zeit veränderlich ist. Weiterhin schwankt die Korrelation zwischen Kreditausfällen mit der Branche von Unternehmen oder deren regionaler Herkunft (Crouhy *et al.* (2009, Seite 332)), was zu Risikokonzentration in Bankenportfolios führen kann. In *CreditRisk+* wird diese durch o.g. zufälligen Effekt auf die Ausfallsrate je Branche implizit modelliert (Rosenow und Weißbach (2009)).

Banken sind nach Kreditwesengesetz (KWG) verpflichtet, das Ausfallrisiko ihrer Kreditnehmer zu beurteilen. Das KWG enthält insbesondere Vorschriften zur Vermeidung von Klumpenrisiken z. B. im Bereich von Kreditnehmereinheiten (Bundesministerium der Justiz und für Verbraucherschutz (2015, §19, Abs. 2, Satz 1-5)) und Länderrisikokonzentrationen (Bundesministerium der Justiz und für Verbraucherschutz (2015, §25, Abs. 3)).

Konzentrationen von Ausfallrisiken (Kreditrisikokonzentrationen) können durch Geschäftsbeziehungen mit einzelnen Schuldner oder Schuldnergruppen entstehen, die eine Reihe gemeinsamer Merkmale aufweisen und deren Fähigkeit zur Schuldentilgung gleichermaßen von der Veränderung bestimmter wirtschaftlicher Rahmenbedingungen abhängt. (Commerzbank (2010, Seite 303))

Die betreffende Bank führt deshalb eine Reihe zusätzlicher Maßnahmen ein, um ihre Verpflichtungen im Falle der Insolvenz eines Kunden erfüllen zu können.

Ein Bestreben der Kreditrisikosteuerung ist daher die aktive Diversifikation zwischen Kreditnehmern, Branchen etc. unter Beachtung des Klumpenrisikos zur Vermeidung von

Adressen- und Branchenkonzentration, regionaler Konzentration und sonstiger Konzentration im Kreditgeschäft. Dabei ist das Rating eines Schuldners ein wichtiges Hilfsmittel (McNeil *et al.* (2005, Seite 338 ff)). Es sagt jedoch nichts über eventuell vorhandene Klumpenkorrelationen aus.

Das verallgemeinerte Ziel einschlägiger Forschung, konservativere Handlungsempfehlungen auszusprechen und damit Vertrauensintervalle konservativer zu schätzen ist zugleich ein moralisches Hemmnis dafür, (Klumpen-)Design-Effekte in Hypothesen-Tests einzu beziehen. Denn der Nachweis signifikanter Wirkungen wird dadurch unwahrscheinlicher. Zusätzlich besteht in vielen praktischen Anwendungen Interesse daran, Kausalzusammenhänge zu verstehen. Der ICC weist – wie andere Zusammenhangsmaße – jedoch nur einen statistischen Zusammenhang nach. Ob dieser zufällig besteht oder durch äußere Einflüsse zu Stande kommt, kann darüber allein nicht erklärt werden. Aalen und Frigessi (2007) geben einen Überblick darüber, wie Statistik dennoch dazu beitragen kann, Kausalzusammenhänge zu verstehen.

Der ICC ist in der Praxis unbekannt und muss geschätzt werden. Die Literatur bietet dazu eine sehr große Anzahl an Schätzern, die verschiedenste Annahmen und Spezialisierungen unterstützen.

Eldridge *et al.* (2009) stellen fest, dass die Ansätze davon abhängig sind, ob das beobachtete Merkmal stetig oder dichotom skaliert ist. Den ICC für stetige Zielvariablen interpretieren sie als Anteil der zwischen Klumpen erzeugten Varianz σ_a^2 an der gesamten Varianz σ^2 :

$$\rho = \frac{\sigma_a^2}{\sigma^2}.$$

Die Darstellung resultiert aus dem Varianzanalyse-Modell

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

mit Erwartungswert μ , gruppenspezifischem Effekt α_i und individuellem Effekt ϵ_{ij} . Neuere Modelle zielen auf Spezialisierungen dieses linearen Modells, auch *random effects model* genannt, ab. Zur Berechnung des ICC für dichotome Variablen schlagen die Autoren ein hierarchisches Modell vor.

Das *intraclass correlation model* (auch: *common correlation model*), welches häufig verwendetes Modell für dichotome Daten ist, wird erstmals in Snedecor und Cochran (1989, Kapitel 13.5) unter diesem Namen erwähnt. Es besagt, dass alle Beobachtungen X_{ij} die gleiche Verteilung besitzen und dass für je zwei Elemente eines Klumpens i der Korrelationskoeffizient $\rho = Kor(X_{ij}, X_{ij'})$ ist.

Fleiss *et al.* (2003, Seite 442) geben eine Einleitung zu einzeln und mehrfach korrelierten Stichproben sowie Methoden wie Logistische Regression für korrelierte Daten. Wooldridge (2003) beschreibt ein lineares, ökonometrisches Modell zur Beachtung von Korrelationen in Daten mit Klumpen-Design. Eine Übersicht über Methoden der statistischen Inferenz bei Anwesenheit von Gruppeneffekten wird gegeben.

Eine bisher weniger beachtete Methode basiert auf einem Bayes-Ansatz. Bei Beachtung von Korrelationen von Kreditausfällen, verschiebt sich Masse von der Binomialverteilung hin zu einer neuen Verteilung mit höherer Wahrscheinlichkeit für höhere Ausfallraten (Tasche (2013)). Yamamoto und Yanagimoto (1992) stellen einen auf der Momentenmethode basierenden, unverzerrten Schätzer für balancierte Klumpen innerhalb einer Beta-Binomial-Verteilung vor. Die Autoren verweisen auf drei Beispiele für extreme Situationen, in denen der Schätzer nicht existiert oder einen Wert außerhalb des zu erwartenden Intervalls hat.

Speziell bei dichotomen Merkmalen wurden in der Vergangenheit eine Vielzahl von Schätzern der Intra-Klumpen-Korrelation vorgestellt. Die Eigenschaften dieses „Potentials“ hängen stark ab vom Studiendesign (balanciert, unbalanciert), Anzahl und Größe der Klumpen sowie der Homogenität der Klumpen hinsichtlich der Zielvariablen. Einige der Schätzer weisen Äquivalenzen oder identisches asymptotisches Verhalten auf (Ridout *et al.* (1999)).

Ziel dieser Arbeit ist die Vorstellung bzw. Herleitung und Analyse von elf grundverschiedenen Methoden zur Schätzung der Intra-Klumpen-Korrelation (Kapitel 3), sowie deren Katalogisierung und Einordnung in verschiedene Modellklassen (Kapitel 2). In Kapitel 5.1 werden zwei neue Datensätze mit Klumpen-Design vorgestellt. Anhand der Ergebnisse, die die Schätzer angewandt auf diese Daten liefern (Kapitel 5), und einem Simulationenvergleich in Kapitel 4 wird die Güte der Schätzer diskutiert. Schließlich wird aufgezeigt, welchen kausalen Ursprung die Intra-Klumpen-Korrelation im Beispiel haben kann (u. a. Kapitel 5.4). So zeigt sich, dass – je nach Verteilungsannahme des unterliegenden Modells – der Intra-Klumpen-Korrelation unterschiedliche kausale Herkunft zugewiesen werden kann. Den Beobachtungen X_{ij} seien dabei drei unterschiedliche Verteilungen unterstellt:

1. $X_{ij} \sim Ber(\pi)$: X_{ij} sind unabhängig, identisch verteilt: keine Intra-Klumpen-Korrelation
2. $X_{ij} \sim CCM(\rho, \pi)$: X_{ij} sind nicht unabhängig, identisch verteilt: Modell mit pauschaler Klumpenkorrelation (auch: *common correlation model* (CCM), Kapitel 2.1)
3. $X_{ij} |_{\pi_i} \sim Ber(\Pi(\alpha, \beta))$: X_{ij} sind bedingt unabhängig und nicht identisch verteilt: Beta-Binomial-Modell (Kapitel 2.2).

Kapitel 6 betrachtet die vorgestellten Methoden noch einmal kritisch und gibt einen Ausblick darauf, wie einige dieser Kritikpunkte auszuräumen sind.

Die Datenanalysen erfolgen mit der Statistik-Software SAS (siehe z. B. Krämer *et al.* (2008)). Relevante Quellcodes sind im Anhang A.3 aufgeführt.

2 Modelle

Die Modellierung von Querschnittsdaten ist grundsätzlich mit einem linearen Ansatz möglich. Binäre Zielgrößen lassen sich mit Logistischen Modellen effizienter abbilden.

Für die Modellierung binärer Intra-Klumpen-korrelierter Daten seien in dieser Arbeit zwei unterschiedliche Grundannahmen getroffen, die entsprechend zu zwei verschiedenen mathematischen Modellen führen. Zum einen sei angenommen, dass die Klumpenkorrelation endogen ist, sie also in Form des Parameters ρ pauschal als gegeben angesehen und direkt modelliert wird (siehe Kapitel 2.1, Modell mit pauschaler Klumpenkorrelation). Zweitens sei angenommen, dass die Klumpen sich in ihren Eigenschaften genau so unterscheiden, dass in Ihnen Zusammenhänge nachgewiesen werden können (siehe Kapitel 2.2, Beta-Binomial-Modell). Die Ursächlichkeit der Zusammenhänge kann in beiden Modellen jeweils auf zwei verschiedene Arten erklärt werden: zum einen in Form von gegenseitiger Beeinflussung der Elemente und zum anderen als Beeinflussung der Elemente durch externe Faktoren.

2.1 Modell mit pauschaler Klumpenkorrelation

Eine Stichprobe bestehe aus k zufällig gezogenen Klumpen. Jeder Klumpen enthalte n_i ($i = 1, \dots, k$) Elemente mit $\sum_i n_i = N$, die Stichprobengröße. Die Zufallsvariable X_{ij} ($i = 1, \dots, k; j = 1, \dots, n_i$) beschreibt jeweils für das j -te Element des i -ten Klumpens den Ausgang eines binären Bernoulli-Experimentes mit den Ausgängen „Erfolg“ und „Misserfolg“, kodiert als 1 bzw. 0.

Im Modell mit pauschaler Klumpenkorrelation (auch: *common correlation model* (CCM)) wird diese Klumpenstruktur zu einer Anpassung der nominalen Varianz in der Zufallsvariable der Gesamtzahl an Erfolgen in Klumpen i , $Y_i = \sum_j X_{ij}$, führen, wenn folgende zentrale Bedingung für $\rho > 0$ gilt:

$$\text{Kor}(X_{ij}, X_{i'j'}) = \begin{cases} \rho, & i = i', j \neq j' \\ 1, & i = i', j = j' \\ 0, & i \neq i'. \end{cases} \quad (2.1)$$

D.h. innerhalb eines jeden Klumpens herrsche ein jeweils paarweiser Zusammenhang unter den Elementen in Höhe von ρ . Elemente aus jeweils verschiedenen Klumpen seien nicht korreliert. ρ wird auch Intra-Klumpen-Korrelationskoeffizient (engl. *intraclass correlation coefficient* (ICC)) genannt. Der ICC beschreibt nicht zwingend einen kausalen Zusammenhang zwischen den Elementen. In Form einer Scheinkorrelation kann ein äußerer Faktor Ursache für den gemessenen Zusammenhang sein (vgl. Kapitel 6). Plausible Werte für die Intra-Klumpen-Korrelation liegen zwischen Null und Eins, wobei die Grenzen angenommen werden dürfen.

Die durch Bedingung 2.1 eingeführte Zusammenhgangsstruktur bedingt nun, dass die Verteilung der aggregierten Erfolge Y_i von der Binomialverteilung merklich abweicht. Im Fall $k = 1$ und $n \equiv n_1 = 2$ lassen sich die Wahrscheinlichkeiten der Gesamtanzahl von Erfolgen in einem Klumpen anschaulich mit Hilfe der Vierfeldertafel bestimmen.

X_1	X_2		Σ
	0	1	
0	p_{11}	p_{12}	$p_{1\cdot} = (1 - \pi)$
1	p_{21}	p_{22}	$p_{2\cdot} = \pi$
Σ	$p_{\cdot 1} = (1 - \pi)$	$p_{\cdot 2} = \pi$	

Y kann nun die Werte 0, 1 und 2 (Erfolge) annehmen. Ausgehend von der Pearson'schen Definition des Korrelationskoeffizienten gilt

$$\rho = \frac{Kov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}$$

und somit

$$Kov(X_1, X_2) = \rho\pi(1 - \pi).$$

Mit der Definition der Kovarianz $Kov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$ und $E(X_1) = E(X_2) = \pi$ lässt sich nun die erste gesuchte Wahrscheinlichkeit berechnen als

$$\begin{aligned} Pr(Y = 2) &= p_{22} = Pr(X_1 = 1, X_2 = 1) = E(X_1X_2) \\ &= Kov(X_1, X_2) + \pi^2 = \rho\pi(1 - \pi) + \pi^2 \\ &= \rho\pi + (1 - \rho)\pi^2. \end{aligned}$$

Mit der Randwahrscheinlichkeit $p_{2\cdot} = \pi$ ergibt sich

$$\begin{aligned} p_{21} &= \pi - p_{22} \\ &= \pi - (\rho\pi + (1 - \rho)\pi^2) \\ &= (1 - \rho)\pi(1 - \pi) \\ &= p_{12}, \end{aligned}$$

wobei das letzte Gleichheitszeichen aus Symmetriegründen besteht. Die Wahrscheinlichkeit für genau einen Erfolg ist nun

$$\begin{aligned} Pr(Y = 1) &= p_{12} + p_{21} \\ &= 2(1 - \rho)\pi(1 - \pi). \end{aligned}$$

Daher ist die Wahrscheinlichkeit für genau null Erfolge (Gegenwahrscheinlichkeit)

$$\begin{aligned} Pr(Y = 0) = p_{11} &= 1 - (p_{22} + p_{12} + p_{21}) \\ &= 1 - (\rho\pi + (1 - \rho)\pi^2 + 2(1 - \rho)\pi(1 - \pi)) \\ &= 1 - \rho\pi - (1 - \rho)\pi^2 - 2(1 - \rho)(\pi - \pi^2) \\ &= 1 - \rho\pi + (1 - \rho)(-2\pi + \pi^2) \\ &= 1 - \rho\pi + (1 - \rho)(1 - 2\pi + \pi^2) - (1 - \rho) \\ &= \rho(1 - \pi) + (1 - \rho)(1 - \pi)^2. \end{aligned}$$

Der allgemeine Fall geht auf [Madsen \(1993\)](#) zurück.

Definition 2.1 (*CCM-Verteilung*)

Für alle $1 \leq i \leq k$, $k \geq 1$ und $n_i \geq 2$ ist

$$Pr_{\rho, \pi}(y_i) = Pr(Y_i = y_i) = \begin{cases} \rho(1 - \pi) + (1 - \rho)(1 - \pi)^{n_i}, & y_i = 0, \\ \binom{n_i}{y_i} (1 - \rho)\pi^{y_i}(1 - \pi)^{n_i - y_i}, & 1 \leq y_i \leq n_i - 1, \\ \rho\pi + (1 - \rho)\pi^{n_i}, & y_i = n_i \end{cases} \quad (2.2)$$

eine Wahrscheinlichkeitsfunktion, wenn gilt:

$$\max \left[-\frac{(1 - \pi)^{n_i}}{(1 - \pi) - (1 - \pi)^{n_i}}, -\frac{\pi^{n_i}}{\pi - \pi^{n_i}} \right] \leq \rho \leq 1.$$

Die zu dieser Wahrscheinlichkeitsfunktion gehörende Verteilung heißt CCM-Verteilung.

Die Wahrscheinlichkeitsfunktion ist abschnittsweise definiert mit drei disjunkten Definitionsbereichen, wobei der Schwerpunkt der Wahrscheinlichkeitsmasse auf dem 2. Abschnitt liegt. Die Masse verschiebt sich hin zum 1. und 3. Abschnitt, wenn ρ gegen 1 geht¹:

$$\begin{array}{ll}
 1. & \rho(1 - \pi) + (1 - \rho)(1 - \pi)^{n_i}, \quad y_i = 0 \\
 2. & \binom{n_i}{y_i} (1 - \rho)\pi^{y_i}(1 - \pi)^{n_i - y_i}, \quad 1 \leq y_i \leq n_i - 1 \\
 3. & \rho\pi + (1 - \rho)\pi^{n_i}, \quad y_i = n_i.
 \end{array}
 \quad \rho \nearrow 1 \quad \left. \begin{array}{c} \updownarrow \\ \text{Masse} \end{array} \right\}$$

Das erste und zweite Moment der CCM-verteilten Zufallsvariable Y_i sind²

$$E(Y_i) = n_i\pi, \quad Var(Y_i) = n_i\pi(1 - \pi) \underbrace{[1 + (n_i - 1)\rho]}_{\text{VIF}}, \quad 0 \leq i \leq n_i. \quad (2.3)$$

Die CCM-Verteilung besitzt somit den gleichen Erwartungswert wie die Binomialverteilung. Die Varianz ist jedoch um den sogenannten Varianzinflationsfaktor VIF erhöht, falls $\rho > 0$ ist (Überdispersion). Werte von $\rho < 0$ führen zu Unterdispersion. Da negative Werte von ρ in der Praxis häufig unplausibel sind, wird dieser Fall hier nicht weiter betrachtet. Bei $\rho = 0$ liegt keine Klumpenkorrelation vor und die Y_i sind jeweils binomialverteilt mit Parametern (n_i, π) . $\rho = 1$ ist der Spezialfall perfekter positiver Abhängigkeit. Er entspricht einer Bernoulliverteilung auf $\{0, n_i\}$ mit Parameter π . Dass selbst äußerst kleine Werte von ρ zu einem relativ großen VIF führen können, zeigt Tabelle 2.1.

		Klumpengröße		
		100	1000	5000
ρ	0,01	2	11	51
	0,001	1,1	2	6

Tabelle 2.1: VIF in Abhängigkeit von ICC und Klumpengröße

So wird beispielsweise bei einer Intra-Klumpen-Korrelation von $\rho = 0,01$ die Varianz in einem Klumpen mit 1000 Elementen um den Faktor 11 höher liegen als bei einem Vergleichsklumpen, bei dem die Elemente nicht paarweise korrelieren. Adams *et al.* (2004) betrachten verschiedene medizinische Klumpen-Design-Studien mit ICCs zwischen 0,007

¹Der Grenzwert von $Pr_{\rho, \pi}(y_i)$ für ρ gegen 1 existiert, ist aber nicht gleich dem Funktionswert $Pr_{\rho, \pi}(y_i)|_{\rho=1}$

²Die Herleitungen befinden sich im Anhang A.2.1.

und 0,66 bei durchschnittlichen Klumpengrößen zwischen 6 und 125. Bei entsprechenden Stichproben ist der Varianzinflationsfaktor mit bis zu Faktor 19 zu bedenken.

Aus der CCM-Verteilung ergeben sich folgende nützliche Einzelwahrscheinlichkeiten für $k > 1$ (analog zum Fall $k = 1$):

$$Pr(X_{ij} = 1, X_{i'j'} = 1) = \begin{cases} \pi^2 + \pi(1 - \pi)\rho, & i = i', j \neq j' \\ \pi^2, & i \neq i', j \neq j' \end{cases} \quad (2.4)$$

$$Pr(X_{ij} = 0, X_{i'j'} = 0) = \begin{cases} (1 - \pi)^2 + \pi(1 - \pi)\rho, & i = i', j \neq j' \\ (1 - \pi)^2, & i \neq i', j \neq j' \end{cases}$$

$$\begin{aligned} Pr(X_{ij} = X_{i'j'}) &= Pr(X_{ij} = 1, X_{i'j'} = 1) \\ &\quad + Pr(X_{ij} = 0, X_{i'j'} = 0) \\ &= \begin{cases} 1 - 2\pi(1 - \pi)(1 - \rho), & i = i', j \neq j' \\ 1 - 2\pi(1 - \pi), & i \neq i', j \neq j' \end{cases} \end{aligned} \quad (2.5)$$

$$Pr(X_{ij} \neq X_{i'j'}) = 1 - Pr(X_{ij} = X_{i'j'}) = \begin{cases} 2\pi(1 - \pi)(1 - \rho), & i = i', j \neq j' \\ 2\pi(1 - \pi), & i \neq i', j \neq j'. \end{cases}$$

Die CCM-Verteilung ist abhängig von drei Parametern: n_i, π und ρ , wobei n_i über die Klumpengröße gegeben ist. Im Folgenden wird jeweils auf den Stichprobenschätzer für die Parameter Erfolgswahrscheinlichkeit und Intra-Klumpen-Korrelation eingegangen.

Punktschätzung für die Erfolgswahrscheinlichkeit π

Ein intuitiver Schätzer für π ist der Maximum-Likelihood-Schätzer (ML). Die Likelihood-Funktion der CCM-Verteilung ist gegeben durch

$$\begin{aligned} L(y_1, \dots, y_n | \rho, \pi) &= \prod_{i=1}^k Pr_{\rho, \pi}(y_i) \\ &= \prod_{i=1}^k \left[\rho(1 - \pi) + (1 - \rho)(1 - \pi)^{n_i} \right]^{I_{\{y_i=0\}}} \\ &\quad \cdot \left[\binom{n_i}{y_i} (1 - \rho)\pi^{y_i}(1 - \pi)^{n_i - y_i} \right]^{I_{\{1 \leq y_i \leq n_i - 1\}}} \\ &\quad \cdot \left[\rho\pi + (1 - \rho)\pi^{n_i} \right]^{I_{\{y_i=n_i\}}}. \end{aligned}$$

Da $\binom{n_i}{y_i}$ nicht von ρ abhängt, wird es vernachlässigt und der Log-Likelihood-Kern ist

$$\begin{aligned} l(y_1, \dots, y_n | \rho, \pi) \propto & \sum_{i=1}^k \left[\ln [(\rho(1-\pi) + (1-\rho)(1-\pi)^{n_i})^{I_{\{y_i=0\}}}] \right. \\ & + \ln [((1-\rho)\pi^{y_i}(1-\pi)^{n_i-y_i})^{I_{\{1 \leq y_i \leq n_i-1\}}}] \\ & \left. + \ln [(\rho\pi + (1-\rho)\pi^{n_i})^{I_{\{y_i=n_i\}}}] \right]. \end{aligned} \quad (2.6)$$

Für die dazugehörige Score-Funktion ergibt sich

$$\begin{aligned} \frac{\partial l(y_1, \dots, y_n | \rho, \pi)}{\partial \pi} \propto & \sum_{i=1}^k \left[-\frac{\rho + (1-\rho)n_i(1-\pi)^{n_i-1}}{\rho(1-\pi) + (1-\rho)(1-\pi)^{n_i}} \cdot I_{\{y_i=0\}} + \frac{y_i - n_i\pi}{\pi(1-\pi)} \cdot I_{\{1 \leq y_i \leq n_i-1\}} \right. \\ & \left. + \frac{\rho + (1-\rho)n_i\pi^{n_i-1}}{\rho\pi + (1-\rho)\pi^{n_i}} \cdot I_{\{y_i=n_i\}} \right]. \end{aligned}$$

Nullstellensuche in der Score-Funktion hat den ML-Schätzer zum Ergebnis:

$$\hat{\pi}_{ML} = \frac{\sum_{i=1}^k y_i}{N}, \quad \text{falls } 1 \leq y_i \leq n_i - 1, \forall i.^3$$

Die Varianz dieses Schätzers lässt sich leicht berechnen, da das zweite Moment der CCM-Verteilung bekannt ist (siehe Gleichung 2.3) und $\sum_{i=1}^k n_i = N$ definiert wurde.

$$\begin{aligned} \text{Var}(\hat{\pi}_{ML}) &= \frac{\sum_{i=1}^k \text{Var}(Y_i)}{N^2} \\ &\stackrel{2.3}{=} \frac{1}{N^2} \sum_{i=1}^k n_i \pi (1-\pi) [1 + (n_i - 1)\rho] \\ &= \frac{1}{N^2} \left[\sum_{i=1}^k n_i \pi (1-\pi) + \sum_{i=1}^k n_i^2 \pi (1-\pi) \rho - \sum_{i=1}^k n_i \pi (1-\pi) \rho \right] \\ &= \frac{\pi(1-\pi)}{N} \underbrace{\left[1 + \left(\frac{\sum n_i^2}{N} - 1 \right) \rho \right]}_{\text{VIF}^*}. \end{aligned}$$

Damit ist die Varianz des Anteilsschätzers von Y_i $\hat{\pi}_{ML}$ gleich $\frac{\pi(1-\pi)}{N}$ multipliziert mit dem Varianzinflationsfaktor VIF*, der im Gegensatz zum klumpenspezifischen VIF durch die gewichtete durchschnittliche Klumpengröße dominiert wird (vgl. auch Fleiss *et al.* (2003, Kapitel 15.1)).

³Diese Einschränkung ist in vielen praktischen Anwendungen gerechtfertigt.

Punktschätzung für die Intra-Klumpen-Korrelation ρ

Für die Punktschätzung von ρ ist zunächst wieder Maximum-Likelihood die natürliche Methode. Die Score-Funktion ergibt sich aus der Log-Likelihood 2.6 wie folgt:

$$\begin{aligned} \frac{\partial l(y_1, \dots, y_n | \rho, \pi)}{\partial \rho} = & \sum_{i=1}^k \left[\frac{(1-\pi) - (1-\pi)^{n_i}}{\rho(1-\pi) + (1-\rho)(1-\pi)^{n_i}} \cdot I_{\{y_i=0\}} - \frac{1}{(1-\rho)} \cdot I_{\{1 \leq y_i \leq n_i-1\}} \right. \\ & \left. + \frac{\pi - \pi^{n_i}}{\rho\pi - (1-\rho)\pi^{n_i}} \cdot I_{\{y_i=n_i\}} \right]. \end{aligned} \quad (2.7)$$

Der ML-Schätzer von ρ ist damit impliziert durch

$$\frac{k}{1 - \hat{\rho}_{ML}} = 0, \quad \text{falls } 1 \leq y_i \leq n_i - 1, \forall i.$$

Wenn in jedem der k Klumpen die Anzahl der Erfolge mindestens 1 und maximal $n_i - 1$ ist, dann ist der ML-Schätzer von ρ entweder $-\infty$ oder $+\infty$. Da ρ per Definition im Intervall $[0, 1]$ liegt, ist der ML-Schätzer dann keine zulässige Lösung.⁴ Simulationen basierend auf der CCM-Verteilung zeigen, dass es höchst unwahrscheinlich ist, einen Datensatz zu erzeugen, in dem es Klumpen mit Null oder n_i Erfolgen gibt.⁵ Auch die in dieser Arbeit verwendeten Daten rechtfertigen diese Annahme. Deshalb ist es wahrscheinlich, dass es bei vielen praktischen Anwendungen keinen zulässigen Maximum-Likelihood-Schätzer für die Intra-Klumpen-Korrelation gibt.

Im Gegensatz dazu lässt sich leicht zeigen, dass das Modell identifizierbar ist und somit zumindest theoretisch eine Möglichkeit der wahren Schätzung von ρ gegeben ist, da die Wahrscheinlichkeitsfunktion 2.2 strikt monoton für $n_i \geq 2, \forall i$ und $\pi \neq \{0, 1\}$ ist.

Lemma 2.2 *Die Wahrscheinlichkeitsfunktion der CCM-Verteilung ist strikt monoton in ρ .*

Beweis:

Zu zeigen ist, dass jedes Teilstück der stückweise definierten Wahrscheinlichkeitsfunktion bei Variation von ρ niemals konstant ist. Für alle i und $n_i \geq 2$ gilt folgende Unterscheidung:

⁴Für weitere Beispiele von inkonsistenten Schätzern siehe Lancaster (2000) sowie Lehmann und Casella (1998).

⁵SAS-Programmcode von Simulationen mit Klumpengrößen zwischen 11 und 107 befinden sich im Anhang A.3.1.

- $y_i = 0$:

$$\begin{aligned} Pr_{\rho, \pi}(y_i) &= \rho(1 - \pi) + (1 - \rho)(1 - \pi)^{n_i} \\ \frac{\partial Pr_{\rho, \pi}(y_i)}{\partial \rho} &= (1 - \pi) - (1 - \pi)^{n_i} \neq 0 \iff \pi \notin \{0, 1\} \end{aligned}$$

- $y_i \in [1, n_i - 1]$:

$$\begin{aligned} Pr_{\rho, \pi}(y_i) &= \binom{n_i}{y_i} (1 - \rho)\pi^{y_i}(1 - \pi)^{n_i - y_i} \\ \frac{\partial Pr_{\rho, \pi}(y_i)}{\partial \rho} &= -\pi^{y_i}(1 - \pi)^{n_i - y_i} \neq 0 \iff \pi \notin \{0, 1\} \end{aligned}$$

- $y_i = n_i$:

$$\begin{aligned} Pr_{\rho, \pi}(y_i) &= \rho\pi + (1 - \rho)\pi^{n_i} \\ \frac{\partial Pr_{\rho, \pi}(y_i)}{\partial \rho} &= \pi - \pi^{n_i} \neq 0 \iff \pi \notin \{0, 1\} \end{aligned}$$

Mathematisch gesehen ist ein Modell genau dann identifizierbar, wenn für unterschiedliche Parameterwerte auch die Verteilungen unterschiedlich sind. Da in allen der drei o. g. Fälle die Verteilungsfunktion bei Veränderung von ρ variiert, ist dies hier erfüllt. Erfolgswahrscheinlichkeiten in Höhe von 0 oder 1 sind in der Praxis nicht relevant. \square

Gouriéroux und Monfort (1995, Seite 95) legen nun in Property 3.13 dar, dass die Funktion $\rho \mapsto E_{\rho_0} \ln Pr_{\rho, \pi}(Y)$ ein eindeutiges Maximum an der Stelle $\rho = \rho_0$ hat, wenn das Modell identifizierbar ist. Folgendes Lemma deutet den Zusammenhang zwischen $E_{\rho_0} \ln Pr_{\rho, \pi}(Y)$ und der Likelihood-Funktion der CCM-Verteilung an und beweist, dass zumindest asymptotisch ein zulässiger Maximum-Likelihood-Schätzer existiert.

Lemma 2.3 $\arg \max_{\rho} E_{\rho_0} \ln Pr_{\rho, \pi}(Y_i)$ ist asymptotisch gleich $\arg \max_{\rho} L(y_1, \dots, y_n | \rho, \pi)$ und es existiert ein eindeutiger Maximum-Likelihood-Schätzer, der gleich dem wahren Parameterwert ρ_0 ist.

Beweis:

Es ist $Pr_{\rho, \pi}(y_i) = Pr_{\rho, \pi}(y_i | n_i) = \frac{Pr_{\rho, \pi}(y_i, n_i)}{Pr(n_i)}$. Dann folgt aus dem Gesetz der großen Zahlen und der Eigenschaft der Zufallsvariablen Y_i , unabhängig und identisch verteilt zu sein, dass

$$\begin{aligned} \arg \max_{\rho} \prod_{i=1}^k Pr_{\rho, \pi}(y_i | n_i) &= \arg \max_{\rho} \sum_{i=1}^k \ln Pr_{\rho, \pi}(y_i | n_i) \\ &= \arg \max_{\rho} \left[\sum_{i=1}^k \ln Pr_{\rho, \pi}(y_i, n_i) - \sum_{i=1}^k \ln Pr(n_i) \right] \end{aligned}$$

und weiter, da $Pr(n_i)$ nicht von ρ abhängt

$$\begin{aligned} &= \arg \max_{\rho} \frac{1}{k} \sum_{i=1}^k \ln Pr_{\rho, \pi}(y_i, n_i) \\ &\xrightarrow{k \rightarrow \infty} \arg \max_{\rho} E_{\rho_0} \ln Pr_{\rho, \pi}(y, n) \quad =: \rho_0. \end{aligned}$$

Zusätzlich sichert die o. s. Gleichung die Konsistenz des ML-Schätzers für ρ . Asymptotische Normalität kann gezeigt werden (siehe dazu [Kremer et al. \(2014\)](#)). \square

Ein Datensatz mit mindestens einem Klumpen, der *nicht* in die Gruppe 1 bis $n_i - 1$ Erfolge fällt, hat einen zulässigen ML-Schätzer. Diese Situation wird umso wahrscheinlicher je größer k ist. Für $k \rightarrow \infty$ existiert garantiert ein zulässiger ML-Schätzer.

Beispiel 2.4 (zulässiger Maximum-Likelihood-Schätzer)

Ein Datensatz bestehe aus k Klumpen wobei genau ein Klumpen i^* n_{i^*} Erfolge hat und die restlichen $k - 1$ Klumpen jeweils zwischen 1 und $n_i - 1$ Erfolge haben. Dann enthält die Score-Funktion für ρ Beiträge aus $-\frac{1}{(1-\rho)}$ und $\frac{\pi - \pi^{n_{i^*}}}{\rho\pi - (1-\rho)\pi^{n_{i^*}}}$ (vgl. Ableitung der Log-Likelihood 2.7). Der ML-Schätzer ist in diesem Fall explizit angegeben mit

$$\hat{\rho}_{ML} = \frac{(2 - k)\pi^{n_{i^*}} - \pi}{(2 - k)\pi^{n_{i^*}} - \pi k}. \quad \square$$

Der oben erklärte Umstand, dass die bloße Erhöhung des Stichprobenumfanges nicht genügt, um den ML-Schätzer näher an den wahren Parameterwert zu bringen, rechtfertigt die Suche nach anderen Schätzern für die Intra-Klumpen-Korrelation. Kapitel 3 widmet sich diesem Thema ausführlicher.

2.2 Beta-Binomial-Modell

Die Ausgangssituation gleicht der des Modells mit pauschaler Klumpenkorrelation: Eine Stichprobe bestehe aus k zufällig gezogenen Klumpen. Jeder Klumpen enthalte n_i ($i = 1, \dots, k$) Elemente mit $\sum_i n_i = N$. Die Zufallsvariable X_{ij} ($i = 1, \dots, k; j = 1, \dots, n_i$) beschreibt jeweils für das j -te Element des i -ten Klumpens den Ausgang eines binären Bernoulli-Experimentes mit den Ausgängen „Erfolg“ und „Misserfolg“ mit der Erfolgswahrscheinlichkeit π_i . Der Wert π_i sei für jeden Klumpen i jeweils Realisation der Beta-verteilten Zufallsvariable Π mit Parametern α und β , d. h.

1. Stufe: $\Pi \sim \text{Beta}(\alpha, \beta)$ ⁶,
2. Stufe: $X_{ij} |_{\pi_i} \sim \text{Ber}(\pi_i)$.

Elemente innerhalb eines Klumpens besitzen somit die gleiche Erfolgswahrscheinlichkeit. Elemente unterschiedlicher Klumpen unterscheiden sich in ihrer Erfolgswahrscheinlichkeit. Diese Eigenschaft veränderlicher Erfolgswahrscheinlichkeiten wird in der Literatur als Regionale oder Latente⁷ Heterogenität bezeichnet. Im Beta-Binomial-Modell (BBM) führt eine solche Beschaffenheit dazu, dass die individuelle Größe X_{ij} Beta-Bernoulliverteilt ist und die aggregierte Größe der Gesamtzahl an Erfolgen in Klumpen i , $Y_i = \sum_j X_{ij}$, Beta-Binomialverteilt ist mit Parametern (α, β, n_i) , d. h.

$$Y_i \sim \text{Beta} - \text{Bin}(\alpha, \beta, n_i),$$

(vgl. Cox und Snell (1989, Kapitel 3.2)). Latente Heterogenität führt weiter zu Varianzinflation in Y_i , wie die folgenden Berechnungen zeigen. Zunächst sei bemerkt, dass Y_i bedingt auf die Realisation von Π binomial verteilt ist. Die ersten beiden Momente sind dann gegeben durch

$$E(Y_i |_{\pi_i}) = n_i \pi_i, \quad \text{Var}(Y_i |_{\pi_i}) = n_i \pi_i (1 - \pi_i), \quad \forall i.$$

Der unbedingte Erwartungswert von Y_i ist (Satz von der iterierten Erwartung)

$$E(Y_i) = E(E(Y_i |_{\Pi})) = n_i E(\Pi) = \frac{n_i \alpha}{\alpha + \beta}.$$

⁶Erwartungswert einer Beta-verteilten Zufallsvariable ist $\frac{\alpha}{\alpha + \beta}$, Varianz $\frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$.

⁷i. S. v. unbeobachtet

Mit Hilfe der Varianzzerlegung nach der ersten Ordnung (Herleitung im Anhang A.2.2) ist die (unbedingte) Varianz von Y_i

$$\begin{aligned}
\text{Var}(Y_i) &= E(\text{Var}(Y_i|\Pi)) + \text{Var}(E(Y|\Pi)) \\
&= E(n_i\Pi(1-\Pi)) + \text{Var}(n_i\Pi) \\
&= n_i E(\Pi) - n_i E(\Pi^2) + n_i^2 \text{Var}(\Pi) \\
&= n_i E(\Pi) - n_i E(\Pi^2) + \underbrace{n_i E(\Pi^2) - n_i E(\Pi)^2}_{=-n_i \text{Var}(\Pi)} + n_i^2 \text{Var}(\Pi) \\
&= n_i E(\Pi)(1 - E(\Pi)) + n_i(n_i - 1)\text{Var}(\Pi). \tag{2.8}
\end{aligned}$$

Für die Varianz von Π gilt aber auch

$$\begin{aligned}
\text{Var}(\Pi) &= \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} \\
&= \frac{1}{\alpha + \beta + 1} \left(\frac{\alpha^2 + \alpha\beta - \alpha^2}{(\alpha + \beta)^2} \right) \\
&= \frac{1}{\alpha + \beta + 1} \left(\frac{\alpha}{(\alpha + \beta)} - \frac{\alpha^2}{(\alpha + \beta)^2} \right) \\
&= \frac{1}{\alpha + \beta + 1} \left(\frac{\alpha}{\alpha + \beta} \left(1 - \frac{\alpha}{\alpha + \beta} \right) \right) \\
&= \frac{1}{\alpha + \beta + 1} \left(E(\Pi)(1 - E(\Pi)) \right).
\end{aligned}$$

Dieses Ergebnis mit $\rho := \frac{1}{\alpha + \beta + 1}$ eingesetzt in Gleichung 2.8 ergibt eine gewichtige Darstellung der Varianz in Abhängigkeit vom Erwartungswert der Erfolgswahrscheinlichkeit.

$$\begin{aligned}
\text{Var}(Y_i) &= n_i E(\Pi)(1 - E(\Pi)) + n_i(n_i - 1)\rho E(\Pi)(1 - E(\Pi)) \\
&= n_i E(\Pi)(1 - E(\Pi)) \underbrace{(1 + (n_i - 1)\rho)}_{\text{VIF}}.
\end{aligned}$$

Das zweite Moment der Beta-Binomialverteilten Variable Y_i gleicht demnach formell dem einer CCM-verteilten Größe (vgl. Gleichung 2.3). Für den i -ten Klumpen bedeutet dies, dass die Varianz der Gesamtzahl an Erfolgen um den Varianzinflationsfaktor VIF erhöht ist, falls $\rho > 0$ ist.

Lemma 2.5 ρ ist ein Maß für die paarweise Abhängigkeit Beta-Bernoulliverteilter Elemente innerhalb eines Klumpens, d. h. es gilt

$$\rho = \text{Kor}(X_{ij}, X_{ij'}), \quad \forall i.$$

Beweis:

X_{ij} sei eine Beta-Bernoulliverteilte Zufallsvariable, d. h. $\Pi \sim \text{Beta}(\alpha, \beta)$, $X_{ij} |_{\pi_i} \sim \text{Ber}(\pi_i)$.
Dann gilt für die ersten Momente

$$\begin{aligned}
 E(X_{ij}) &= E(\underbrace{E(X_{ij} | \Pi)}_{\Pi}) \\
 &= \frac{\alpha}{\alpha + \beta} \\
 \text{Var}(X_{ij}) &= E(\underbrace{\text{Var}(X_{ij} | \Pi)}_{(1-\Pi)\Pi}) + \text{Var}(\underbrace{E(X_{ij} | \pi_i)}_{\pi_i}) \\
 &= E(\Pi) - E(\Pi^2) + \text{Var}(\Pi) \\
 &= E(\Pi) - \left(E(\Pi^2) - E(\Pi)^2 + E(\Pi)^2 \right) + \text{Var}(\Pi) \\
 &= E(\Pi) - \left(\text{Var}(\Pi) + E(\Pi)^2 \right) + \text{Var}(\Pi) \\
 &= E(\Pi) - E(\Pi)^2 \\
 &= \frac{\alpha\beta}{(\alpha + \beta)^2}.
 \end{aligned}$$

Damit schließt sich der Beweis mittels

$$\begin{aligned}
 \text{Kor}(X_{ij}, X_{ij'}) &= \frac{\text{Kov}(X_{ij}, X_{ij'})}{\text{Var}(X_{ij})} \\
 &= \frac{E(X_{ij}X_{ij'}) - E(X_{ij})E(X_{ij'})}{\text{Var}(X_{ij})} \\
 &= \frac{E(E(X_{ij}X_{ij'}) | \Pi) - E(X_{ij})E(X_{ij'})}{\text{Var}(X_{ij})} \\
 &= \frac{E(E(X_{ij}X_{ij'} | \Pi)) - E(X_{ij})E(X_{ij'})}{\text{Var}(X_{ij})}.
 \end{aligned}$$

X_{ij} sind bedingt auf Π unabhängig (unbedingt aber abhängig)

$$\begin{aligned}
 &= \frac{E(\overbrace{E(X_{ij} | \Pi)}^{\Pi} \overbrace{E(X_{ij'} | \Pi)}^{\Pi})) - E(X_{ij})E(X_{ij'})}{\text{Var}(X_{ij})} \\
 &= \frac{E(\Pi^2) - E(\Pi)^2}{\text{Var}(X_{ij})} \\
 &= \frac{\text{Var}(\Pi)}{\text{Var}(X_{ij})} \\
 &= \frac{1}{\alpha + \beta + 1} = \rho. \quad \square
 \end{aligned}$$

Bemerkung: ρ liegt stets im Intervall $(0, 1)$. Insbesondere enthält das Beta-Binomial-Modell nicht die Möglichkeit der Unkorreliertheit ($\rho = 0$). Diese Eigenschaft vererbt sich aus der Forderung positiver Parameter α und β in der Beta-Verteilung um dort Normierbarkeit zu gewährleisten.

Da ρ ein Maß für die paarweise Abhängigkeit der Elemente innerhalb eines Klumpens ist, ist auch im BBM Varianzinflation durch Intra-Klumpen-Korrelation nachgewiesen. Im Gegensatz zum CCM resultiert sie hier aus der Verschiedenheit der Klumpen hinsichtlich der Erfolgswahrscheinlichkeit.

Zusammenfassung

Das Modell mit pauschaler Klumpenkorrelation und das Beta-Binomial-Modell sind zwei Möglichkeiten der Modellierung binärer, Intra-Klumpen-korrelierter Daten. Die Frage nach der kausalen Ursache der Intra-Klumpen-Korrelation können sie jedoch nicht beantworten (siehe Kapitel 6, Ausblick). Die Notwendigkeit, zwei Modelle zu betrachten, ergibt sich mindestens aus der Tatsache, dass das Beta-Binomial-Modell keinen Test auf $\rho = 0$ erlaubt, da diese Spezifikation nicht Teil des Modells ist. Dieser Aspekt wird sich in den nächsten Kapiteln im Vergleich der beiden Modelle widerspiegeln.

Zwei weitere Varianten sind zum einen die Modellierung über Quotenverhältnisse (*Odds Ratio*) und zum anderen Generalisierte Lineare Gemischte Modelle. Zum letzteren sei auf einschlägige Literatur verwiesen. Das *Odds Ratio* für Paare $(X_{ij}, X_{ij'})$ ist konstant für alle i und besitzt die Form

$$\text{Odds Ratio} = \frac{\Pr(X_{ij} = 1, X_{ij'} = 1)\Pr(X_{ij} = 0, X_{ij'} = 0)}{\Pr(X_{ij} = 1, X_{ij'} = 0)\Pr(X_{ij} = 0, X_{ij'} = 1)}.$$

Mit den aus Kapitel 2.1 bekannten Wahrscheinlichkeiten ergibt sich die gewünschte Beziehung zu ρ .

$$\begin{aligned} \text{Odds Ratio} &= \frac{(\rho\pi + (1 - \rho)\pi^2)(\rho(1 - \pi) + (1 - \rho)(1 - \pi)^2)}{(1 - \rho)^2\pi^2(1 - \pi)^2} \\ &= \frac{\rho^2\pi(1 - \pi) + (1 - \rho)\rho\pi(1 - \pi)^2 + (1 - \rho)\rho\pi^2(1 - \pi) + (1 - \rho)^2\pi^2(1 - \pi)^2}{(1 - \rho)^2\pi^2(1 - \pi)^2} \\ &= 1 + \frac{\rho}{(1 - \rho)^2} \left(\frac{1}{\pi} + \frac{1}{1 - \pi} \right). \end{aligned}$$

Modelle, die auf *Odds Ratios* basieren, werden in dieser Arbeit nicht näher betrachtet.

3 Methoden

Für die Schätzung des interessierenden Parameters ρ bietet die Literatur eine Vielzahl unterschiedlicher Methoden. Eine über alle möglichen Datenskalierungen hinweg übliche Methode ist derzeit nicht bekannt. So ist es verbreitet, kardinal skalierte Daten mit Hilfe linearer, gemischter Modelle zu beschreiben und die Intra-Klumpen-Korrelation zu bestimmen (Eldridge *et al.* (2009); Donner und Klar (2000, Kapitel 7)). Für dichotome Zielgrößen ist ein linearer Ansatz mit zufälligem Störterm nicht praktikabel.

Im Kapitel „Methoden“ werden elf verschiedene Varianten, den ICC bei binärer Zielgröße zu schätzen, eingeführt: Maximum-Likelihood (ML) und abgewandelte Likelihood-Methoden (Quasi-Likelihood (QL), Extended-Quasi-Likelihood (EQL), Pseudo-Likelihood (PL)), Pairwise-Equal-Weights (PEQ), ANOVA, Moment-Method-Based (MMB), Bayes (B), Beta-Binomial (DBB), Weighted-Empirical-Pairwise (WEP) sowie Kappa (K). Bayes- und Beta-Binomial-Schätzer sind bisher nicht zur ICC-Schätzung in Klumpen-Studien geläufig, während beispielsweise der ANOVA-Schätzer ein Standardwerkzeug geworden ist.

Alle vorgestellten Schätzmethode sind sowohl auf balancierte als auch auf unbalancierte Daten anwendbar. Weiterhin basiert jede Methode entweder auf dem Modell mit pauschaler Klumpenkorrelation bzw. auf dem Beta-Binomial-Modell oder ist modellunabhängig konstruiert. Tabelle 3.1 gibt Auskunft über die entsprechende Zuordnung der untersuchten Methoden.

Tabelle 3.1: Schätzmethode nach zu Grunde liegendem Modell

CCM	ML, QL, EQL, PL
BBM	B, DBB
modellunabhängig	PEQ, ANOVA, MMB, WEP, K

3.1 Auf dem Modell mit pauschaler Klumpenkorrelation basierende Schätzer

Im Modell mit pauschaler Klumpenkorrelation ist die Verteilung der Gesamtanzahl an Erfolgen pro Klumpen abhängig von den Parametern ρ und π bekannt (siehe Definition 2.1). Daher bieten sich Likelihood-Methoden zur Schätzung des interessierenden Parameters ρ an.

3.1.1 Maximum-Likelihood-Schätzer

Mit Hilfe der Maximum-Likelihood-Methode wird im Allgemeinen genau der Schätzwert als optimal akzeptiert, bei dem die beobachtete Stichprobe die größte Wahrscheinlichkeit besitzt (Bleymüller und Weißbach (2015, Kapitel 15.5)).

Wie in Kapitel 2.1 gesehen, ist der ML-Schätzer für den ICC, $\hat{\rho}_{ML}$, im Modell mit pauschaler Klumpenkorrelation abhängig von der konkreten Stichprobe, genauer gesagt von der konkreten Anzahl an Erfolgen in jedem einzelnen Klumpen. In der Praxis häufig anzutreffende Stichprobenkonstellationen (Anzahl der Erfolge mindestens 1 und maximal $n - 1$ für alle Klumpen) führen dazu, dass es keinen zulässigen ML-Schätzer in diesem Modell gibt. Im allgemeinen Fall ergibt sich der ML-Schätzer im CCM aus der Score-Funktion angewandt auf den konkreten Datensatz (vgl. auch Beispiel 2.4).

$$\begin{aligned} \frac{\partial l(y_1, \dots, y_n | \rho, \pi)}{\partial \rho} = & \sum_{i=1}^k \left[\frac{(1 - \pi) - (1 - \pi)^{n_i}}{\rho(1 - \pi) + (1 - \rho)(1 - \pi)^{n_i}} \cdot I_{\{y_i=0\}} - \frac{1}{(1 - \rho)} \cdot I_{\{1 \leq y_i \leq n_i - 1\}} \right. \\ & \left. + \frac{\pi - \pi^{n_i}}{\rho\pi - (1 - \rho)\pi^{n_i}} \cdot I_{\{y_i=n_i\}} \right] \end{aligned}$$

3.1.2 Quasi-Likelihood-Schätzer

Eine Schätzfunktion ist im Allgemeinen eine spezielle Stichprobenfunktion, d. h. sie überführt den Stichprobenvektor (y_1, \dots, y_n) in einen Skalar (Rinne (2008, Seite 437)). Der Quasi-Likelihood-Schätzer (QL) ergibt sich aus der Quasi-Score-Schätzfunktion, welche im Folgenden hergeleitet wird.

Sei π gegeben bzw. vorab aus den zu analysierenden Daten geschätzt. Die Varianz von Y_i sei definiert als

$$\text{Var}(Y_i) = \phi_i V(n_i \pi),$$

wobei $\phi_i = \frac{1+\rho(n_i-1)}{n_i} = \frac{VIF}{n_i}$ und $V(n_i \pi) = n_i \pi(1 - \pi)$ sind.

McCullagh und Nelder (1989, Seite 325) definieren

$$U_i(Y_i, E(Y_i), \text{Var}(Y_i)) = \frac{Y_i - n_i \pi}{\phi_i V(n_i \pi)}$$

für jede Beobachtung der Gesamtzahl an Erfolgen in Klumpen i .

U_i hängt ausschließlich von den Daten und deren erstem und zweitem Moment ab. Dennoch hat die Funktion folgende Eigenschaften mit einer Log-Likelihood-Ableitung gemeinsam:

$$\begin{aligned} E(U_i) &= 0, \\ \text{Var}(U_i) &= \frac{1}{\text{Var}(Y_i)}, \\ -E\left(\frac{\partial U_i}{\partial n_i \pi}\right) &= \frac{1}{\text{Var}(Y_i)}. \end{aligned}$$

Weiterhin gilt unter einigen sehr weichen Annahmen, dass das Integral

$$Q_i = \int_{y_i}^{n_i \pi} \frac{y_i - t}{\phi_i V(t)} dt = \int_{y_i}^{n_i \pi} \frac{y_i - t}{\phi_i V(t)} dt + \text{function of } y_i$$

in seiner Beschaffenheit einer Log-Likelihood-Funktion ähnelt (vergleiche Wedderburn (1974) sowie McCullagh und Nelder (1989, Seite 325 ff.) für eine Berechnung der Quasi-Score-Funktion als auch einem Beweis der Unverzerrtheit und asymptotischer Normalität des erhaltenen Schätzers).

Definition 3.1 (*Log-Quasi-Likelihood*)

Sei y der Vektor der Realisierungen von $Y = (Y_1, \dots, Y_k)$ und $\mu := (n_1 \pi, \dots, n_k \pi)$ der Vektor der Erwartungswerte von Y . Da die Komponenten von Y unabhängig sind, heißt

$$l_{QL}(y, \mu) := \sum_{i=1}^k Q_i = \sum_{i=1}^k \int_{y_i}^{n_i \pi} \frac{y_i - t}{\phi_i V(t)} dt$$

die Log-Quasi-Likelihood für $V(\mu)$ basierend auf den Daten y .

Die Log-Quasi-Likelihood für $\phi_i = \frac{1+\rho(n_i-1)}{n_i}$ und $V(n_i\pi) = n_i\pi(1-\pi)$ ist

$$l_{QL}(y_1, \dots, y_k | \rho, \pi) = \sum_{i=1}^k \frac{\Phi_i}{VIF},$$

wobei

$$\Phi_i = n_i \cdot \log\left(\frac{n_i - n_i\pi}{n_i - y_i}\right) + y_i \cdot \log\left(\frac{(n_i - y_i)n_i\pi}{(n_i - n_i\pi)y_i}\right).$$

Die o. g. Eigenschaften von U_i können nur für die partiellen Ableitungen nach μ_i gezeigt werden (nicht für Ableitungen nach ρ und/oder $Var(Y_i)$). Deshalb ist die Anwendbarkeit der vorgestellten, einfachen Quasi-Likelihood-Methode zur Konstruktion von Schätzern für ρ an dieser Stelle zweifelhaft. In Kapitel 5.2 stellt sich heraus, dass die Antwort auf diese Frage im Einzelfall negativ beantwortet werden muss.

3.1.3 Extended-Quasi-Likelihood-Schätzer

Extended-Quasi-Likelihood ist ursprünglich als Erweiterung der Quasi-Likelihood-Methode zum formellen Vergleich verschiedener Regressionsmodelle mit unterschiedlicher Varianzfunktion entwickelt worden (McCullagh und Nelder (1989, Kapitel 9.6)). Daher erfüllt eine Extended-Quasi-Likelihood-Funktion zusätzlich zu den Eigenschaften einer Quasi-Likelihood-Funktion auch Eigenschaften hinsichtlich der Varianz $Var(Y_i)$.

Die Extended-(Log)-Quasi-Likelihood-Funktion für $Var(Y_i) = \underbrace{1 + (n_i - 1)\rho}_{VIF} n_i\pi(1 - \pi)$

ist definiert als

$$l_{EQL}(y_1, \dots, y_k | \rho, \pi) = -\frac{1}{2} \sum_{i=1}^k \left(\log(VIF) + \frac{\tilde{\Phi}_i}{VIF} \right),$$

mit

$$\tilde{\Phi}_i = 2 \left(y_i \log\left(\frac{y_i}{n_i\pi}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - n_i\pi}\right) \right)$$

(siehe Ridout *et al.* (1999)). Dieser Ausdruck erfüllt die Eigenschaften einer (herkömmlichen) Log-Likelihood für VIF-Ableitungen (siehe McCullagh und Nelder (1989, Kapitel 9.6)), wobei $VIF = Var(Y_i)/V(n_i\pi)$ im Wesentlichen der Varianz von Y_i entspricht.

Die partiellen Ableitungen von l_{EQL} nach π und ρ sind

$$\begin{aligned}\frac{\partial l_{EQL}}{\partial \pi} &= \sum_{i=1}^k -\frac{1}{VIF} \left(\frac{-y_i n_i \pi}{y_i} \frac{y_i n_i}{(n_i \pi)^2} + \frac{(n_i - y_i)(n_i - n_i \pi)}{(n_i - y_i)} \frac{(n_i - y_i) n_i}{(n_i - n_i \pi)^2} \right) \\ &= \sum_{i=1}^k \frac{1}{VIF} \left(\frac{y_i}{n_i} - \frac{n_i - y_i}{1 - \pi} \right) \\ &= \sum_{i=1}^k \frac{y_i - n_i \pi}{VIF \cdot \pi(1 - \pi)}\end{aligned}$$

und

$$\begin{aligned}\frac{\partial l_{EQL}}{\partial \rho} &= \sum_{i=1}^k \frac{\Phi_i(n_i - 1)}{VIF^2} - \frac{n_i - 1}{VIF} \\ &= \sum_{i=1}^k (n_i - 1) \left(\frac{\Phi_i - VIF}{VIF^2} \right).\end{aligned}$$

Ridout *et al.* (1999) dokumentieren, dass der Schätzer $\hat{\rho}_{EQL}$ angemessen effizient ist, wenn ρ klein ist, wie in diesem Beispiel. Konsistent ist der Extended-Quasi-Likelihood-Schätzer dagegen nicht (siehe Davidian und Carroll (1988)).

3.1.4 Pseudo-Likelihood-Schätzer

Die Pseudo-Likelihood-Methode (vgl. Carroll und Ruppert (1988, Kapitel 3.2)) bietet eine weitere Möglichkeit der Approximation an die (herkömmliche) Likelihood-Funktion eines gegebenen Datensatzes.

Die Pseudo-(Log)-Likelihood-Funktion im Modell mit pauschaler Klumpenkorrelation lautet

$$l_{PL}(y_1, \dots, y_k | \rho, \pi) = -\frac{1}{2} \sum_{i=1}^k \left(\log(VIF) + \frac{\bar{\Phi}_i}{VIF} \right),$$

mit

$$\bar{\Phi}_i = \frac{(y_i - n_i \pi)^2}{n_i \pi (1 - \pi)}$$

(siehe [Ridout et al. \(1999\)](#)). Sie gleicht somit formell der Extended-(Log)-Quasi-Likelihood-Funktion, verwendet aber mit $\bar{\Phi}_i$ ein anderes Abstandsmaß. Der resultierende Schätzer für die Intra-Klumpen-Korrelation, $\hat{\rho}_{PL}$, ist unverzerrt, da die Schätzfunktion

$$\sum_{i=1}^k (n_i - 1) \left(\frac{\bar{\Phi}_i - VIF}{VIF^2} \right) = 0$$

den Erwartungswert 0 hat. Dafür ist zu zeigen, dass $E(\bar{\Phi}_i) = VIF$ ist:

$$E(\bar{\Phi}_i) = E\left(\frac{(Y_i - n_i\pi)^2}{n_i\pi(1-\pi)}\right) = \frac{1}{n_i\pi(1-\pi)} E((Y_i - n_i\pi)^2)$$

Mit dem Verschiebungssatz für Erwartungswerte gilt weiter

$$\begin{aligned} \frac{1}{n_i\pi(1-\pi)} E((Y_i - n_i\pi)^2) &= \frac{1}{n_i\pi(1-\pi)} \left(\text{Var}(Y_i - n_i\pi) + E(Y_i - n_i\pi)^2 \right) \\ &= \frac{1}{n_i\pi(1-\pi)} \left(\text{Var}(Y_i) + E(Y_i - n_i\pi)^2 \right) \end{aligned}$$

(mit Gleichung 2.3)

$$\begin{aligned} &= \frac{1}{n_i\pi(1-\pi)} \left(n_i\pi(1-\pi)(1 + (n_i - 1)\rho) + (n_i\pi - n_i\pi)^2 \right) \\ &= 1 + (n_i - 1)\rho \\ &= VIF. \end{aligned}$$

□

3.2 Auf dem Beta-Binomial-Modell basierende Schätzer

Im Folgenden werden zwei Methoden (Beta-Binomial und Bayes) zur Schätzung der Intra-Klumpen-Korrelation vorgestellt, deren Herleitung auf dem Beta-Binomial-Modell beruht. Im Gegensatz zu o. g. Methoden ist die grundsätzliche Idee in diesem Abschnitt die Schätzung der Parameter α und β mit anschließendem Schluss auf den interessierenden Parameter ρ .

3.2.1 Direkter Beta-Binomial-Modell-Schätzer

Eine Möglichkeit α und β direkt aus der Verteilung der Daten zu schätzen, geht auf Dupuis (1995, Seite 768) zurück. Der Autor schlägt vor, zunächst den Parameter α zu schätzen und β mit Hilfe der Relationen

$$\hat{\pi}_{ML} = \frac{\sum_{i=1}^k y_i}{N} = E(\Pi_i) \underset{\text{Beta-Verteilung}}{=} \frac{\alpha}{\alpha + \beta}$$

zu berechnen. Denn es gilt

$$\beta = \frac{\alpha - \alpha \hat{\pi}_{ML}}{\hat{\pi}_{ML}}.$$

Die Intra-Klumpen-Korrelation wird schließlich bestimmt durch

$$\hat{\rho}_{DBB} = \text{Kor}(X_{ij}, X_{ij'}) = \frac{1}{\alpha + \beta + 1}$$

(Lemma 2.5). Das Verfahren im Einzelnen:

Ablauf zur Schätzung von (α, β) :

1. Berechnung von $\hat{\pi}_{ML} = \frac{\sum_{i=1}^k y_i}{N}$,
2. Berechnung von $\hat{\pi}_{i,ML} = \frac{y_i}{n_i}$, $i = 1, \dots, k$,
3. Sortierung der π_i der Größe nach und Berechnung des (empirischen) 95% Vertrauensintervalls $[\pi_{k \times 0,025}, \pi_{k \times 0,975}]$, dessen Grenzen genau die Werte π_i sind, deren Index $[0, 025k]$ bzw. $[0, 975k]$ ist,
4. Für α von 0,1 bis 20 in Schritten 0,1
 - a) Berechnung von $\beta = \frac{\alpha - \alpha \hat{\pi}_{ML}}{\hat{\pi}_{ML}}$,
 - b) Berechnung des mittleren 95% Vertrauensintervalls $[Q_{0,025}, Q_{0,975}]$ der Beta-Verteilung,
 - c) Ermittlung von (α, β) , die den Abstand $(Q_{0,025} - \pi_{k \times 0,025})^2 + (Q_{0,975} - \pi_{k \times 0,975})^2$ minimieren.

3.2.2 Bayes-Schätzer

Der Bayes-Ansatz beruht darauf, zwei gleichwertige Bestimmungsstücke zur Parameterschätzung zu nutzen - ein Vorwissen über die Verteilung des Parameters (A-priori-Verteilung) und das Stichprobenergebnis. Dabei wird der Parameter wie eine Zufallsvariable behandelt. Als Ergebnis steht die Verteilung dieses Parameters korrigiert um die Daten (A-posteriori-Verteilung).

Im Beta-Binomial-Modell ist Π der zu schätzende Parameter, dessen A-priori-Verteilung die Betaverteilung mit den Parametern α und β ist,

$$\Pi \sim \text{Beta}(\alpha, \beta).$$

Der tatsächlich gesuchte Schätzer für ρ ergibt sich in diesem Verfahren, indem die Beziehung

$$\rho = \text{Korr}(X_{ij}, X_{ij'}) = (\alpha + \beta + 1)^{-1}$$

(vgl. Lemma 2.5) auf die A-posteriori-Parameter von Π nach einer Bayes-Iteration angewandt wird. Der Satz von Bayes weist auf, wie diese Parameter berechnet werden (vgl. Bortz und Döring (2002, Kapitel 7.2.5)).

Satz 3.2 (*Satz von Bayes*)

Seien A und B zwei Ereignisse. Es gilt

$$Pr(\underbrace{A}_{\Pi} \mid \underbrace{B}_{\text{Daten}}) = \frac{\overbrace{Pr(B|A)}^{L(\Pi)} \cdot \overbrace{Pr(A)}^{Pr(\Pi)}}{Pr(B)}.$$

Beweis: Siehe z. B. Hesse (2003, Seite 52). □

Im Satz von Bayes wird das Ereignis A nun als der gesuchte Parameter Π und B als Stichprobenergebnis interpretiert. Somit lautet der Satz für die stetige Zufallsvariable Π (und ohne den Normierungsfaktor $Pr(B)$)

$$Pr(\Pi | \text{Daten}) \propto Pr(\Pi) \cdot \underbrace{L(\Pi)}_{\prod_{i=1}^k Pr_{\pi}(y_i)}.$$

Die A-posteriori-Verteilung von Π ist die Dichte von Π multipliziert mit der Likelihood-Funktion des Stichprobenergebnisses y_1, \dots, y_k bei gegebener Verteilung von Π .

Die Dichte der Betaverteilung auf $[0, 1]$ ist – bis auf einen Normierungsfaktor – gegeben durch

$$Pr(\pi) = \pi^{\alpha-1} (1 - \pi)^{\beta-1}.$$

Die bedingte Wahrscheinlichkeitsfunktion der Gesamtanzahl an bedingt unabhängigen Erfolgen in Klumpen i , Y_i , ist (Binomialverteilung)

$$Pr(Y_i = y_i) = \binom{n_i}{y_i} \pi^{y_i} (1 - \pi)^{n_i - y_i}, \quad y_i = 0, \dots, n_i.$$

Für die Likelihood-Funktion ergibt sich somit

$$\begin{aligned} L(y_1, \dots, y_n | \pi) &= \prod_{i=1..k} \binom{n_i}{y_i} \pi^{y_i} (1 - \pi)^{n_i - y_i} \\ &\propto \pi^{\sum y_i} (1 - \pi)^{\sum n_i - \sum y_i}. \end{aligned}$$

Damit ist die A-posteriori-Dichte von Π

$$\begin{aligned} Pr(\Pi | \text{Daten}) &= Pr(\Pi | y_1, \dots, y_n) = \pi^{\alpha-1} (1 - \pi)^{\beta-1} \pi^{\sum y_i} (1 - \pi)^{\sum n_i - \sum y_i} \\ &= \pi^{\sum y_i + \alpha - 1} (1 - \pi)^{\sum n_i - \sum y_i + \beta - 1}. \end{aligned}$$

Diese Darstellung lässt einen leichten Schluss auf die A-posteriori-Parameter von Π zu, denn Π ist nun nach einer Bayes-Iteration betaverteilt:

$$\Pi^{post} \sim \text{Beta}\left(\underbrace{\sum y_i}_Y + \alpha, \underbrace{\sum n_i}_N - \sum y_i + \beta\right).$$

Für die um die Stichprobe y_1, \dots, y_k korrigierte Intra-Klumpen-Korrelation bedeutet dies

$$\begin{aligned} \hat{\rho}_B = \text{Korr}^{post}(X_{ij}, X_{ij'}) &= (\alpha + Y + \beta + N - Y + 1)^{-1} \\ &= (N + \alpha + \beta + 1)^{-1}. \end{aligned}$$

Der Bayes-Schätzer ist daher für große N fast unabhängig von der Wahl der Hyperparameter.

Nachrichtlich sind im Anhang [A.2.3](#) (Momente der A-posteriori-Verteilung von Π) weitere interessante Eigenschaften der A-posteriori-Verteilung von Π aufgeführt.

3.3 Modellunabhängige Schätzer

Ausgangspunkt der bisher vorgestellten Schätzer für die Intra-Klumpen-Korrelation waren Modelle, die entweder direkt oder indirekt die Existenz Intra-Klumpen-korrelierter Daten rechtfertigen.

Die folgenden Schätzmethode rücken von dieser Modellabhängigkeit ab. So basieren einige stichprobentheoretische Schätzer (K, WEP, PEQ) auf der Pearson'schen Definition der Korrelation zweier Merkmale

$$\rho(X, Y) = \frac{Kov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

und zielen somit ausschließlich auf die jeweils paarweisen Zusammenhänge innerhalb der Daten ab. Es stellt sich die Frage der Anwendbarkeit der Pearson-Korrelation auf dichotome Daten.

Lemma 3.3 (*Anwendbarkeit*)

Der Pearson-Korrelationskoeffizient ist auf dichotome Daten anwendbar.

Beweis: Der Pearson-Korrelationskoeffizient (ρ) ist ein Maß für den Grad des Zusammenhanges zwischen zwei kardinalskalierten Merkmalen. Für dichotome Daten existiert das Konzept des Phi-Koeffizienten (Φ). Es ist zu zeigen, dass ρ gleich Φ ist.

Seien X und Y zwei binäre Zufallsvariablen mit $P(X = 1) = P(Y = 1) = \pi$ und jeweils positiver Varianz. Die weiteren Wahrscheinlichkeiten sind in folgender Vierfeldertafel eingetragen.

X	Y		Σ
	0	1	
0	p_{11}	p_{12}	$p_{1\cdot} = (1 - \pi)$
1	p_{21}	p_{22}	$p_{2\cdot} = \pi$
Σ	$p_{\cdot 1} = (1 - \pi)$	$p_{\cdot 2} = \pi$	

Es gilt nach Definition:

$$\begin{aligned}
 \rho &= \frac{Kov(X, Y)}{\sqrt{Var(X)Var(Y)}} \\
 &= \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} \\
 &= \frac{p_{22} - (p_{21} + p_{22})(p_{12} + p_{22})}{\sqrt{p_1 \cdot p_2 \cdot p_1 \cdot p_2}} \\
 &= \frac{p_{22} - p_{21}p_{12} - p_{21}p_{22} - p_{22}p_{12} - p_{22}p_{22}}{\sqrt{p_1 \cdot p_2 \cdot p_1 \cdot p_2}} \\
 &= \frac{p_{22}(1 - p_{21} - p_{12} - p_{22}) - p_{21}p_{12}}{\sqrt{p_1 \cdot p_2 \cdot p_1 \cdot p_2}} \\
 &= \frac{p_{22}p_{11} - p_{21}p_{12}}{\sqrt{p_1 \cdot p_2 \cdot p_1 \cdot p_2}} = \Phi
 \end{aligned}$$

□

3.3.1 Kappa-Schätzer

Der Kappa-Schätzer für die Intra-Klumpen-Korrelation verwendet ebendiese Eigenschaft des Pearson-Korrelationskoeffizienten.

Nach Annahme gilt für zwei Elemente X_{ij} und $X_{ij'}$

$$Kor(X_{ij}, X_{ij'}) = \rho = \frac{Kov(X_{ij}, X_{ij'})}{\sqrt{Var(X_{ij})Var(X_{ij'})}}$$

Mit dem Verschiebungssatz für Kovarianzen

$$Kov(X_{ij}, X_{ij'}) = E(X_{ij}X_{ij'}) - E(X_{ij})E(X_{ij'})$$

und den Beziehungen $E(X_{ij}) = \pi$ sowie $E(X_{ij}X_{ij'}) = Pr(X_{ij} = 1, X_{ij'} = 1)$ besteht ebenso

$$\rho = \frac{Pr(X_{ij} = 1, X_{ij'} = 1) - \pi^2}{\pi(1 - \pi)}$$

Zusammen mit den Gleichungen (2.4) und (2.5) kann dieser Ausdruck umgestellt werden und es ist

$$\rho = \frac{Pr(X_{ij} = X_{ij'}) - Pr(X_{ij} = X_{i'j'})}{1 - Pr(X_{ij} = X_{i'j'})}. \quad (3.1)$$

In dieser Form kann ρ als die Übereinstimmung in Wahrscheinlichkeit zwischen zwei dichotomen Zufallsvariablen interpretiert werden.¹

In Gleichung (3.1) wird $Pr(X_{ij} = X_{ij'}) =: Pr_{observed}$ nun als die beobachtete Übereinstimmung zwischen Elementen innerhalb eines Klumpens interpretiert.² Dabei werden Abhängigkeiten zwischen den Elementen berücksichtigt. $Pr(X_{ij} = X_{i'j'})$ wird gedeutet als Pr_{chance} , die hypothetische Wahrscheinlichkeit der Übereinstimmung zwischen zwei Elementen in verschiedenen Klumpen (keine Abhängigkeiten zwischen Elementen).³ Der Term $1 - Pr(X_{ij} = X_{i'j'})$ ist dann der „maximale Abstand zweier Elemente“ (in Wahrscheinlichkeit) (vgl. Eldridge *et al.* (2009)).

Der Kappa-Schätzer für die Intra-Klumpen-Korrelation lautet

$$\hat{\rho}_K = \frac{Pr_{observed} - Pr_{chance}}{1 - Pr_{chance}}.$$

Dabei ist

$$Pr_{observed} = \frac{1}{k} \sum_{i=1}^k P_i$$

mit $P_i = \frac{Y_i^2 + (n_i - Y_i)^2 - n_i}{n_i(n_i - 1)}$ die Anzahl aller möglichen Elemente-Paare in Klumpen i mit gleichem Ausgang geteilt durch die Anzahl aller möglichen Paare in Klumpen i und

$$\begin{aligned} Pr_{chance} &= Pr(X_{ij} = 1, X_{i'j'} = 1) + Pr(X_{ij} = 0, X_{i'j'} = 0) \\ &= \pi^2 + (1 - \pi)^2. \end{aligned}$$

Der Kappa-Schätzer berücksichtigt also die Anzahl der Paare „Erfolg“ ($X_{ij} = X_{ij'} = 1$) und „Mißerfolg“ ($X_{ij} = X_{ij'} = 0$). Bei der Paarbildung wird die Reihenfolge berücksichtigt (Variationen). Die Schätzung von π kann über die Maximum-Likelihood-Methode erfolgen und ist meist unkritisch. Ridout *et al.* (1999) zeigen, dass die Effizienz von $\hat{\rho}_K$ gut ist, wenn für ρ ein kleiner Wert erwartet wird.

¹Vgl. Fleiss' Kappa, wonach der Kappa-Schätzer benannt ist.

²Zwischen Bewertern bei Fleiss' Kappa.

³Rein zufällige Übereinstimmung zweier Bewertungen bei Fleiss' Kappa.

3.3.2 Weighted-Empirical-Pairwise-Schätzer

Analog zum Kappa-Schätzer ist der Weighted-Empirical-Pairwise-Schätzer vom (Pearsonschen-) Korrelationskoeffizienten ausgehend entwickelt:

$$\begin{aligned} \text{Kor}(X_{ij}, X_{ij'}) &= \rho_{WEP} = \frac{\text{Kov}(X_{ij}, X_{ij'})}{\sqrt{\text{Var}(X_{ij})\text{Var}(X_{ij'})}} \\ &= \frac{\text{Pr}(X_{ij} = 1, X_{ij'} = 1) - \pi^2}{\pi(1 - \pi)}. \end{aligned}$$

π wird durch $\hat{\pi}_{ML} = \frac{Y}{N}$ ersetzt. Die Wahrscheinlichkeit, dass zwei unterschiedliche Elemente eines Klumpens i gleichzeitig den Ausgang „Erfolg“ haben, $\text{Pr}(X_{ij} = 1, X_{ij'} = 1)$, ist nun empirisch zu schätzen:

$$\text{Pr}(X_{ij} = 1, X_{ij'} = 1) \approx \text{emp } E(X_{ij}X_{ij'}) := \frac{\sum_{i=1}^k P_i}{\sum_{i=1}^k N_i}.$$

Dabei ist $P_i = \binom{y_i}{2}$ die Anzahl aller möglichen Elemente-Paare in Klumpen i mit Ausgang „Erfolg“ ($X_{ij} = X_{ij'} = 1$) und $N_i = \binom{n_i}{2}$ die Anzahl aller möglichen Paare in Klumpen i . Im Gegensatz zum Kappa-Schätzer werden hier nur die Paare mit Ausgang „Erfolg“ berücksichtigt. Bei der Paarbildung bleibt die Reihenfolge unberücksichtigt (Kombinationen). Das hat zur Folge, dass große Klumpen stärker gewichtet werden als bei der Methode nach Kappa.

3.3.3 Pairwise-Equal-Weights-Schätzer

Mit dem Pairwise-Equal-Weights-Schätzer reiht sich eine weitere Methode in die Klasse stichprobentheoretischer Schätzer ein. Das Spezielle an diesem Verfahren ist, dass die Abhängigkeiten innerhalb der Stichprobe grob über die paarweise Korrelation aller möglichen Paare gemessen wird.

Ridout *et al.* (1999) geben in einer Variante des Schätzers jedem (Pearsonschen-) Korrelationskoeffizienten zwischen Elementen-Paaren innerhalb der Klumpen das gleiche Gewicht. Der resultierende Schätzer ist

$$\hat{\rho}_{PEQ} = \frac{1}{\mu_{PEQ}(1 - \mu_{PEQ})} \left(\frac{\sum_{i=1}^k Y_i(Y_i - 1)}{\sum_{i=1}^k n_i(n_i - 1)} - \mu_{PEQ}^2 \right),$$

wobei

$$\mu_{PEQ} = \frac{\sum_{i=1}^k Y_i(n_i - 1)}{\sum_{i=1}^k n_i(n_i - 1)}.$$

Der Arbeit von [Ridout et al. \(1999\)](#) zufolge ist die Leistung dieses Schätzers ebenfalls gut, wenn ρ klein ist.

3.3.4 ANOVA-Schätzer

Die wohl älteste Methode, die Intra-Klumpen-Korrelation zu schätzen, findet ihre erste relevante Erwähnung in [Kish \(1965, Kapitel 5\)](#). Der Autor bezeichnet die *intraclass correlation* schlicht mit „roh“ und verwendet zur Schätzung Techniken der Varianzanalyse. Das Verfahren ist in vielen – gerade anwendungsorientierten – Problemen (z.B. klinischen Studien) zum Standard geworden, da es mit relativ einfacher Rechen-Ausstattung durchzuführen ist. [Ganninger et al. \(2007\)](#) definieren den ANOVA-Schätzer als

$$\hat{\rho}_{ANOVA} = \frac{MSB - MSW}{MSB + (K - 1)MSW}$$

mit

$$\begin{aligned} MSB &= \frac{SQA}{k - 1} \\ MSW &= \frac{SQR}{N - k} \\ K &= \frac{N - \frac{1}{N} \sum_{i=1}^k n_i^2}{k - 1}. \end{aligned}$$

Darin ist SQA die Summe der quadratischen Abweichungen zwischen den Klumpen $\sum_{i=1}^k n_i (\bar{X}_i - \bar{X}_{..})^2$ und SQR die Summe der quadratischen Abweichungen innerhalb der Klumpen $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$. Die Autoren zeigen, dass der Schätzer effizient, konsistent und erwartungstreu ist.

3.3.5 Moment-Method-Based-Schätzer

Im Folgenden wird ein Verfahren entwickelt, welches den Ansatz von [Cox und Snell \(1989, Kapitel 3.2.4\)](#) weiterentwickelt und auf der Momentenmethode basiert. Die Autoren schlagen vor, den Chi-Quadrat-Unabhängigkeitstest zu invertieren. Die Hypothesen lauten

$$\begin{aligned} H_0 : \rho &= 0 && \text{Die Daten } X_{ij} \text{ sind unabhängig} \rightarrow \text{Verteilung von } Y_i \text{ binomial,} \\ H_1 : \rho &\neq 0. \end{aligned}$$

Lemma 3.4 (*Goodness-of-fit-Teststatistik*)

Es sei $T_{H_0} := \sum_{i=1}^k \frac{(Y_i - n_i \pi)^2}{n_i \pi (1 - \pi)}$. Dann gilt für den Erwartungswert von T_{H_0} bezüglich ρ :

$$E_\rho T_{H_0} = k \cdot VIF = k \cdot (1 + (n_i - 1)\rho).$$

Beweis:

Die Statistik T_{H_0} ist unter H_0 Chi-Quadrat-verteilt mit k Freiheitsgraden (Cox und Snell, 1989, Kapitel 3.2.4)

$$T_{H_0} \sim \chi^2(k).$$

Dann gilt im Falle balancierter Klumpen, d. h. $n_i = n, \forall i$

$$\begin{aligned} E_\rho T_{H_0} &= \sum_{i=1}^k \frac{E_\rho(Y_i - n\pi)^2}{n\pi(1-\pi)} \\ &= \sum_{i=1}^k \frac{\overbrace{\text{Var}(Y_i)}^{\text{beobachtete Varianz}}}{n\pi(1-\pi)} \\ &= \sum_{i=1}^k \frac{n\pi(1-\pi)(1+(n-1)\rho)}{n\pi(1-\pi)} \\ &= k \cdot VIF. \end{aligned}$$

Damit ist gezeigt, dass $\frac{T_{H_0}}{k}$ ein Schätzer für die Varianzinflation VIF ist, denn

$$E\left(\frac{T_{H_0}}{k}\right) = VIF.$$

Im unbalancierten Fall ist der Schätzer implizit gegeben durch das Lösen der Gleichung

$$\sum_{i=1}^k \frac{(Y_i - n_i \pi)^2}{n_i \pi (1 - \pi) (1 + (n_i - 1)\rho)} = k,$$

da analog zum balancierten Fall gilt

$$\frac{T_{H_0}}{VIF} \approx k.$$

□

3.4 Konfidenzintervall für die Intra-Klumpen-Korrelation

Lui *et al.* (1996) schlagen einen Intervall-Schätzer für den ICC im Beta-Binomial-Modell vor. Ausgehend von der aus Abschnitt 2.2 bekannten Beziehung $Kor(X_{ij}, X_{ij'}) = \rho = 1/(\alpha + \beta + 1)$, $i \neq i'$, d. h.

$$\alpha + \beta = \rho^{-1} - 1, \quad (3.2)$$

definieren sie im Falle fixer Klumpengrößen n zwei Zufallsvariablen $B_i = n(Y_i/n - \pi)^2$ und $W_i = \sum_j (X_{ij} - Y_i/n)^2 / (n - 1)$. Dabei ist $\pi = \alpha / (\alpha + \beta)$ der Erwartungswert der betaverteilten Erfolgswahrscheinlichkeiten. Y_i folgt der Beta-Binomialverteilung mit Parametern (α, β, n) und Erwartungswert $n\alpha / (\alpha + \beta)$. Weiter sei $\theta := E(B_i) / E(W_i)$ der Quotient der Erwartungswerte von B_i und W_i und ist gegeben durch $\theta = 1 + n / (\alpha + \beta)$ (siehe Lui *et al.*, 1996, Appendix A). Die Zufallsvariable

$$Z_i := B_i - \theta W_i$$

hat nun den Erwartungswert $E(Z_i) = 0$ und eine konstante Varianz $Var(Z_i)$. In gleicher Weise ist eine weitere Zufallsvariable Z_i^* definiert als

$$Z_i^* = B_i^* - \theta W_i,$$

wobei $B_i^* = n(Y_i/n - (\sum Y_i/N)^2)$ ist. Lui *et al.* (1996) zeigen, dass über den Zentralen Grenzwertsatz und das Slutsky-Theorem⁴ neben $\sqrt{k} \sum_{i=0}^k Z_i/k$ auch $\sqrt{k} \sum_{i=0}^k Z_i^*/k$ asymptotisch einer Normalverteilung mit Erwartungswert 0 und konstanter Varianz $Var(Z_i)$ folgt. Diese Erkenntnis ermöglicht die Berechnung eines Konfidenzintervalles für den ICC, da durch

$$\sum_{i=0}^k Z_i^*/k = ((k-1)/k) \cdot MQA - \theta \cdot MQR$$

die Zufallsvariable

$$((k-1)/k) \cdot MQA - \theta \cdot MQR$$

ebenfalls normalverteilt ist und nach deren Normierung und Erhebung ins Quadrat gilt:

$$\frac{\left(((k-1)/k) \cdot MQA - \theta \cdot MQR \right)^2}{\left(var(Z)/k \right)} \leq Z_{\alpha/2}^2. \quad (3.3)$$

⁴Der Zentrale Grenzwertsatz und das Slutsky-Theorem sind Konvergenz-Sätze der Wahrscheinlichkeitstheorie.

Dabei sind MQR die mittlere Abweichungsquadratsumme des Restes (innerhalb der Klumpen) und MQA die mittlere Abweichungsquadratsumme zwischen den Klumpen. Z_a entspricht dem Perzentil $100 \cdot a$ der Standardnormalverteilung.

Die untere/obere Grenze für ein Konfidenzintervall von ρ ergibt sich nun durch Umstellen und Auflösen nach θ bzw. $\rho (= (\theta - 1)/(n + \theta - 1))$ und Fieller's Theorem⁵ als $(\theta_u - 1)/(\theta_u + n - 1)$ bzw. $(\theta_o - 1)/(\theta_o + n - 1)$ für den balancierten Fall gleicher Klumpengrößen. Dabei sind θ_u und θ_o als Lösung der quadratischen Gleichung 3.3 modifizierte Quantile der Normalverteilung (siehe Lui *et al.*, 1996, Appendix B).

Im Falle unbalancierter Klumpen ergeben sich die Grenzen:

$$\begin{aligned} \text{untere Intervallgrenze: } & (\theta_u - 1)/(\theta_u + \bar{n} - 1), \\ \text{obere Intervallgrenze: } & (\theta_o - 1)/(\theta_o + \bar{n} - 1), \end{aligned}$$

mit $\bar{n} = N/k$ die durchschnittliche Klumpengröße (siehe Lui *et al.*, 1996).

Zusammenfassung

Für die Analyse von Querschnittsabhängigkeiten in Datensätzen mit Klumpen-Design wurden in diesem Kapitel zehn Methoden vorgestellt, die entweder auf dem Modell mit pauschaler Klumpenkorrelation oder dem Beta-Binomial-Modell basieren oder modellfrei konstruiert sind. Sie sollen in den kommenden Kapiteln für die weitere Studie der Intra-Klumpen-Korrelation der in Kapitel 5.1 vorgestellten Daten genutzt werden. Die Ergebnisse dafür sowie die Berechnungen der Konfidenzintervalle befinden sich in Kapitel 5.

Gerade bei den stichprobentheoretischen Schätzmethoden (K, WEP, PEQ), die darauf abstellen, die Wahrscheinlichkeit $P(X_{ij} = 1, X_{ij'} = 1)$ zu schätzen, gibt es Spielraum für weitere Ansätze, diese Wahrscheinlichkeit zu schätzen. Die vorgestellten Varianten sind jedoch einfach zu implementieren und versprechen eine hohe Genauigkeit, wie in Kapitel 4 gezeigt wird.

⁵Satz zur Berechnung von Konfidenzintervallen von Quotienten zweier Mittelwerte (siehe Fieller (1954)).

4 Simulationen

In diesem Abschnitt werden die in Kapitel 3 vorgestellten Methoden in vergleichenden Simulationen hinsichtlich ihrer Güte getestet, bevor in Kapitel 5 die exakte Intra-Klumpen-Korrelation für reale Daten berechnet wird. Geeignete Qualitätskriterien sind Erwartungstreue und Standardfehler. Die Berechnungen erfolgen mit Hilfe der Statistik-Software SAS. Die entsprechenden Quellcodes befinden sich im Anhang A.3.1.

Für vergleichbare Simulations-Ergebnisse werden zunächst binäre Intra-Klumpen-korrelierte Daten mit ein und denselben Bedingungen erzeugt (Schritt 1) und anschließend deren Klumpenkorrelation berechnet (Schritt 2).

Schritt 1 stützt sich zum einen auf das Modell mit pauschaler Klumpenkorrelation und zum anderen auf das Beta-Binomial-Modell, so dass zwei Sätze von Zufallsdaten zur Verfügung stehen. Die Ausgangsbedingungen für beide Teil-Simulationen sind:

- 170 Klumpen mit Klumpengrößen zwischen 10 und 107 (durchschnittliche Klumpengröße 47), Grundgesamtheit beträgt 7.978,
- Einhaltung einer vorgegebenen Erfolgswahrscheinlichkeit von $\pi = 0,38$ (bzw. $E(\pi) = 0,38$ beim Beta-Binomial-Modell),
- Einhaltung einer vorgegebenen Korrelation von $\rho = 0,03$ innerhalb der Klumpen,
- Erzeugung von je 10.000 Datensätzen.

Die vorgegebene Korrelation wird im Modell mit pauschaler Klumpenkorrelation direkt über die Verteilungsfunktion sichergestellt. Im Beta-Binomial-Modell bedeutet sie für die Parameter (α, β) das Wertepaar $(1843/150, 3007/150)$.

Im 2. Schritt wird jeweils die Intra-Klumpen-Korrelation mit den Methoden K, WEP, PEQ, ANOVA, MMB (modellunabhängig), DBB (auf Beta-Binomial Modell basierend) und ML, EQL, PL (auf CCM basierend) geschätzt. Der Bayes-Schätzer (B) findet keine Anwendung, da seine Berechnung bis auf die Parameter α , β und N identisch ist. Die Quasi-Likelihood-Methode (QL) liefert keine plausiblen Ergebnisse und ist deshalb in den Ergebnissen ebenfalls nicht aufgeführt. Für alle anderen Schätz-Methoden fasst Tabelle 4.1 die Mittelwerte der Punktschätzer und die Summen der quadratischen Abweichungen der Punktschätzer zum vorgegebenen ρ zusammen.

Die auf dem CCM basierenden Schätzer haben bei der Anwendung auf Daten, die im Beta-Binomial-Modell erzeugt sind, generell eine kleinere Standardabweichung als im

Tabelle 4.1: Simulierter Schätz-Mittelwert (\pm Standardfehler); unbalancierte Klumpen; Anzahl Simulationsläufe: 10.000; Anzahl Klumpen: 170; Grundgesamtheit: 7.978; Rho: 0,03

Schätzer	Beta-Binomial-Modell	Modell mit pauschaler Klumpenkorrelation
exakt	0,0300	0,0300
ML	0	0
EQL	0,0314 \pm 0,006	0,0491 \pm 0,026
PL	0,0996 \pm 0,071	0,1108 \pm 0,088
DBB	0,0305 \pm 0,004	0,1000 \pm 0,134
WEP	0,0308 \pm 0,032	0,0285 \pm 0,020
K	0,0300 \pm 0,010	0,0303 \pm 0,015
PEQ	0,0297 \pm 0,006	0,0456 \pm 0,023
ANOVA	0,0299 \pm 0,006	0,0381 \pm 0,018
MMB	0,0130 \pm 0,022	0,0152 \pm 0,024

eigenen Modell und liegen näher am exakten Wert. Die Schätzer DBB und MMB haben ebenfalls gute Eigenschaften bei BBM-Daten, wobei der DBB im CCM den wahren Wert um den Faktor 3 verfehlt und zudem die mit Abstand höchste Varianz aufweist.

Alle modellunabhängigen Schätzer weichen im BBM nur minimal vom wahren Wert ab und variieren zudem nur geringfügig. Im CCM dagegen liegt der Schätz-Mittelwert zum Teil deutlich entfernt vom wahren Wert (z.B. PEQ). Die Standardabweichung ist ausnahmslos höher als bei Daten aus dem BBM.

Unabhängig von der Herkunft der Daten spricht das Simulationsergebnis dafür, dass die Schätzer WEP, K und ANOVA empfehlenswerte Methoden sind.

5 Anwendungen

Gegenstand dieses Kapitels ist die Schätzung und Darstellung des Intra-Klumpen-Korrelationskoeffizienten als Maß für das Vorhandensein paarweiser Zusammenhänge für zwei Datensätze aus Finanzwirtschaft und Demografie. Beide Datensätze werden im folgenden Abschnitt ausführlich beschrieben.

Die Abschnitte 5.3 und 5.4 widmen sich Ergebnissen aus der Stichprobenplanung bzw. der Einbeziehung von Kovariablen. Das Kapitel schließt mit der Berechnung von Konfidenzintervallen für die beiden Datensätze.

5.1 Datenbeschreibung

Für die Analyse stehen zwei Klumpenstichproben zur Verfügung: 1. eine Erhebung der Zahngesundheitsstatus von Kindern in zufällig ausgewählten Kindergärten (Zahngesundheitsstudie (ZAHN)) und 2. eine amtliche Insolvenz-Statistik aus Mecklenburg-Vorpommern (INSOL). Hier repräsentiert die Unterteilung nach Branchen das Klumpen-Design. Die Daten existieren auf zwei unterschiedlichen Verdichtungsebenen:

1. Elemente der Klumpen (ZAHN):
 - binäre Zufallsvariable X_{ij} , oder
2. Gesamtanzahl der Erfolge im Klumpen (INSOL):
 - aggregierte Zufallsvariable $Y_i = \sum_j X_{ij}$.

5.1.1 Zahngesundheitsstudie

Die Daten der Zahngesundheitsstudie (ZAHN) stammen aus zahnärztlichen Untersuchungen der Jahre 1990 bis 2000 in 170 Kindergärten. Die meisten dieser Untersuchungen kommen aus dem hessischen Landkreis Groß-Gerau (139 Kindergärten). Die restlichen Daten wurden aus Landkreisen der Region Frankfurt zur Verfügung gestellt. Im ganzen Land Hessen gab es in dieser Zeit etwa 3.500 Kindergärten (Statistik-Hessen, 2013) mit etwa 186.000 Kindern im Alter von 3 bis 5 Jahren (DAJ-Mitgliederversammlung, 1998). Allerdings schwanken diese Zahlen über den Beobachtungszeitraum beträchtlich (Geburtenrückgang, Zuzug, Abwanderungen usw.).

Der lange Untersuchungszeitraum und die damit verbundene Untersuchung an unterschiedlichen Kohorten sowie die relativ große regionale Streuung der untersuchten Kindergärten lässt die Annahme zu, dass die Auswahl der Kindergärten unabhängig und zufällig erfolgte.¹ Merkmalsträger sind die Kinder im Alter von 3 bis 5 Jahren. Es wird angenommen, dass alle Kinder eines Kindergartens, die diesem Merkmal entsprechen, untersucht wurden. Insgesamt wurden 7.978 3 bis 5-Jährige registriert, die sich ungleichmäßig auf die 170 Kindergärten verteilen. Damit bilden die Kindergärten natürliche Klumpen mit variabler Klumpengröße (unbalanciertes Design).

Im Rahmen der Gruppenprophylaxe in Kindergärten wurde der dmft-Wert eines jeden Kindes erfasst. Dieser Wert bezeichnet einen Summen-Score im Milchgebiss und gibt an, wie viele Milchzähne kariös (*decayed*), wegen Karies extrahiert (*missing*) oder gefüllt (*filled*) sind. Dabei sind Werte zwischen 0 und 20 möglich, wobei 0 „gebissgesund/keine Karieserfahrung“ bedeutet.

Operationalisierung Für die Analyse der Intra-Klumpen-Korrelation wird für jedes Element (Kind i) eines Klumpens (Kindergarten j) der dmft-Wert dichotomisiert und in der Zufallsvariable X_{ij} gespeichert, sodass X_{ij} die Werte 0 (gebissgesund) und 1 (nicht gebissgesund) annehmen kann. Die Gesamtanzahl an Kindern mit Karieserfahrung im Klumpen i wird mit $Y_i := \sum_j X_{ij}$ bezeichnet.

Die Grundgesamtheit besteht also aus $k = 170$ Klumpen mit den Klumpengrößen n_i , $i = 1, \dots, k$, wobei $\sum_i n_i = 7978 = N$ die Stichprobengröße ist.

Deskription Die Klumpengrößen schwanken zwischen 10 und 107 mit einer mittleren Größe von 46,9 Kindern. Innerhalb der Klumpen beträgt die Karies-Prävalenz, also die Häufigkeit des Ereignisses „Kind hat Karieserfahrung“, zwischen 12% und 75%. Die mittlere Prävalenz beträgt 38%.

Von allen 7.978 Kindern sind 1.710 3-jährig, 3.077 4-jährig und 3.191 im Alter von 5 Jahren. Grundsätzlich gilt die Tendenz, dass ältere Kinder eine höhere Karies-Prävalenz aufweisen (siehe Tabelle 5.1). So steigt der Anteil der Kinder mit Karieserfahrung in der untersuchten Altersklasse jeweils nach einem Jahr um etwa 10%-Punkte. Der komplette Datensatz ist auf aggregierter Verdichtungsebene in Anhang A.1.1 aufgeführt.

Tabelle 5.1: Karies-Prävalenz nach Alter

Alter (in Jahren)	3	4	5
Karies-Prävalenz (in %)	26,1	36,6	44,7

¹Dies ist auf Grund von praktischen und organisatorischen Zwängen allerdings nicht der Fall.

5.1.2 Insolvenzen in Mecklenburg-Vorpommern 2006

Der Datensatz „Insolvenzen in Mecklenburg-Vorpommern 2006“ (INSOL) ist eine Vollerhebung, die sich aus zwei Berichten des Statistischen Amtes Mecklenburg-Vorpommern ergibt: erstens „Unternehmen und Betriebe in Mecklenburg-Vorpommern, Stand 31.12.2006“ (Statistisches Amt MV, 2015a) und zweitens „Insolvenzen in Mecklenburg-Vorpommern, 01.01. bis 31.12.2006“ (Statistisches Amt MV, 2015b).

Die erste Erhebung, das Unternehmensregister, stellt fest, wie viele wirtschaftlich aktive Unternehmen und Betriebe es zum Stichtag 31.12.2006 in Mecklenburg-Vorpommern gab. Unternehmen ohne Umsatzsteuerpflicht und ohne sozialversicherungspflichtig Beschäftigte bleiben dabei weitgehend unberücksichtigt. Die Unternehmen sind u. a. nach Branchen klassifiziert. Die maßgebliche Systematik dafür folgt der „Klassifikation der Wirtschaftszweige, Ausgabe 2003 (WZ 2003)“, eine einheitliche statistische Einordnung von Wirtschaftszweigen in der Europäischen Gemeinschaft. Die Branchen bilden somit eine natürliche Grundlage für die Zusammensetzung von Klumpen in der weiteren Analyse.

Die zweite Erhebung ist die Insolvenzstatistik Mecklenburg-Vorpommerns für das Jahr 2006. Dort sind alle eröffneten sowie mangels Masse abgelehnten oder mit Schuldenbereinigungsplan beendeten Insolvenzverfahren erfasst. Im Allgemeinen ist die Zahlungsunfähigkeit eines Unternehmens der Eröffnungsgrund für ein Insolvenzverfahren. Diese sogenannten Regelinsolvenzverfahren sind im Bericht u. a. wieder nach Branchen klassifiziert (WZ 2003).

Die o. g. Unternehmen sind Merkmalsträger für den Datensatz INSOL. Neun Branchen werden ausgewiesen. Die Anzahl der Unternehmen jeweils einer Branche addiert mit der Anzahl der Insolvenzen bis zum Stichtag ergibt die Grundgesamtheit dieser Branche. Insgesamt stehen damit $N = 53391$ Unternehmen für die Analyse zur Verfügung.

Operationalisierung Für die Analyse der Intra-Klumpen-Korrelation werden die im amtlichen Bericht genannten Insolvenz-Zahlen für jede der k Klumpen (Branchen) ($k = 1, \dots, 9$) in der Zufallsvariable der Gesamtanzahl an Ausfällen Y_i , $i = 1, \dots, k$ gespeichert. Die Ergebnisse auf Ebene der Elemente der Klumpen werden hier also nicht beobachtet. Die Klumpengröße wird mit n_i , $i = 1, \dots, k$ bezeichnet.

Deskription Die Klumpengrößen liegen zwischen 12.363 und 966 Unternehmen. Die durchschnittliche Größe beträgt 5.932 Unternehmen. Im Mittel sind innerhalb eines Jahres 1,12% der Unternehmen zahlungsunfähig gemeldet. Dieser Wert schwankt innerhalb der Branchen zwischen 0,07% und 2,27%. Der komplette Datensatz kann in Anhang A.1.2 eingesehen werden.

Zusammenfassung Die vorgestellten Datensätze, einer aus der Demografie und einer aus der Finanzwirtschaft, unterscheiden sich in ihrer Struktur stark. Die Zahngesundheitsdaten bestehen aus relativ vielen Klumpen mit einer relativ hohen Eintrittshäufigkeit des definierten Ereignisses. Die Insolvenzdaten sind auf wenige Klumpen verteilt, wobei es relativ viele Elemente pro Klumpen gibt. Die Eintrittshäufigkeit ist niedrig. Die Entstehung beider Datensätze lässt sich exemplarisch durch Beobachtung der entsprechenden Ereignis-Prävalenz an einem festen Zeitpunkt bei Verweildauerdaten illustrieren (siehe Abbildung 5.1).

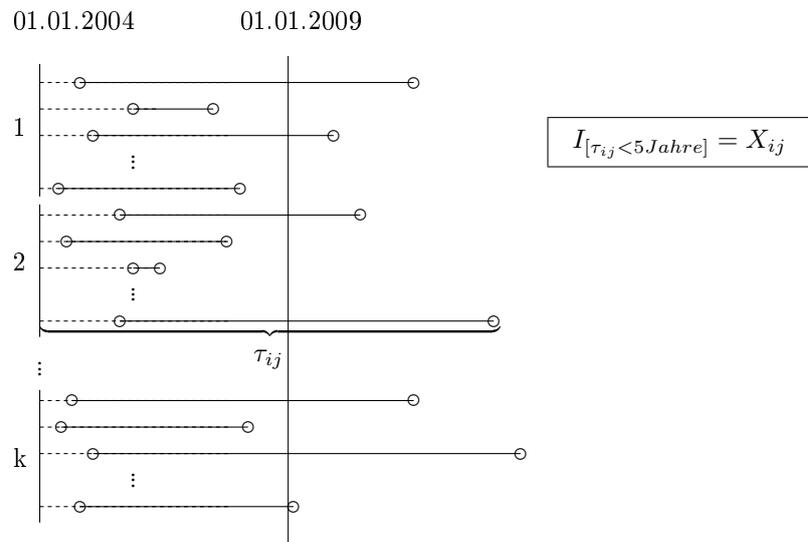


Abbildung 5.1: Daten mit Klumpen-Design im Zeitverlauf

Dort ist der 01.01.2004 ein modellhaftes Startdatum. Für $1, \dots, k$ Klumpen steht jede waagerechte Linie für ein Element des Klumpens. Jedes Element beginnt im Zeitablauf von links nach rechts seinen Lebenszyklus (Geburt bei ZAHN, Unternehmensgründung bei INSOL), markiert durch einen Kreis. Das Auftreten des zu untersuchenden Ereignisses (Karies-Prävalenz bei ZAHN, Insolvenz bei INSOL) ist wiederum durch einen Kreis am rechten Ende der Linie gekennzeichnet. Dann ist die Intra-Klumpen-Korrelation genau die Querschnittsabhängigkeit am modellhaften Beobachtungszeitpunkt 01.01.2009. Die Frage, ob für das j -te Element im i -ten Klumpen das Ereignis eingetreten ist, wird gleichsam bejaht, wenn im Klumpen i die j -te Verweildauer τ_{ij} den Beobachtungszeitraum nicht übersteigt.

5.2 ICC-Schätzung für die Datensätze ZAHN und INSOL

Für die Anwendung aller Methoden, deren ICC-Schätzer von der Erfolgswahrscheinlichkeit π abhängen, gilt die Konvention, dass π in einem vorgelagerten Schritt mit der Maximum-Likelihood-Methode geschätzt wird. Für die Anwendung der Methoden, die auf dem CCM basieren, gilt dabei die Annahme der Gleichheit der Erfolgswahrscheinlichkeit π in allen Klumpen. Sie unterscheidet sich in den Stichproben nur durch die natürliche Streuung (vgl. Kapitel 6 (Ausblick)). Die Schätzer für π sind $\hat{\pi}_{ML} = 0,38$ für die Zahngesundheitsdaten und $\hat{\pi}_{ML} = 0,011$ für die Insolvenzdaten. Alle Berechnungen erfolgen mit den ursprünglich für die Simulationen entwickelten SAS-Quellcodes in angepasster Form. Die Original-Quellcodes sind im Anhang A.3.1 aufgeführt.

Maximum-Likelihood-Schätzer Die Maximum-Likelihood-Methode „versagt“ wie in den vorgehenden Kapiteln angedeutet bei den Datensätzen ZAHN und INSOL. Da es in keinem der Datensätze einen oder mehrere Klumpen mit entweder keinem oder n_i Erfolgen gibt, entfällt der entsprechende Anteil aus der Score-Funktion. Dies führt zum mathematischen Problem, dass der ML-Schätzer von ρ Nullstelle einer Funktion ist, deren Lösung gegen $-\infty$ bzw. $+\infty$ strebt.

Quasi-Likelihood-Schätzer Die Quasi-Likelihood-Methode konnte in den Simulationen keine plausiblen Schätzwerte für die Intra-Klumpen-Korrelation hervorbringen. Sie verhält sich auch für die Datensätze ZAHN und INSOL nicht evident.

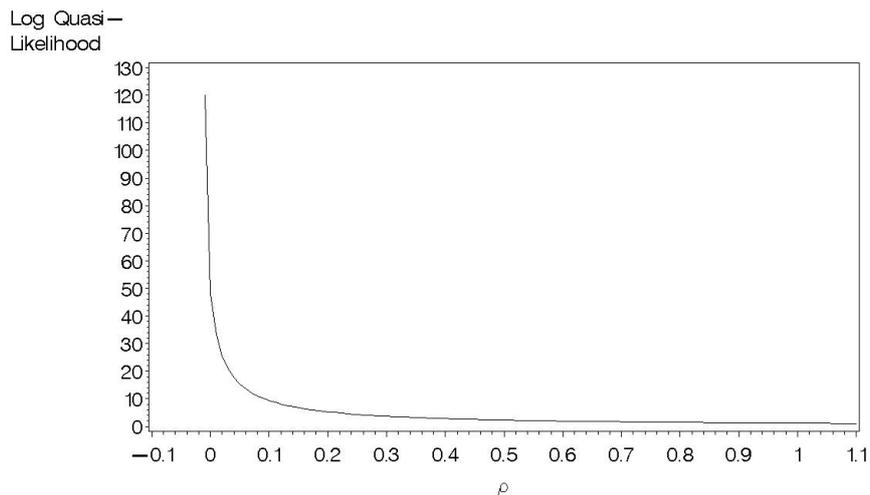


Abbildung 5.2: Log-Quasi-Likelihood-Funktion (Datensatz ZAHN)

5 Anwendungen 5.2 ICC-Schätzung für die Datensätze ZAHN und INSOL

Abbildung 5.2 zeigt die Log-Quasi-Likelihood-Funktion für die Daten aus der Zahngesundheitsstudie.

Die Erfolgswahrscheinlichkeit beträgt $\pi = 0,38$ und ρ nimmt Werte in einer Spanne von $-0,01$ bis $1,1$ an. Da ρ im CCM links von 0 und rechts von 1 begrenzt ist, gibt es nur ein zulässiges (lokales) Maximum am linken Rand des Intervalls bei $\rho = 0$. Der Wert der Log-Quasi-Likelihood-Funktion steigt ungeachtet dessen für Werte $\rho < 0$ weiter an.

Extended-Quasi-Likelihood-Schätzer In Abbildung 5.3 ist der Wert der Extended-(Log)-Quasi-Likelihood gegen Werte von ρ im Intervall $[-0,005; 1,1]$ grafisch dargestellt. Datengrundlage sind die Zahngesundheitsdaten ($\pi = 0,38$). Werte für ρ , die kleiner als $-0,005$ sind, ergeben keine zulässigen Likelihood-Werte. In der Abbildung ist ein akzeptables Maximum für ρ bei $\hat{\rho}_{EQL} = 0,04$ ersichtlich.

Analog dazu schätzt die gleiche Methode die Intra-Klumpen-Korrelation für die Insolvenzdaten bei $\hat{\rho}_{EQL} = 0,005$ (siehe Abbildung 5.4).

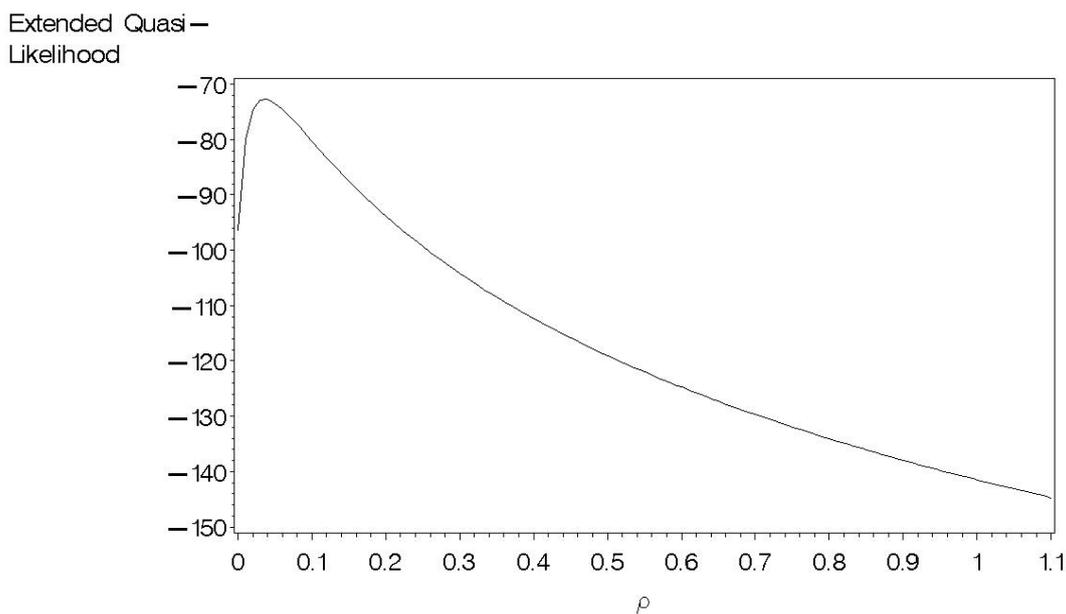


Abbildung 5.3: Extended-Quasi-Likelihood-Funktion (Datensatz ZAHN)

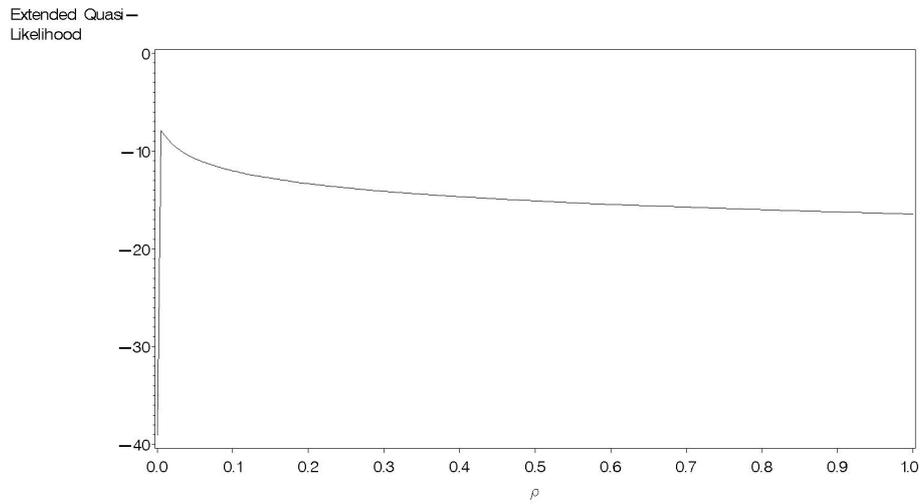


Abbildung 5.4: Extended-Quasi-Likelihood-Funktion (Datensatz INSOL)

Pseudo-Likelihood-Schätzer Der Pseudo-Likelihood-Schätzer beträgt für ZAHN $\hat{\rho}_{PL} = 0,1$ und ist damit um den Faktor 2,5 größer als das Ergebnis der EQL-Methode. Für die Insolvenzdaten ist in Abbildung 5.5 die Pseudo-Likelihood-Funktion für Werte von ρ im Intervall $[0; 1]$ dargestellt.

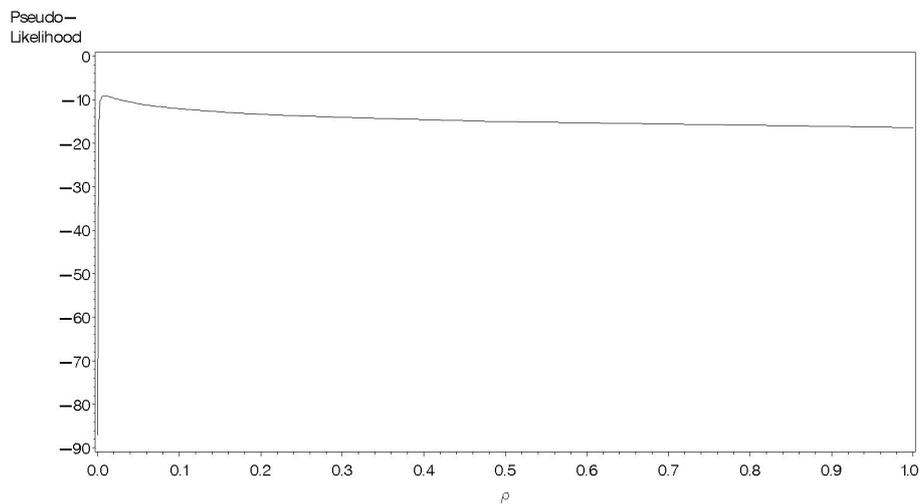


Abbildung 5.5: Pseudo-Likelihood-Funktion (Datensatz INSOL)

Direkter Beta-Binomial-Modell-Schätzer Der Algorithmus zur direkten Schätzung der Parameter der Beta-Verteilung ergibt beim Datensatz ZAHN das Wertepaar $(\hat{\alpha}; \hat{\beta}) = (4; 6,5)$, was zum Schätzer $\hat{\rho}_{DBB} = 0,086$ führt. Für den Datensatz INSOL wird das Paar $(\hat{\alpha} = 4; \hat{\beta} = 365)$ bestimmt ($\hat{\rho}_{DBB} = 0,0027$).

Bayes-Schätzer Das Reziproke des Bayes-Schätzers ist durch die Stichprobengröße $N = 7978$ (ZAHN) bzw. $N = 53391$ (INSOL) dominiert. Der Schätzer geht somit gegen Null, wenn N gegen plus unendlich strebt. Im Fall der beiden Datensätze besitzt $\hat{\rho}_B$ jeweils erst ab der vierten Nachkommastelle positive Ziffern, weshalb der Schätzer hier mit gerundet Null angegeben wird. Im Gegensatz zur ML/QL-Methode unterstreicht dieser Schätzer zumindest das Vorhandensein einer Intra-Klumpen-Korrelation.

Weighted-Empirical-Pairwise-Schätzer Die Schätzung von ρ mittels WEP-Methode hat folgende Werte zum Ergebnis: $\hat{\rho}_{WEP} = 0,007$ (ZAHN) bzw. $\hat{\rho}_{WEP} = 0,0025$ (INSOL).

Kappa-Schätzer Durch Anwendung der Gleichungen zur Kappa-Schätzung auf ZAHN bzw. INSOL errechnen sich $\hat{\rho}_K = 0,029$ bzw. $\hat{\rho}_K = 0,22$. Der Schätzwert der Insolvenzdaten liegt somit um den Faktor 100 höher als bspw. der des WEP-Schätzers. Generell liegen beide Werte über dem jeweiligen Schätzer der WEP-Methode, da große Klumpen dort ein größeres Gewicht bekommen als bei der Kappa-Methode.

Pairwise-Equal-Weights-Schätzer Für den Fall der Zahngesundheitsdaten beträgt der Equal-Weights-Schätzer $\hat{\rho}_{PEQ} = 0,031$. In den Insolvenzdaten dagegen entspricht die Intra-Klumpen-Korrelation dem Schätzer zufolge $\hat{\rho}_{PEQ} = 0,0023$.

ANOVA-Schätzer Der ANOVA-Schätzer misst die gewichteten quadratischen Abweichungen der Erfolge zwischen und innerhalb der Klumpen. Die Formel ergibt folgende Schätzwerte: $\hat{\rho}_{ANOVA} = 0,034$ (ZAHN) und $\hat{\rho}_{ANOVA} = 0,0033$ (INSOL).

Moment-Method-Based-Schätzer Die Momentenmethode lässt zunächst einen Test auf die Hypothese $H_0 : \rho = 0$ zu. Die Prüfgröße beträgt für die Daten der Zahngesundheitsstudie $T_{H_0} = 430,75$. Der kritische Wert zum Signifikanzniveau 5% ist 208 und berechnet sich als das 97,5% Quantil der Chi-Quadrat-Verteilung mit Parameter 170 (zweiseitiger Test). Die Prüfgröße überschreitet den kritischen Wert, weshalb die Nullhypothese abgelehnt wird. Die Daten sind nicht unabhängig.

Im Falle der Insolvenzdaten ist $T_{H_0} = 175,6$. Der kritische Wert zum 5% Signifikanzniveau beträgt 23,3 (12 Freiheitsgrade). Auch hier wird H_0 abgelehnt und es kann von signifikant abhängigen Daten ausgegangen werden.

5 Anwendungen 5.2 ICC-Schätzung für die Datensätze ZAHN und INSOL

Der ICC-Schätzer nach Momentenmethode ist implizit gegeben und entspricht genau dem ρ , welches den Wert des Terms

$$\sum_{i=1}^k \frac{(Y_i - n_i\pi)^2}{n_i\pi(1-\pi)(1+(n_i-1)\rho)}$$

am nächsten an die Klumpenanzahl von 170 (ZAHN) bzw. 12 (INSOL) führt. Das Resultat ergibt für die Zahn-Daten eine Korrelation in Höhe von $\hat{\rho}_{MMB} = 0,036$ ($\hat{\rho}_{MMB} = 0,0033$ für die Insolvenzdaten).

Zusammenfassung Alle Ergebnisse sind in Tabelle 5.2 zusammengefasst. Die mathematisch korrekten jedoch sachlich unplausiblen Werte für $\hat{\rho}_{ML}$ und $\hat{\rho}_{QL}$ sind nachrichtlich ebenso aufgeführt.

Insgesamt ist festzustellen, dass viele Schätzer auf dem jeweiligen Datensatz sehr ähnliche Ergebnisse hervorbringen. Fünf der elf Schätzmethode vermuten den wahren Wert des Parameters zwischen 0,029 und 0,04 (ZAHN). Bei den Insolvenzdaten sind sogar sechs Schätzer im engen Intervall $[0,0023; 0,005]$ zu finden.² Bis auf den EQL-Schätzer der Zahn-Daten weisen alle Likelihood-Methoden auf beiden Datensätzen stark von diesen Intervallen abweichende Werte auf. Modellunabhängige Schätzer (WEP, K, PEQ, ANOVA, MMB) liegen dagegen innerhalb des engen Bandes (außer WEP bei ZAHN und K bei INSOL).

Tabelle 5.2: Schätzwerte der Intra-Klumpen-Korrelation für die Datensätze ZAHN und INSOL

Schätzer	ZAHN	INSOL
$\hat{\pi}_{ML}$	0,38	0,011
$\hat{\rho}_{ML}$	0,0	0,0
$\hat{\rho}_{QL}$	0,0	0,0
$\hat{\rho}_{EQL}$	0,04	0,005
$\hat{\rho}_{PL}$	0,1	0,009
$\hat{\rho}_{DBB}$	0,086	0,0027
$\hat{\rho}_B$	≈ 0	≈ 0
$\hat{\rho}_{WEP}$	0,007	0,0025
$\hat{\rho}_K$	0,029	0,22
$\hat{\rho}_{PEQ}$	0,031	0,0023
$\hat{\rho}_{ANOVA}$	0,034	0,0033
$\hat{\rho}_{MMB}$	0,036	0,0033

²Die Häufung von Schätzern in der Nähe eines bestimmten Wertes kann allerdings kein Anhaltspunkt für die Lage des wahren Parameters sein.

5.3 Stichprobenplanung (ZAHN)

Ein verbreiteter Fall für die Notwendigkeit einer ICC-Schätzung ist die Fallzahlplanung in *Cluster*-randomisierten Studien, wie sie in der Medizin üblich sind, aber auch z. B. in der Wirtschaft Anwendung finden. Hier muss abgewogen werden zwischen einer – aus Kostengründen – möglichst kleinen Stichprobe und einer Stichprobe, die groß genug ist, um das vorher festgelegte Vertrauensniveau einzuhalten. Wird eine Klumpenstichprobe gezogen, so besteht jedoch eine gewisse Wahllosigkeit beim Ziehen der Merkmalsträger, da die Klumpen zufällig ausgewählt werden und nicht deren Elemente. Eine Möglichkeit dies zu berücksichtigen ist es, zunächst die Fallzahl bei einfacher Zufallsstichprobe (n_{SRS} , *Simple Random Sample*) zu berechnen und diese um den Varianzinflationsfaktor zu inflationieren. Ergebnis ist die benötigte Mindest-Fallzahl für die Klumpenstichprobe (n_{CRS} , *Cluster Randomized Sample*, vgl. Donner und Klar (2000, Kapitel 1.3))

$$\hat{n}_{CRS} = n_{SRS} \cdot VIF. \quad (5.1)$$

Die Berechnung von n_{SRS} für den Datensatz ZAHN erfolgt über

$$n_{SRS} = \frac{N}{1 + \frac{Ne^2}{u_{\alpha/2}^2 \pi_L (1 - \pi_L)}}$$

mit

$$\begin{aligned} N &= 7978, \\ e &= 0,03 \text{ (vorgegebene Präzision 3\%)}, \\ u_{\alpha/2} &= 1,96 \text{ ((1 - } \alpha/2) \text{- Quantil der Standardnormalverteilung (} \alpha = 5\% \text{))}, \\ \hat{\pi} &= 0,62 \text{ (Anteil kariesfreier Kinder)}. \end{aligned}$$

Die Berechnungsformel ist Weißbach *et al.* (2015) entnommen. Dort ist das Thema Stichprobenplanung für zahnmedizinische Untersuchungen in Kindergärten ausführlicher dargestellt.

Für die einfache Stichprobe ergibt sich nun $n_{SRS} = 893$. Zum Berechnen der Klumpenstichprobe muss der Varianzinflationsfaktor geschätzt werden³. Es sei

$$\begin{aligned} \widehat{VIF} &= \widehat{VIF}^* \\ &= \left[1 + \left(\frac{\sum n_i^2}{N} - 1 \right) \hat{\rho}_{ANOVA} \right] \\ &= 2,768. \end{aligned}$$

³Hier ist dafür die Grundgesamtheit bekannt. In anderen Fällen wird auf Werte vergleichbarer Studien oder Vollerhebungen zurückgegriffen.

Mit Gleichung 5.1 ergibt sich für die Klumpenstichprobe eine benötigte Fallzahl von mindestens 1.580 Kindern um mit einer maximalen zufälligen Abweichung von 3% in im Mittel 95 von 100 Fällen eine wahre Aussage über die Grundgesamtheit zu treffen. Dazu müssen $1580/47 \approx 34$ Kindergärten der durchschnittlichen Größe von 47 Kindern untersucht werden.

5.4 Das Alter als Kovariable (ZAHN)

Für die Altersstufen $a \in \{3, 4, 5\}$ sei

$$X_{ij}^a = \begin{cases} 1, & \text{Kind im Alter } a \text{ hat Karieserfahrung} \\ 0, & \text{Kind im Alter } a \text{ hat keine Karieserfahrung} \end{cases}$$

eine Zufallsvariable, die die Zahngesundheit des j -ten Kindes im i -ten Kindergarten in einem bestimmten Alter beschreibt. Dann ist

$$Y_i = \sum_{j=0}^{n_i} X_{ij} = \sum_{j=0}^{n_i^3} X_{ij}^3 + \sum_{j=0}^{n_i^4} X_{ij}^4 + \sum_{j=0}^{n_i^5} X_{ij}^5$$

die Gesamtzahl an Kindern mit Karieserfahrung im i -ten Kindergarten. Die Definitionen

$$P(X_{ij}^a = 1) =: \pi^a, \quad \text{Kor}(X_{ij}^a, X_{ij'}^a) =: \rho^a$$

vervollständigen ein Untermodell des Modells mit pauschaler Klumpenkorrelation.

Mit den Daten der Zahngesundheitsstudie ergibt sich pro Altersstufe ein neuer Schätzer für die Intra-Klumpen-Korrelation (siehe Tabelle 5.3).

Tabelle 5.3: Zahngesundheitsdaten (ZAHN) nach Alter

Alter in Jahren	Anzahl Kinder	$\hat{\pi}$	$\hat{\rho}$
3	1.709	$\hat{\pi}^3 = 0,26$	$\hat{\rho}_{MMB}^3 = 0,034$
4	3.066	$\hat{\pi}^4 = 0,37$	$\hat{\rho}_{MMB}^4 = 0,045$
5	3.191	$\hat{\pi}^5 = 0,45$	$\hat{\rho}_{MMB}^5 = 0,038$
(alle)	(7.978)	$(\hat{\pi} = 0,38)$	$(\hat{\rho}_{MMB} = 0,036)$

Generell ist ersichtlich, dass die Zugehörigkeit zu einer Altersgruppe einen wesentlichen Einfluss auf die Schätzer hat. Obwohl die geschätzte Erfolgswahrscheinlichkeit mit steigendem Alter deutlich zunimmt, steigt der ρ -Schätzer zunächst bei steigendem Alter an, sinkt dann aber wieder für die Altersstufe „5“.

In der Altersstufe „3“ liegt die Intra-Klumpen-Korrelation unterhalb der des gesamten Datensatzes. Dies kann ein Hinweis auf den Ursprung der gemessenen, paarweisen Korrelation sein. Sie kann einerseits hauptsächlich der gegenseitigen Beeinflussung der Kinder untereinander geschuldet sein, andererseits kann primär das soziale Umfeld auf die Kinder wirken. Für die Altersstufe „3“ scheint das soziale Umfeld ein wichtiger Faktor zu sein. Da $\hat{\rho}_{MMB}^3$ (Intra-Klumpen-Korrelation für reduzierte Anzahl Kinder pro Klumpen) kleiner ist als $\hat{\rho}_{MMB}$ mit der vollen Anzahl an möglichen Paarbildungen im Klumpen, ist im Umkehrschluss die Interaktion unwichtiger.

Im Alter von vier bzw. fünf Jahren steigt die Bedeutung der Interaktion an und das soziale Umfeld verliert an Einfluss ($\hat{\rho}_{MMB}^4 > \hat{\rho}_{MMB}$ und $\hat{\rho}_{MMB}^5 > \hat{\rho}_{MMB}$).

5.5 Konfidenzintervall für die Intra-Klumpen-Korrelation

Alternative Methode In die Intervallschätzung geht über Gleichung 3.2 ein Schätzer von $\alpha + \beta$ bzw. den unbekanntem Parameter ρ ein. [Lui et al. \(1996\)](#) wählen den ANOVA-Schätzer für ICCs, was zu $\alpha + \beta = 103$ für die Zahngesundheitsstudie führt (0,44 für die Insolvenzdaten). Eine direkte Schätzung von α und β mit Hilfe der aus Abschnitt 3.2.1 bekannten Methode liefert den Wert von $\alpha + \beta = 10,5$ (ZAHN) bzw. 369 (INSOL) und passt die Beta-Verteilung wesentlich besser an die vorhandenen Daten an.

Die sich ergebenden Konfidenzintervalle für die Datensätze ZAHN und INSOL befinden sich in Tabelle 5.4. Alle zur Berechnung benötigten Programm-Quellcodes in SAS 9.4 sind Teil des Anhangs A.3.2.

Das Intervall für die Zahngesundheitsdaten ist relativ breit, wobei es in der alternativen Berechnung etwas kleiner ist. Der Wert des rechten Randes ist jedoch bei beiden Methoden mindestens drei mal so groß wie der ANOVA-Punktschätzer (0,034) und lässt selbst im günstigsten Fall der in der Tabelle erscheinenden Intervalle einen Varianzinflationsfaktor von 11 bei durchschnittlicher Klumpengröße von 100 zu.

Für die Insolvenzdaten liegt das berechnete Intervall nach [Lui et al. \(1996\)](#) außerhalb plausibler Grenzen. Der Grund ist in der Beschaffenheit der Daten zu finden. Die hohe Variation in der Größe der einzelnen Klumpen sorgt für eine hohe Varianz in Z . Es ist unklar, ob das von den Autoren benutzte Theorem von Fieller an dieser Stelle zulässig ist.

5 Anwendungen 5.5 Konfidenzintervall für die Intra-Klumpen-Korrelation

Die alternative Berechnungsmethode erzeugt ein Intervall, welches jedoch in Relation zum ANOVA-Punktschätzer (0,003) extrem große Schwankungen zulässt (mindestens Faktor 19). Die Ergebnisse zeigen insgesamt, dass die Sicherheit über die Intra-Klumpen-Korrelation sehr gering ist.

Tabelle 5.4: Konfidenzintervalle für ρ zu den Niveaus 90 und 95 Prozent. Der ANOVA-Punktschätzer liegt bei 0,034 (ZAHN) bzw. 0,003 (INSOL).

Datensatz	$\widehat{\alpha + \beta}$	90%-Konfidenzintervall	95%-Konfidenzintervall
Methode von Lui <i>et al.</i> (1996)			
ZAHN	103	[0; 0,111]	[0; 0,127]
INSOL	0,44	—	—
Alternative Methode			
ZAHN	10,6	[0; 0,105]	[0; 0,118]
INSOL	369	[0; 0,059]	[0; 0,225]

Zusammenfassung

Essenz dieses Kapitels ist vor allem die Erkenntnis, dass die Güte der einzelnen Schätzer schwer zu beurteilen ist und vom konkreten Datensatz abhängt. Es gibt einen bemerkenswerten Unterschied in der Lage der Schätzergebnisse in Relation zueinander bei authentischen Daten im Vergleich zu jener in den Simulationsergebnissen aus Kapitel 4. So ist bspw. der MMB-Schätzer in den Simulationen noch um den Faktor 2,5 niedriger als der Häufungspunkt⁴ aller Schätzer (0,03 im BBM). Der gleiche Schätzer angewandt auf den Datensatz ZAHN, welcher eine ähnliche Struktur wie die Simulationsdaten hat, befindet sich jedoch im Einklang mit vier weiteren Schätzern im Bereich des Häufungspunktes.

Umgekehrt ist der Fall beim WEP-Schätzer. Er fügt sich bei den Simulationen in beiden Datenmodellen in den Bereich des jeweiligen Häufungspunktes (korrekter Wert für ρ) ein. Bei der Anwendung auf INSOL kann der Schätzer diese Eigenschaft ebenfalls aufweisen, auf die ZAHN-Daten angewandt jedoch weicht WEP deutlich von der Masse der Schätzer ab.

Ein dritter Fall ist der PL-Schätzer, welcher in den Simulationen unabhängig vom Datenmodell die höchste Varianz aufweist und auch in den Ergebnissen für ZAHN/INSOL weit entfernt vom Häufungspunkt liegt.

⁴Das Wort „Häufungspunkt“ bezeichnet hier den gerundeten Durchschnitt einer Menge von Schätzwerten, die „nah“ an diesem Punkt liegen.

5 Anwendungen 5.5 Konfidenzintervall für die Intra-Klumpen-Korrelation

Weiterhin stellt sich die Frage, warum der Kappa-Schätzer in den Simulationen sehr gut abschneidet, in INSOL aber nicht. Die Antwort scheint in der speziellen Datenstruktur der Insolvenzdaten zu liegen (relativ wenige Klumpen, stark unterschiedliche Klumpengröße).

Alle aufgeführten Beispiele begründen den Fakt, dass es schwer ist vorherzusagen, welcher Schätzer für welche Datenstruktur die „besten“ Ergebnisse liefert. Die Vielzahl an vorhandenen Schätzern ist somit gerechtfertigt und eine adäquate Modellwahl wird zum zentralen Thema weiterer Forschung.

6 Ausblick

In dieser Arbeit wurden elf verschiedene ICC-Schätzer vorgeschlagen. Ihr Anwendungsgebiet liegt stets auf korrelierten binären Daten mit Klumpenstruktur. Es gibt in der Literatur weitergehende Betrachtungen, die einerseits speziellere Anwendungsgebiete suchen oder andererseits eine Reihe von Kritikpunkten an o. g. Schätzern ausräumen wollen.

Ein solcher Kritikpunkt ist, dass die vorgestellten Modelle nur auf paarweise Korrelation prüfen. [Stefanescu und Turnbull \(2003\)](#) beachten Korrelationen höherer Ordnung. Ausgehend von der Pearson'schen Definition bezeichnen die Autoren die Korrelation r -ter Ordnung in Klumpen i als

$$\rho_r = \frac{E((X_{i1} - \pi) \dots (X_{ir} - \pi))}{\pi(1 - \pi)^{r/2}}, \quad r = 2, \dots, n_i.$$

Damit ist modelliert, dass o. B. d. A. Element 1 auf Element 2 wirkt und beide zusammen auf Element 3 (Korrelation 3. Ordnung) usw. bis zur Korrelation n_i -ter Ordnung. Somit variiert die Korrelation auch mit der Klumpengröße. Für jede Untermenge von r Elementen besteht die gleiche Korrelation r -ter Ordnung. $r = 2$ entspricht der paarweisen Korrelation. [Stefanescu und Turnbull \(2003\)](#) geben an, dass die Schätzung mittels EM-Algorithmus gute Konvergenz-Eigenschaften bietet - sogar bei Klumpengrößen von 150.

Sowohl das CCM als auch das BBM gehen von der Gleichheit der Korrelation für alle möglichen Paare von Elementen innerhalb eines Klumpens aus. Dass dies zumindest im Zahngesundheitsdatensatz nicht so ist, wurde in Kapitel 5.4 gezeigt. Die Korrelation ist dort nicht stationär im Alter. Denkbar ist weiterhin, die Stationarität in der Kalenderzeit zu untersuchen und somit zur Querschnitt-Untersuchung von Intra-Klumpen-korrelierten Daten die Längsschnitt-Analyse hinzuzufügen.

Gleichermaßen stellt sich die Frage, ob die Annahme der Gleichheit der Erfolgswahrscheinlichkeit π für alle Klumpen im CCM richtig ist. Ein Test zur Beantwortung der Frage würde bei der hohen Anzahl an Klumpen sicher verneint. Die Hinzunahme des zusätzlichen Parameters ρ ist daher ein richtiger Schritt, um mehr Streuung zu erklären. Das erweitert jedoch die Fragestellung auf die mögliche Ungleichheit von (π, ρ) zwischen den Klumpen. Mit Hilfe der gemeinsamen Verteilung beider Parameter ist es vorstellbar, einen Test dahingehend zu entwickeln.

Ein weiterer Kritikpunkt ist das Fehlen zusätzlicher Regressoren in den vorgestellten Modellen. Dabei ist bspw. im Beta-Binomial-Modell gar nicht wichtig, was auf die Korrelation wirkt, sondern welche Einflüsse es auf die Erfolgswahrscheinlichkeit π gibt. Diese können in einer Funktion von Regressoren zusammengefasst werden:

$$P(X_{ij} = 1) = f(\pi) + \dots$$

Einflüsse auf den Zahlungsausfall eines Unternehmens sind z.B. das Kreditrating oder die Anzahl konkurrierender Unternehmen. Das Alter eines Kindes oder die Häufigkeit des Zähneputzens können Einfluss auf die Karieswahrscheinlichkeit haben.

Da ein Modell, das Unkorreliertheit abbilden kann, nicht im Beta-Binomial-Modell genestet ist, macht mindestens die Beachtung des Modells mit pauschaler Klumpenkorrelation Sinn. Dort ist die Möglichkeit auf $\rho = 0$ zu testen gegeben. In der Arbeit wird gezeigt, dass Schätzer je nach Modell zum Teil weit entfernte Ergebnisse liefern. Zentrales Interesse liegt daher auf der Modellwahl. Von welchem der (mindestens zwei) Modelle kann ausgegangen werden? Welches Modell kann im Sinne konsistenter Parameterschätzung überzeugen?

Der ICC könnte einerseits tatsächlich einen kausalen Zusammenhang zwischen den Elementen beschreiben. Andererseits kann ein äußerer Faktor ursächlich für den gemessenen Zusammenhang sein. Woher kommt diese sogenannte Scheinkorrelation? Welche Mittel gibt es, die Ursache vorhandener Zusammenhänge zu erkennen? In Kapitel 5.4 wurde ein erster Hinweis zur Beantwortung der Fragen gefunden.

Weiterhin ist die Beschreibung von Verweildauern über die Analyse Intra-Klumpenkorrelierter Querschnittsdaten möglich. Dazu muss allerdings Korrelations-Stationarität gelten, d. h. die Intra-Klumpen-Korrelation muss zu unterschiedlichen Zeitpunkten gleich groß sein. Weitere Forschung kann hier zusätzliche Erkenntnisse zur Nutzbarkeit liefern.

Anhang

A.1 Daten

A.1.1 Zahngesundheitsstudie

Tabelle A.1: Datensatz

ZAHN

Kindergarten	n_i	Y_i	Kindergarten	n_i	Y_i	Kindergarten	n_i	Y_i
1	50	7	21	57	21	41	59	26
2	59	21	22	61	25	42	47	32
3	42	12	23	63	28	43	40	28
4	50	16	24	23	10	44	52	11
5	33	19	25	29	11	45	50	20
6	67	23	26	95	54	46	32	14
7	61	31	27	27	11	47	58	13
8	58	21	28	39	20	48	107	44
9	52	17	29	47	18	49	21	12
10	40	11	30	43	14	50	61	31
11	45	18	31	21	15	51	40	10
12	50	15	32	23	12	52	34	16
13	58	29	33	10	4	53	68	35
14	42	24	34	60	15	54	51	13
15	41	8	35	90	25	55	42	17
16	41	25	36	64	27	56	65	24
17	41	26	37	24	11	57	49	19
18	44	15	38	57	28	58	36	18
19	22	7	39	31	14	59	34	12
20	24	8	40	37	11	60	73	26

Anhang

Kindergarten	n_i	Y_i	Kindergarten	n_i	Y_i	Kindergarten	n_i	Y_i
61	35	11	98	43	20	135	29	11
62	58	23	99	37	10	136	59	20
63	42	14	100	60	16	137	52	22
64	19	13	101	72	24	138	54	20
65	20	11	102	23	10	139	60	41
66	60	22	103	70	23	140	68	28
67	79	23	104	56	32	141	59	22
68	48	22	105	51	22	142	68	18
69	48	22	106	45	19	143	32	8
70	18	6	107	38	17	144	66	25
71	53	20	108	64	20	145	54	16
72	45	20	109	36	11	146	48	22
73	73	24	110	64	22	147	51	22
74	25	13	111	44	15	148	56	26
75	60	16	112	62	26	149	71	19
76	60	19	113	25	14	150	55	21
77	40	16	114	53	25	151	18	7
78	55	18	115	40	11	152	35	6
79	39	12	116	11	5	153	64	18
80	68	18	117	29	9	154	43	18
81	49	16	118	31	4	155	53	15
82	63	19	119	40	15	156	13	7
83	59	18	120	66	19	157	58	17
84	48	29	121	43	9	158	41	13
85	43	20	122	26	4	159	57	7
86	58	21	123	43	18	160	62	21
87	75	25	124	17	5	161	39	5
88	50	17	125	75	14	162	23	12
89	46	17	126	33	10	163	61	26
90	30	8	127	29	4	164	15	6
91	42	18	128	65	27	165	35	17
92	72	33	129	42	16	166	46	8
93	51	24	130	26	12	167	19	6
94	30	13	131	38	19	168	55	17
95	67	26	132	50	19	169	45	17
96	58	27	133	38	19	170	32	4
97	59	16	134	12	9			

A.1.2 Insolvenzen in Mecklenburg-Vorpommern 2006

Tabelle A.2: Datensatz INSOL

Branche	n_i	Y_i
Baugewerbe	7915	180
Handel, Instandhaltung und Reparatur von Kraftfahrzeugen	12363	121
Gastgewerbe	5625	75
Verkehr und Lagerei; Information und Kommunikation	2899	49
Kredit- und Versicherungsgewerbe	966	8
Grundstücks- und Wohnungswesen	11050	107
Erziehung und Unterricht	1353	1
Gesundheits-, Veterinär- und Sozialwesen	5484	11
Erbringung von sonstigen öffentl. und persönl. Dienstleistungen	5736	48

A.2 Herleitungen

A.2.1 Momente einer CCM-verteilten Zufallsvariable

Aus Gründen der besseren Lesbarkeit wird das Subskript i für die Klumpennummer weggelassen. Die folgenden Rechnungen gelten für alle Klumpen $1 \leq i \leq k$.

$$\begin{aligned}
 E(Y) &= \sum_{y=1}^{n-1} y \binom{n}{y} (1-\rho)\pi^y(1-\pi)^{n-y} + n(\rho\pi + (1-\rho)\pi^n) \\
 &= (1-\rho) \sum_{y=1}^{n-1} y \frac{n!}{(n-y)!y!} \pi^y(1-\pi)^{n-y} + n(\rho\pi + (1-\rho)\pi^n) \\
 &= (1-\rho)n\pi \sum_{y=1}^{n-1} \frac{(n-1)!}{(n-y)!(y-1)!} \pi^{y-1}(1-\pi)^{(n-1)-(y-1)} + n(\rho\pi + (1-\rho)\pi^n) \\
 &= (1-\rho)n\pi \sum_{y=1}^{n-1} \binom{n-1}{y-1} \pi^{y-1}(1-\pi)^{(n-1)-(y-1)} + n(\rho\pi + (1-\rho)\pi^n)
 \end{aligned}$$

Anhang

(Substitution $l := y - 1, m := n - 1$)

$$= (1 - \rho)n\pi \left[\sum_{l=0}^{m-1} \binom{m}{l} \pi^l (1 - \pi)^{m-l} + \pi^m - \pi^m \right] + n(\rho\pi + (1 - \rho)\pi^n)$$

(Binomischer Lehrsatz)

$$= (1 - \rho)n\pi((\pi + (1 - \pi))^m - \pi^m) + n(\rho\pi + (1 - \rho)\pi^n)$$

(Rücksubstitution)

$$\begin{aligned} &= (1 - \rho)n\pi(1 - \pi^{n-1}) + n(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho)n\pi - (1 - \rho)n\pi^n + n\rho\pi + (1 - \rho)n\pi^n \\ &= n\pi. \end{aligned}$$

$$\begin{aligned} E(Y^2) &= \sum_{y=1}^{n-1} y^2 \binom{n}{y} (1 - \rho)\pi^y (1 - \pi)^{n-y} + n^2(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho) \sum_{y=1}^{n-1} y^2 \frac{n!}{(n-y)!y!} \pi^y (1 - \pi)^{n-y} + n^2(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho)n\pi \sum_{y=1}^{n-1} y \binom{n-1}{y-1} \pi^{y-1} (1 - \pi)^{(n-1)-(y-1)} + n^2(\rho\pi + (1 - \rho)\pi^n) \end{aligned}$$

(Substitution $l := y - 1$)

$$\begin{aligned} &= (1 - \rho)n\pi \sum_{l=0}^{n-2} (l+1) \binom{n-1}{l} \pi^l (1 - \pi)^{n-1-l} + n^2(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho)n\pi \left[\sum_{l=1}^{n-2} l \binom{n-1}{l} \pi^l (1 - \pi)^{n-1-l} + \sum_{l=0}^{n-2} \binom{n-1}{l} \pi^l (1 - \pi)^{n-1-l} \right] \\ &\quad + n^2(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho)n\pi \left[\sum_{l=1}^{n-2} l \binom{n-1}{l} \pi^l (1 - \pi)^{n-1-l} + \underbrace{\sum_{l=0}^{n-1} \binom{n-1}{l} \pi^l (1 - \pi)^{n-1-l} - \pi^{n-1}}_{=1 \text{ (Binomischer Lehrsatz)}} \right] \\ &\quad + n^2(\rho\pi + (1 - \rho)\pi^n) \\ &= (1 - \rho)n\pi \left[(n-1)\pi \sum_{l=1}^{n-2} \binom{n-2}{l-1} \pi^{l-1} (1 - \pi)^{(n-2)-(l-1)} + 1 - \pi^{n-1} \right] + n^2(\rho\pi + (1 - \rho)\pi^n) \end{aligned}$$

(Indexverschiebung in Laufindex l)

$$\begin{aligned}
 &= (1 - \rho)n\pi \left[(n - 1)\pi \sum_{k=0}^{n-3} \binom{n-2}{k} \pi^k (1 - \pi)^{(n-2)-k} + 1 - \pi^{n-1} \right] + n^2(\rho\pi + (1 - \rho)\pi^n) \\
 &= (1 - \rho)n\pi \left[(n - 1)\pi \left[\sum_{k=0}^{n-2} \binom{n-2}{k} \pi^k (1 - \pi)^{(n-2)-k} - \pi^{n-2} \right] + 1 - \pi^{n-1} \right] \\
 &\quad + n^2(\rho\pi + (1 - \rho)\pi^n)
 \end{aligned}$$

(Binomischer Lehrsatz)

$$\begin{aligned}
 &= (1 - \rho)n\pi \left[(n - 1)\pi(1 - \pi^{n-2}) + 1 - \pi^{n-1} \right] + n^2(\rho\pi + (1 - \rho)\pi^n) \\
 &= (1 - \rho) \left[n\pi^2(n - 1) + \pi n \right] + n^2\rho\pi.
 \end{aligned}$$

$$\begin{aligned}
 \text{Var}(Y) &= E(Y^2) - \underbrace{E(Y)^2}_{=n^2\pi^2} \\
 &= n\pi^2(n - 1) + \pi n - n\rho\pi^2(n - 1) - \pi\rho n + n^2\rho\pi - n^2\pi^2 \\
 &= \pi n - n\pi^2 + (n\pi - n\pi^2)(n\rho - \rho) \\
 &= n\pi(1 - \pi)(1 + (n - 1)\rho)
 \end{aligned}$$

A.2.2 Zerlegung der Varianz nach der 1. Ordnung

Gegeben seien zwei Zufallsvariablen X und Y . Es gilt

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2$$

(Satz über iterierte Erwartungen)

$$\begin{aligned}
 &= E(E(Y^2|X) - (E(E(Y|X)))^2) \\
 &= E(E(Y^2|X) - \underbrace{E(E(Y|X)^2)}_{=0} + E(E(Y|X)^2) - (E(E(Y|X)))^2) \\
 &= E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))
 \end{aligned}$$

A.2.3 Momente der A-posteriori-Verteilung von Π

A-posteriori-Parameter:

$$\alpha + Y, \beta + N - Y$$

A-posteriori-Erwartungswert:

$$E(\Pi^{post}) = \frac{\alpha + Y}{\beta + N} \approx \frac{Y}{N} \quad (= \hat{\pi}_{ML})$$

A-posteriori-Varianz:

$$\begin{aligned} \text{Var}(\Pi^{post}) &= \frac{(\alpha + Y)(\beta + N - Y)}{(\alpha + \beta + N)^2(\alpha + \beta + N - 1)} \xrightarrow{\text{gro\ss e } N} \frac{Y(N - Y)}{N^3} \\ &= \frac{1}{N} \cdot \frac{Y}{N} \left(1 - \frac{Y}{N}\right) \\ &= \frac{1}{N} \cdot \hat{\pi}_{ML} (1 - \hat{\pi}_{ML}) \\ &\approx \text{Var}\left(\frac{\alpha + Y}{\beta + N}\right) \quad \left(= E(\Pi^{post})\right) \end{aligned}$$

A.3 SAS Quellcode

A.3.1 Simulationen

SAS Quellcode für Kapitel 4 (Teil-Simulation Modell mit pauschaler Klumpenkorrelation)

```
1 /*Fallzahlensimulation mit Zufallszahlen aus Proc RandMultinomial*/
2 /*vorgegebenes Rho, Erfolgswahrscheinlichkeit berechnet über CCM-
   Verteilung*/
3
4
5 data n; /*Initialisierung n-Vektor = Anzahl Individuen
   in Cluster i=1..k */
6 input n_i;
7 cards;
8 50
9 59
10 42
11 50
12 33
13 67
14 61
15 58
16 52
```

Anhang

17	40
18	45
19	50
20	58
21	42
22	41
23	41
24	41
25	44
26	22
27	24
28	57
29	61
30	63
31	23
32	29
33	95
34	27
35	39
36	47
37	43
38	21
39	23
40	10
41	60
42	90
43	64
44	24
45	57
46	31
47	37
48	59
49	47
50	40
51	52
52	50
53	32
54	58
55	107
56	21
57	61
58	40
59	34
60	68
61	51
62	42
63	65
64	49
65	36
66	34

Anhang

67	73
68	35
69	58
70	42
71	19
72	20
73	60
74	79
75	48
76	48
77	18
78	53
79	45
80	73
81	25
82	60
83	60
84	40
85	55
86	39
87	68
88	49
89	63
90	59
91	48
92	43
93	58
94	75
95	50
96	46
97	30
98	42
99	72
100	51
101	30
102	67
103	58
104	59
105	43
106	37
107	60
108	72
109	23
110	70
111	56
112	51
113	45
114	38
115	64
116	36

Anhang

117	64
118	44
119	62
120	25
121	53
122	40
123	11
124	29
125	31
126	40
127	66
128	43
129	26
130	43
131	17
132	75
133	33
134	29
135	65
136	42
137	26
138	38
139	50
140	38
141	12
142	29
143	59
144	52
145	54
146	60
147	68
148	59
149	68
150	32
151	66
152	54
153	48
154	51
155	56
156	71
157	55
158	18
159	35
160	64
161	43
162	53
163	13
164	58
165	41
166	57

Anhang

```
167 62
168 39
169 23
170 61
171 15
172 35
173 46
174 19
175 55
176 45
177 32
178 ;
179 run;
180
181
182 Proc IML;
183
184 use n;                /*n-Vektor übernehmen in IML*/
185 show datasets;
186 show contents;
187 read all into n;
188
189 /*****
190     *****/
191 pi = 0.38;            /*vorgegebener Wert für Pi (Zou-Verteilung)*/
192     /
193 rho = 0.03;          /*vorgegebener Wert für Rho (Zou-Verteilung)*/
194     */
195 v_rho = do(0,1,0.005) `;    /*Initialisierung Träger-Vektor für
196     Likelihoodwerte*/
197 loops = 10000;        /*Anzahl Schleifendurchläufe*/
198 /*****
199     *****/
200 *n=j(170,1,50);
201 k=nrow(n);           /*k = Anzahl Cluster*/
202 y=j(k,1,0);          /*Initialisierung y-Vektor = Anzahl der
203     Erfolge in Cluster i=1..k */
204 rho_EQL = j(loops,1,0);    /*Initialisierung Schätzer-Vektor*/
205 rho_PL = j(loops,1,0);    /*Initialisierung Schätzer-Vektor*/
206 rho_ML = j(loops,1,0);    /*Initialisierung Schätzer-Vektor*/
207 rho_MMB = j(loops,1,0);   /*Initialisierung Schätzer-Vektor*/
208 rho_anova = j(loops,1,0); /*Initialisierung Schätzer-Vektor*/
209 rho_PEQ = j(loops,1,0);   /*Initialisierung Schätzer-Vektor*/
210 rho_K = j(loops,1,0);     /*Initialisierung Schätzer-Vektor*/
211 rho_WEP = j(loops,1,0);   /*Initialisierung Schätzer-Vektor*/
212 rho_bbm = j(loops,1,0);   /*Initialisierung Schätzer-Vektor*/
213 v_PL = j(nrow(v_rho),1,0);
214 v_EQL = j(nrow(v_rho),1,0);
```

Anhang

```
211 v_ML = j(nrow(v_rho),1,0);
212 v_MMB = j(nrow(v_rho),1,0);
213 sim_var_pl = 0; /*Initialisierung Simulationsvarianz*/
214 sim_var_eql = 0; /*Initialisierung Simulationsvarianz*/
215 sim_var_mmb = 0; /*Initialisierung Simulationsvarianz*/
216 sim_var_anova = 0; /*Initialisierung Simulationsvarianz*/
217 sim_var_peq = 0; /*Initialisierung Simulationsvarianz*/
218 sim_var_k = 0; /*Initialisierung Simulationsvarianz*/
219 sim_var_wep = 0; /*Initialisierung Simulationsvarianz*/
220 sim_var_bbm = 0; /*Initialisierung Simulationsvarianz*/
221
222 start F(y_dummy,p,rho,pi,n_dummy); /*Wahrscheinlichkeitsfunktion
    definieren (s. Zou(2004))*/
223 if y_dummy=0 then p = rho * (1-pi) + (1-rho) * ((1-pi)**n_dummy);
224 else if y_dummy=n_dummy then p = rho * pi + (1-rho) * (pi**n_dummy);
225 else p = comb(n_dummy,y_dummy) * (1-rho) * (pi**y_dummy) * ((1-pi)**(n_
    dummy-y_dummy));
226 finish;
227
228 /*****
    *****/
229 /*Beginn Schleife für Anzahl Simulationen*/
230 do z = 1 to loops;
231     do i = 1 to k; /*Schleife für alle Cluster 1..k */
232         prob = j(n[i]+1,1,0); /*Vektor der
            Erfolgswahrscheinlichkeiten für 0..n[i] Erfolge*/
233
234         do j = 1 to n[i]+1; /*Schleife für mgl. Anzahlen von
            Erfolgen, Achtung: verschoben um +1, da in SAS 0 nicht mgl*/
235             run F(j-1,p,rho,pi,n[i]); /*Aufruf Funktion F mit Ausgabe p=
                Wahrscheinlichkeit*/
236             prob[j] = p;
237         end;
238
239         %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\rand.sas';
240         *print (n[i]); *print prob; *print (nrow(prob));
241     end;
242
243 /*****
    *****/
244 /*Berechnung Schätzer für rho*/
245
246 /*Schätzer, die modellunabhängig funktionieren*/
247 /*Weighted Empirical Pairwise*/
248 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\WEP.sas';
249
250 /*Kappa*/
251 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\K.sas';
252
253 /*Pairwise Equal Weights*/
```

Anhang

```
254 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\PEQ.sas';
255
256 /*ANOVA*/
257 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\anova.sas';
258
259 /*Beta Binomial Model*/
260 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\bbm.sas';
261
262
263 /*****
264 *****/
265 /*Schätzer, die auf Zou-Verteilung basieren (keine geschlossene Form)*/
266
267 /*Maximum Likelihood*/
268 *%include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\ml.sas';
269
270 /*Pseudo-Likelihood*/
271 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\pl.sas';
272
273 /*Extended Quasi-Likelihood*/
274 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\eqL.sas';
275
276 /*Moment Method Based (Cox/Snell)*/
277 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou\mmb.sas';
278
279 /*Speichern der Schätzer*/
280 rho_Matrix = j(loops,9,0); /*Initialisierung Schätzer-
281 Vektor*/
282 rho_Matrix[,1] = rho_WEP;
283 rho_Matrix[,2] = rho_K;
284 rho_Matrix[,3] = rho_PEQ;
285 rho_Matrix[,4] = rho_anova;
286 rho_Matrix[,5] = rho_ML;
287 rho_Matrix[,6] = rho_BBM;
288 rho_Matrix[,7] = rho_EQL;
289 rho_Matrix[,8] = rho_PL;
290 rho_Matrix[,9] = rho_MMB;
291
292 end;
293 /*Ende Schleife für Anzahl Simulationen*/
294
295 /*Mittelwert der simulierten Schätzer*/
296 rho_EQL_E = 1/loops * j(1,loops,1) * rho_eqL;
297 rho_PL_E = 1/loops * j(1,loops,1) * rho_pl;
298 rho_ML_E = 1/loops * j(1,loops,1) * rho_ml;
299 rho_Mmb_E = 1/loops * j(1,loops,1) * rho_mmb;
300 rho_anova_E = 1/loops * j(1,loops,1) * rho_anova;
301 rho_peq_E = 1/loops * j(1,loops,1) * rho_peq;
302 rho_k_E = 1/loops * j(1,loops,1) * rho_k;
303 rho_WEP_E = 1/loops * j(1,loops,1) * rho_WEP;
```

Anhang

```
302 rho_bbm_E = 1/loops * j(1,loops,1) * rho_bbm;
303
304 /*Varianzen der simulierten Schätzer*/
305 sim_var_pl = 1/loops * sim_var_pl;
306 sim_var_eql = 1/loops * sim_var_eql;
307 sim_var_mmb = 1/loops * sim_var_mmb;
308 sim_var_anova = 1/loops * sim_var_anova;
309 sim_var_peq = 1/loops * sim_var_peq;
310 sim_var_k = 1/loops * sim_var_k;
311 sim_var_wep = 1/loops * sim_var_wep;
312 sim_var_bbm = 1/loops * sim_var_bbm;
313
314 print pi;
315 print rho;
316 print pi_sim_gesamt;
317 print rho_pl_e;
318 print rho_eql_e;
319 print rho_ml_e;
320 print rho_mmb_e;
321 print rho_anova_e;
322 print rho_peq_e;
323 print rho_k_e;
324 print rho_WEP_e;
325 print rho_bbm_e;
326 print sim_var_pl;
327 print sim_var_eql;
328 print sim_var_mmb;
329 print sim_var_anova;
330 print sim_var_peq;
331 print sim_var_k;
332 print sim_var_wep;
333 print sim_var_bbm;
334
335 create v_PL from v_PL;                               /* Matrix in Datensatz
      schreiben */
336 append from v_PL;
337 close v_PL;
338 create v_EQL from v_EQL;                             /* Matrix in Datensatz
      schreiben */
339 append from v_EQL;
340 close v_EQL;
341 create v_ML from v_ML;                               /* Matrix in Datensatz
      schreiben */
342 append from v_ML;
343 close v_ML;
344 create v_rho from v_rho;                             /* Matrix in Datensatz
      schreiben */
345 append from v_rho;
346 close v_rho;
347
```

Anhang

```
348 create y_i from y; /* Matrix in Datensatz schreiben
    */
349 append from y;
350 close y_i;
351 create rho_Matrix from rho_Matrix; /* Matrix in
    Datensatz schreiben */
352 append from rho_Matrix;
353 close rho_Matrix;
354
355 quit;
356
357
358
359 /* Plots (Werte von Rho gegen entspr. Likelihood */
360
361 libname a 'R:\AD\Forschung\Programme\Rho_Schaetzer_Zou';
362
363 data a.y_i;
364 set y_i;
365 run;
366
367 data a.rho_Matrix;
368 set rho_Matrix;
369 run;
370
371 title1 ' ';
372
373 proc format;
374 value smoothmeth 1='RoT' 2='CV' 3='PI (ROT)';
375 run;
376
377 goptions reset=all;
378 goptions ctext=bl ftext=swiss1 htext=2;
379
380 symbol1 color=black line=1 v = none i = spline w = 1 ;
381 symbol2 color=black line=2 v = none i = spline w = 2 ;
382 symbol3 color=black line=5 v = none i = spline w = 2 ;
383 symbol4 color=black line=41 v = none i = spline w = 1 ;
384 symbol5 color=black line=5 v = none i = spline w = 1 ;
385
386
387 axis1 order=(0 to 1 by 0.05)
388 label=(f=greek "r" justify=right h=2 );
389 axis2
390 label=(f=swiss h=2 justify=left "Pseudo-" justify=left "Likelihood")
    ;
391 axis3
392 label=(f=swiss h=2 justify=left "Extended Quasi-" justify=left "
    Likelihood");
393 axis4
```

Anhang

```
394         label=(f=swiss h=2 justify=left "Maximum" justify=left "Likelihood")
395             ;
396 data hilf1 (RENAME=(COL1=COL2));
397 set v_PL;
398 run;
399 data PL;
400 merge v_rho hilf1;
401 run;
402 data hilf2 (RENAME=(COL1=COL2));
403 set v_EQL;
404 run;
405 data EQL;
406 merge v_rho hilf2;
407 run;
408 data hilf3 (RENAME=(COL1=COL2));
409 set v_ML;
410 run;
411 data ML;
412 merge v_rho hilf3;
413 run;
414
415
416 proc gplot data=PL;
417 *label kurve='NN selector';
418 *format kurve smoothmeth.;
419 plot col2*col1 /*=kurve*/ / vaxis=axis2 haxis=axis1 /* overlay frame */;
420 run;
421 quit;
422 proc gplot data=EQL;
423 *label kurve='NN selector';
424 *format kurve smoothmeth.;
425 plot col2*col1 /*=kurve*/ / vaxis=axis3 haxis=axis1 /* overlay frame */;
426 run;
427 quit;
428 proc gplot data=ML;
429 *label kurve='NN selector';
430 *format kurve smoothmeth.;
431 plot col2*col1 /*=kurve*/ / vaxis=axis4 haxis=axis1 /* overlay frame */;
432 run;
433 quit;
```

SAS Quellcode für Prozedur ML

```
1  /*Maximum Likelihood*/
2
3  do j = 1 to nrow(v_rho)-1;                               /*bei v_rho=1 -> log(0)
4     deshalb -1*/
5     l = 0;
```

Anhang

```
6 do i=1 to k;
7   l_hilf = 0;
8   if y[i]=0 then                                /*Fallunterscheidung da
9     sonst Div. durch Null*/
10    l_hilf = log(v_rho[j]*(1-pi) + (1-v_rho[j]) * (1-pi)**n[i]);    /*vgl.
11      Formel in Zou*/
12    else if y[i]=n[i] then
13      l_hilf = log(v_rho[j] * pi + (1-v_rho[j]) * pi ** n[i]);
14    else l_hilf = log( comb(n[i],y[i]) * (1-v_rho[j]) * (pi**y[i]) * (1-pi
15      )** (n[i]-y[i]) );
16    l = l + l_hilf;
17  end;
18  v_ML[j] = l;
19 end;
20 /*Finde Rho*/
21 ML_max = max(v_ML);                                /*finde max. Likelihoodeintrag*/
22 ML = v_rho||v_ML;                                /*Vektoren nebeneinanderschreiben*/
23 do i=1 to nrow(v_rho);
24   Likelihoodeintrag*/                                /*finde entspr. rho zum max.
25   if ml[i,2] = ml_max then rho_ml[z] = ml[i,1];
26 end;
```

SAS Quellcode für Prozedur EQL

```
1 /*Extended Quasi-Likelihood*/
2
3 do j = 1 to nrow(v_rho);
4   l = 0;
5
6   do i=1 to k;
7     l_hilf = 0;
8     if y[i]=0 then                                /*Fallunterscheidung da
9       sonst Div. durch Null*/
10      D = 2 * ((n[i] - y[i]) * log10( (n[i] - y[i]) / (n[i] - n[i]*pi)));
11        /*vgl. Formel in Zou*/
12      else if y[i]=n[i] then
13        D = 2 * (y[i] * log10(y[i] / (n[i] * pi)));
14      else D = 2 * (y[i] * log10( y[i]/(n[i] * pi)) + (n[i] - y[i]) * log10(
15        (n[i] - y[i]) / (n[i] - n[i]*pi)) );
16      l_hilf = -0.5 * (log10(1 + (n[i] - 1) * v_rho[j]) + D / (1 + (n[i] -
17        1) * v_rho[j]));
18      l = l + l_hilf;
19    end;
20    *print l;
21    v_EQL[j] = l;
22  end;
23 end;
24 /*Finde Rho*/
```

Anhang

```
22 EQL_max = max(v_EQL);          /*finde max. Likelihoodbeitrag v.
    EQL*/
23 EQL = v_rho||v_EQL;
24 do i=1 to nrow(v_rho);        /*finde entspr. rho zum max.
    Likelihoodbeitrag*/
25   if eql[i,2] = eql_max then rho_eql[z] = eql[i,1];
26 end;
27
28 sim_var_eql = sim_var_eql + (rho - rho_eql[z]) ** 2;      /*
    Fehlerquadrate*/
```

SAS Quellcode für Prozedur PL

```
1   /*Pseudo-Likelihood*/
2
3   do j=1 to nrow(v_rho);
4     l = 0;
5
6     do i = 1 to k;
7       l_hilf = 0;
8       l_hilf = -0.5 * (log10(1 + (n[i] - 1) * v_rho[j]) + ((y[i] - n[i] *
9         pi)**2) / (n[i] * pi * (1 - pi) * (1 + (n[i] - 1) * v_rho[j])));
10      l = l + l_hilf;
11    end;
12    v_PL[j] = l;
13  end;
14  /*Finde Rho*/
15  PL_max = max(v_PL);          /*finde max. Likelihoodbeitrag*/
16  PL = v_rho||v_PL;          /*Vektoren nebeneinanderschreiben*/
17  do i=1 to nrow(v_rho);      /*finde entspr. rho zum max.
    Likelihoodbeitrag*/
18    if pl[i,2] = pl_max then rho_pl[z] = pl[i,1];
19  end;
20
21  sim_var_pl = sim_var_pl + (rho - rho_pl[z]) ** 2;      /*Fehlerquadrate*/
```

SAS Quellcode für Prozedur DBB

```
1   /*Direkter Beta-Binomial-Modell-Schätzer*/
2
3   pi_sim_gesamt = j(1,k,1) * y / (j(1,k,1) * n);      /*Mittelwert der pi
    (Klumpenstruktur nicht beachtet)*/
4
5   /*95% credible intervall*/
6   ogrenze = round(k * 0.975);
7   ugrenze = round(k * 0.025);
8
9   pi_i = j(k,1,1);
10  do i=1 to k;
```

Anhang

```
11     pi_i[i] = y[i] / n[i];
12     end;
13
14     call sortndx(idpi,pi_i,1);
15     pi_i_sort = pi_i[idpi, ];           /*pi_i_sort ist die sortierte
16     Variante von pi_i*/
17
18     aopt = pi_i_sort[ugrenze];
19     bopt = pi_i_sort[ogrenze];
20     dmin = 100;
21
22     do alpha_hilf = 0.2 to 100 by 0.1;           /*siehe auch
23     bayesianischesverteilungsmodell.R und Vortrag Dortmund*/
24     beta_hilf = alpha_hilf * (1-pi_sim_gesamt) / pi_sim_gesamt; /*hier das
25     pi aus den simulierten Daten geschätzt*/
26     a = betainv(0.025,alpha_hilf,beta_hilf);
27     b = betainv(0.975,alpha_hilf,beta_hilf);
28     d = (a-aopt)**2 + (b-bopt)**2;
29
30     if (d<dmin) then do;
31         dmin = d;
32         alphamin = alpha_hilf;
33         betamin = beta_hilf;
34     end;
35 end;
36 rho_BBM[z] = 1 / (1+alphamin + betamin);
37
38 sim_var_bbm = sim_var_bbm + (rho - rho_bbm[z]) ** 2;           /*
39 Fehlerquadrate*/
```

SAS Quellcode für Prozedur WEP

```
1     /*Weighted Empirical Pairwise*/
2
3     pi_sim_gesamt = j(1,k,1) * y / (j(1,k,1) * n);           /*Mittelwert der pi (
4     Klumpenstruktur nicht beachtet)*/
5
6     F = 0;           /*Anzahl aller möglichen 2er Kombinationen
7     (Paaren)*/
8     start = 1;           /*Startzeile*/
9     E = 0;           /*Erwartungswert E(X_ij,X_ij')*/
10    do i = 1 to k;
11        if (y[i] = 1) | (y[i] = 0)
12            then hilf2 = 0;
13            else hilf2 = comb(y[i],2);           /*Anzahl aller möglichen 2er
14            Kombinationen (Erfolg) in Klumpen i*/
15        E = E + hilf2;
16        hilf3 = comb(n[i],2);           /*Anzahl aller möglichen 2er
17        Kombinationen (Paaren)*/
18        F = F + hilf3;
```

Anhang

```
15 end;
16 rho_WEP[z,1] = ((E/F) - pi_sim_gesamt**2) / (pi_sim_gesamt*(1-pi_sim_
    gesamt)); /*Schätzer WEP*/
17
18 sim_var_wep = sim_var_wep + (rho - rho_wep[z,1]) ** 2; /*
    Fehlerquadrate*/
```

SAS Quellcode für Prozedur K

```
1 /*Kappa*/
2
3 pi_sim_gesamt = j(1,k,1) * y / (j(1,k,1) * n); /*Mittelwert der pi (
    Klumpenstruktur nicht beachtet)*/
4
5 hilf2 = 0;
6 do i = 1 to k;
7     hilf2 = hilf2 + (y[i]**2 + (n[i] - y[i])**2 - n[i]) / (n[i] * (n[i] -
    1));
8 end;
9 rho_K[z,1] = (1/k * hilf2 - pi_sim_gesamt**2 - (1-pi_sim_gesamt)**2) /
    (2 * pi_sim_gesamt - 2*pi_sim_gesamt**2);
10
11 sim_var_k = sim_var_k + (rho - rho_k[z,1]) ** 2; /*Fehlerquadrate*/
```

SAS Quellcode für Prozedur PEQ

```
1 /*Pairwise Equal Weights*/
2
3 hilf2 = 0;
4 do i = 1 to k;
5     hilf2 = hilf2 + y[i] * (y[i] - 1);
6 end;
7 hilf3 = 0;
8 do i = 1 to k;
9     hilf3 = hilf3 + y[i] * (n[i] - 1);
10 end;
11 hilf4 = 0;
12 do i = 1 to k;
13     hilf4 = hilf4 + n[i] * (n[i] - 1);
14 end;
15 hilf5 = hilf3 / hilf4;
16 rho_PEQ[z,1] = 1/(hilf5 * (1 - hilf5)) * (hilf2 / hilf4 - hilf5**2);
17
18 sim_var_peq = sim_var_peq + (rho - rho_peq[z,1]) ** 2; /*
    Fehlerquadrate*/
```

SAS Quellcode für Prozedur ANOVA

```
1 /*ANOVA*/
2
```

Anhang

```
3  hilf = 0;
4  do i = 1 to k;
5      hilf = hilf + ((y[i]**2) / n[i]);
6  end;
7  sqa = (1 / (k-1)) * (hilf - (1/(j(1,k,1) * n) * ((j(k,1,1)` * y)**2)));
      /*s. Ridout*/
8  sqr = (1 / ((j(1,k,1) * n) - k)) * ((j(k,1,1)` * y) - hilf);
9  n_0 = (1 / (k-1)) * ((j(1,k,1) * n) - 1/(j(1,k,1) * n) * n` * n) ;
10 rho_anova[z] = (sqa - sqr) / (sqa + (n_0 - 1) * sqr);
11
12 sim_var_anova = sim_var_anova + (rho - rho_anova[z]) ** 2;      /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur MMB

```
1 /*Moment Method Based*/
2
3 pi_sim_gesamt = j(1,k,1) * y / (j(1,k,1) * n);      /*Mittelwert der pi (
      Klumpenstruktur nicht beachtet)*/
4
5 do j=1 to nrow(v_rho);      /*Schleife für Likelihoodwerte zu
      entspr. Rho*/
6     l = 0;      /*"Likelihood"*/
7
8     do i = 1 to k;      /*Schleife für Summenbildung*/
9         l_hilf = 0;
10        l_hilf = ((y[i] - n[i] * pi_sim_gesamt)**2) / (n[i] * pi_sim_gesamt
            * (1 - pi_sim_gesamt) * (1 + (n[i] + 1) * v_rho[j]));
11        l = l + l_hilf;
12    end;
13    v_MMB[j] = l;
14 end;
15
16 /*Finde Rho*/
17 **rho_MMB[M] = 0;
18 MMB_diff = (v_MMB - k);
19 MMB_min = min(abs(MMB_diff));
20 *print MMB_min; *print MMB_diff;
21 MMB = v_rho || MMB_diff;      /*Vektoren
      nebeneinanderschreiben*/
22 do i = 1 to nrow(v_rho);
23     if MMB[i,2] = MMB_min then rho_MMB[z] = MMB[i,1];
24 end;
25
26 sim_var_mmb = sim_var_mmb + (rho - rho_mmb[z]) ** 2;      /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur rand

```
1 /*Zufallszahl für Anzahl der Erfolge mit Vektor der Wahrscheinlichkeiten
      für bis zu 100 Einträge im Vektor der Wahrscheinlichkeiten prob*/
```

Anhang

```
2
3 l_prob = nrow(prob);
4
5
6 if l_prob = 1 then
7   y[i] = rand('table',prob[1]) - 1;
8   else if l_prob = 2 then
9     y[i] = rand('table',prob[1],prob[2]) - 1;
10  else if l_prob = 3 then
11    y[i] = rand('table',prob[1],prob[2],prob[3]) - 1;
12  else if l_prob = 4 then
13    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4]) - 1;
14  else if l_prob = 5 then
15    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5]) - 1;
16  else if l_prob = 6 then
17    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6]) -
18      1;
19  else if l_prob = 7 then
20    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
21      [7]) - 1;
22  else if l_prob = 8 then
23    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
24      [7],prob[8]) - 1;
25  else if l_prob = 9 then
26    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
27      [7],prob[8],prob[9]) - 1;
28  else if l_prob = 10 then
29    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
30      [7],prob[8],prob[9],prob[10]) - 1;
31  else if l_prob = 11 then
32    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
33      [7],prob[8],prob[9],prob[10],
34      prob[11]) - 1;
35  else if l_prob = 12 then
36    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
37      [7],prob[8],prob[9],prob[10],
38      prob[11],prob[12]) - 1;
39  else if l_prob = 13 then
40    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
41      [7],prob[8],prob[9],prob[10],
42      prob[11],prob[12],prob[13]) - 1;
43  else if l_prob = 14 then
44    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
45      [7],prob[8],prob[9],prob[10],
46      prob[11],prob[12],prob[13],prob[14]) - 1;
47  else if l_prob = 15 then
48    y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
49      [7],prob[8],prob[9],prob[10],
50      prob[11],prob[12],prob[13],prob[14],prob[15]) - 1;
51  else if l_prob = 16 then
```

Anhang

```
42 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
43 [7],prob[8],prob[9],prob[10],
44 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16]) - 1;
45 else if l_prob = 17 then
46 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
47 [7],prob[8],prob[9],prob[10],
48 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17]) - 1;
49 else if l_prob = 18 then
50 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
51 [7],prob[8],prob[9],prob[10],
52 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
53 [18],prob[19]) - 1;
54 else if l_prob = 20 then
55 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
56 [7],prob[8],prob[9],prob[10],
57 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
58 [18],prob[19],prob[20]) - 1;
59 else if l_prob = 21 then
60 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
61 [7],prob[8],prob[9],prob[10],
62 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
63 [18],prob[19],prob[20],
64 prob[21]) - 1;
65 else if l_prob = 22 then
66 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
67 [7],prob[8],prob[9],prob[10],
68 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
69 [18],prob[19],prob[20],
70 prob[21],prob[22]) - 1;
71 else if l_prob = 23 then
72 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
73 [7],prob[8],prob[9],prob[10],
74 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
75 [18],prob[19],prob[20],
76 prob[21],prob[22],prob[23]) - 1;
77 else if l_prob = 24 then
78 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
79 [7],prob[8],prob[9],prob[10],
80 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
81 [18],prob[19],prob[20],
82 prob[21],prob[22],prob[23],prob[24]) - 1;
83 else if l_prob = 25 then
84 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
85 [7],prob[8],prob[9],prob[10],
86 prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
87 [18],prob[19],prob[20],
88 prob[21],prob[22],prob[23],prob[24],prob[25]) - 1;
```

Anhang

```
74     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
75     [18],prob[19],prob[20],
76     prob[21],prob[22],prob[23],prob[24],prob[25]) - 1;
77 else if l_prob = 26 then
78 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
79     [7],prob[8],prob[9],prob[10],
80     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
81     [18],prob[19],prob[20],
82     prob[21],prob[22],prob[23],prob[24],prob[25],prob[26]) - 1;
83 else if l_prob = 27 then
84 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
85     [7],prob[8],prob[9],prob[10],
86     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
87     [18],prob[19],prob[20],
88     prob[21],prob[22],prob[23],prob[24],prob[25],prob[26],prob[27]) - 1;
89 else if l_prob = 28 then
90 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
91     [7],prob[8],prob[9],prob[10],
92     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
93     [18],prob[19],prob[20],
94     prob[21],prob[22],prob[23],prob[24],prob[25],prob[26],prob[27],prob
95     [28]) - 1;
96 else if l_prob = 29 then
97 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
98     [7],prob[8],prob[9],prob[10],
99     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
100    [18],prob[19],prob[20],
101    prob[21],prob[22],prob[23],prob[24],prob[25],prob[26],prob[27],prob
102    [28],prob[29],prob[30],
103    prob[31],prob[32],prob[33],prob[34],prob[35],prob[36],prob[37],prob
104    [38],prob[39],prob[40],
105    prob[41],prob[42],prob[43],prob[44],prob[45],prob[46],prob[47],prob
    [48],prob[49],prob[50],
    prob[51],prob[52],prob[53],prob[54],prob[55],prob[56],prob[57],prob
    [58],prob[59],prob[60],
```

Anhang

```
106     prob[61],prob[62],prob[63],prob[64],prob[65],prob[66],prob[67],prob
107     [68],prob[69],prob[70],
108     prob[71],prob[72],prob[73],prob[74],prob[75],prob[76],prob[77],prob
109     [78],prob[79],prob[80],
110     prob[81],prob[82],prob[83],prob[84],prob[85],prob[86],prob[87],prob
111     [88],prob[89],prob[90],
112     prob[91],prob[92],prob[93],prob[94],prob[95],prob[96],prob[97],prob
113     [98]) - 1;
114 else if l_prob = 99 then
115 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
116     [7],prob[8],prob[9],prob[10],
117     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
118     [18],prob[19],prob[20],
119     prob[21],prob[22],prob[23],prob[24],prob[25],prob[26],prob[27],prob
120     [28],prob[29],prob[30],
121     prob[31],prob[32],prob[33],prob[34],prob[35],prob[36],prob[37],prob
122     [38],prob[39],prob[40],
123     prob[41],prob[42],prob[43],prob[44],prob[45],prob[46],prob[47],prob
124     [48],prob[49],prob[50],
125     prob[51],prob[52],prob[53],prob[54],prob[55],prob[56],prob[57],prob
126     [58],prob[59],prob[60],
127     prob[61],prob[62],prob[63],prob[64],prob[65],prob[66],prob[67],prob
128     [68],prob[69],prob[70],
129     prob[71],prob[72],prob[73],prob[74],prob[75],prob[76],prob[77],prob
130     [78],prob[79],prob[80],
131     prob[81],prob[82],prob[83],prob[84],prob[85],prob[86],prob[87],prob
132     [88],prob[89],prob[90],
133     prob[91],prob[92],prob[93],prob[94],prob[95],prob[96],prob[97],prob
134     [98],prob[99]) - 1;
135 else if l_prob = 100 then
136 y[i] = rand('table',prob[1],prob[2],prob[3],prob[4],prob[5],prob[6],prob
137     [7],prob[8],prob[9],prob[10],
138     prob[11],prob[12],prob[13],prob[14],prob[15],prob[16],prob[17],prob
139     [18],prob[19],prob[20],
140     prob[21],prob[22],prob[23],prob[24],prob[25],prob[26],prob[27],prob
141     [28],prob[29],prob[30],
142     prob[31],prob[32],prob[33],prob[34],prob[35],prob[36],prob[37],prob
143     [38],prob[39],prob[40],
144     prob[41],prob[42],prob[43],prob[44],prob[45],prob[46],prob[47],prob
145     [48],prob[49],prob[50],
146     prob[51],prob[52],prob[53],prob[54],prob[55],prob[56],prob[57],prob
147     [58],prob[59],prob[60],
148     prob[61],prob[62],prob[63],prob[64],prob[65],prob[66],prob[67],prob
149     [68],prob[69],prob[70],
150     prob[71],prob[72],prob[73],prob[74],prob[75],prob[76],prob[77],prob
151     [78],prob[79],prob[80],
152     prob[81],prob[82],prob[83],prob[84],prob[85],prob[86],prob[87],prob
153     [88],prob[89],prob[90],
154     prob[91],prob[92],prob[93],prob[94],prob[95],prob[96],prob[97],prob
155     [98],prob[99],prob[100]) - 1;
```

Anhang

SAS Quellcode für Kapitel 4 (Teil-Simulation Beta-Binomial-Modell)

```
1
2 /*Fallzahlensimulation mit zufälligen Pi (Beta-verteilt)*/
3 /*Vorgabe: Rho = 0,03*/
4 /*1. Schätzer, die Modellunabhängig funktionieren*/
5 /*2. Schätzer, die auf CCM-Verteilung basieren*/
6
7
8 libname a 'R:\AD\Forschung\Daten\Zahnstudie';
9
10
11 data n;                                /*Initialisierung n-Vektor = Anzahl Individuen
12     in Cluster i=1..k */
13 input n_i;
14 cards;
15 50
16 59
17 42
18 50
19 33
20 67
21 61
22 58
23 52
24 40
25 45
26 50
27 58
28 42
29 41
30 41
31 41
32 44
33 22
34 24
35 57
36 61
37 63
38 23
39 29
40 95
41 27
42 39
43 47
44 43
45 21
46 23
47 10
48 60
49 90
```

Anhang

49	64
50	24
51	57
52	31
53	37
54	59
55	47
56	40
57	52
58	50
59	32
60	58
61	107
62	21
63	61
64	40
65	34
66	68
67	51
68	42
69	65
70	49
71	36
72	34
73	73
74	35
75	58
76	42
77	19
78	20
79	60
80	79
81	48
82	48
83	18
84	53
85	45
86	73
87	25
88	60
89	60
90	40
91	55
92	39
93	68
94	49
95	63
96	59
97	48
98	43

Anhang

99	58
100	75
101	50
102	46
103	30
104	42
105	72
106	51
107	30
108	67
109	58
110	59
111	43
112	37
113	60
114	72
115	23
116	70
117	56
118	51
119	45
120	38
121	64
122	36
123	64
124	44
125	62
126	25
127	53
128	40
129	11
130	29
131	31
132	40
133	66
134	43
135	26
136	43
137	17
138	75
139	33
140	29
141	65
142	42
143	26
144	38
145	50
146	38
147	12
148	29

Anhang

```
149 59
150 52
151 54
152 60
153 68
154 59
155 68
156 32
157 66
158 54
159 48
160 51
161 56
162 71
163 55
164 18
165 35
166 64
167 43
168 53
169 13
170 58
171 41
172 57
173 62
174 39
175 23
176 61
177 15
178 35
179 46
180 19
181 55
182 45
183 32
184 ;
185 run;
186
187
188 Proc IML;
189
190 use n;                /*n-Vektor übernehmen in IML*/
191 show datasets;
192 show contents;
193 read all into n_i;
194
195 /*****
      *****/
196 alpha = 1843 / 150;
197 beta = 3007 / 150;
```

Anhang

```
198 loops = 10000;                                /*Anzahl Simulationen*/
199 v_rho = do(0,1,0.005) `;                       /*Vektor für mögliche rho*/
200 /*****
      *****/
201
202 k = nrow(n_i);                                /*k = Anzahl Cluster*/
203 N = j(k,1,1) `*n_i;
204 pi = alpha / (alpha + beta);                  /*determiniertes Pi*/
205 rho = 1/(alpha + beta + 1);                   /*determiniertes Rho*/
206 rho_WEP = j(loops,1,0);                       /*Initialisierung Schätzer-Vektor*/
207 rho_K = j(loops,1,0);                        /*Initialisierung Schätzer-Vektor*/
208 rho_PEQ = j(loops,1,0);                      /*Initialisierung Schätzer-Vektor*/
209 rho_ANO = j(loops,1,0);                     /*Initialisierung Schätzer-Vektor*/
210 rho_EQL = j(loops,1,0);                     /*Initialisierung Schätzer-Vektor*/
211 rho_PL = j(loops,1,0);                      /*Initialisierung Schätzer-Vektor*/
212 rho_ML = j(loops,1,0);                      /*Initialisierung Schätzer-Vektor*/
213 rho_MMB = j(loops,1,0);                     /*Initialisierung Schätzer-Vektor*/
214 rho_BBM = j(loops,1,0);                     /*Initialisierung Schätzer-Vektor*/
215 v_EQL = j(nrow(v_rho),1,0);                  /*Initialisierung Vektor für
      Likelihoodwerte*/
216 v_PL = j(nrow(v_rho),1,0);                   /*Initialisierung Vektor für
      Likelihoodwerte*/
217 v_ML = j(nrow(v_rho),1,0);                   /*Initialisierung Vektor für
      Likelihoodwerte*/
218 v_MMB = j(nrow(v_rho),1,0);                  /*Initialisierung Vektor für "
      Likelihoodwerte"*/
219 sim_var_pl = 0;                              /*Initialisierung Simulationsvarianz*/
220 sim_var_eql = 0;                             /*Initialisierung Simulationsvarianz*/
221 sim_var_mmb = 0;                             /*Initialisierung Simulationsvarianz*/
222 sim_var_anova = 0;                          /*Initialisierung Simulationsvarianz*/
223 sim_var_peq = 0;                             /*Initialisierung Simulationsvarianz*/
224 sim_var_k = 0;                              /*Initialisierung Simulationsvarianz*/
225 sim_var_wep = 0;                             /*Initialisierung Simulationsvarianz*/
226 sim_var_bbm = 0;                             /*Initialisierung Simulationsvarianz*/
227
228 do M = 1 to loops;
229
230     pi_i=j(k,1,1);                            /*Initialisierung pi_i-Vektor =
      Erfolgswahrscheinlichkeit in Cluster i=1..k */
231     klumpen = j(N,2,0);
232
233     do i=1 to k;
234         pi_i[i] = rand('beta',alpha,beta);    /*Erzeuge k zufällige Pi (
      Betaverteilt)*/
235     end;
236
237     start = 1;
238     do i=1 to k;
239         do j=1 to n_i[i];
```

Anhang

```
240     klumpen[start,2] = rand('bernoulli',pi_i[i]); /*Erzeuge bernoulli-
        verteilte X_ij*/
241     klumpen[start,1] = i;           /*Spalte 1 -> Klumpennummer*/
242     start = start + 1;             /*Inkrement*/
243     end;
244 end;
245
246
247 /*****
        *****/
248
249 /*Schätzer, die modellunabhängig funktionieren*/
250 /*Weighted Empirical Pairwise*/
251 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\WEP.sas';
252
253 /*Kappa*/
254 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\K.sas';
255
256 /*Pairwise Equal Weights*/
257 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\PEQ.sas';
258
259 /*ANOVA*/
260 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\anova.sas';
261
262 /*Beta Binomial Model*/
263 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\BBM.sas';
264
265 /*****
        *****/
266 /*Schätzer, die auf CCM-Verteilung basieren*/
267 /*Extended Quasi Likelihood*/
268 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\eqL.sas';
269
270 /*Pseudo Likelihood*/
271 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\pl.sas';
272
273 /*Maximum Likelihood*/
274 *%include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\ml.sas';
275
276 /*Moment Method Based (Cox/Snell)*/
277 %include 'R:\AD\Forschung\Programme\Rho_Schaetzer_Betavert\mmb.sas';
278
279
280
281
282
283 /*****
        *****/
284
```

Anhang

```
285 *pi_sim_gesamt = j(N,1,1) ` * klumpen[,2] / N;      /*Mittelwert der pi (
      Klumpenstruktur nicht beachtet)*/
286 pi_sim_klump = (j(k,1,1) ` * (y_i / n_i)) / k ;      /*Mittelwert der
      pi (Klumpenstruktur beachtet)*/
287 pi_quer = (j(k,1,1) ` * (pi_i)) / k;      /*Mittelwert vor Y_i
      Simulation*/
288
289
290 /*Speichern der Schätzer*/
291 rho_Matrix = j(loops,9,0);      /*Initialisierung Schätzer-
      Vektor*/
292 rho_Matrix[,1] = rho_WEP;
293 rho_Matrix[,2] = rho_K;
294 rho_Matrix[,3] = rho_PEQ;
295 rho_Matrix[,4] = rho_ano;
296 rho_Matrix[,5] = rho_ML;
297 rho_Matrix[,6] = rho_BBM;
298 rho_Matrix[,7] = rho_EQL;
299 rho_Matrix[,8] = rho_PL;
300 rho_Matrix[,9] = rho_MMB;
301
302 end;
303
304 rho_WEP_E = 1/loops * j(1,loops,1) * rho_WEP;      /*Mittelwert der
      simulierten Schätzer*/
305 rho_WEP_V = 1/(loops-1) * ((j(loops,1,1) * rho) - rho_WEP) ` * ((j(loops
      ,1,1) * rho) - rho_WEP); /*Simulationsvarianz*/
306
307 rho_K_E = 1/loops * j(1,loops,1) * rho_K;
308 rho_PEQ_E = 1/loops * j(1,loops,1) * rho_PEQ;
309 rho_ANO_E = 1/loops * j(1,loops,1) * rho_ANO;
310 rho_EQL_E = 1/loops * j(1,loops,1) * rho_EQL;
311 rho_PL_E = 1/loops * j(1,loops,1) * rho_PL;
312 rho_ML_E = 1/loops * j(1,loops,1) * rho_ML;
313 rho_MMB_E = 1/loops * j(1,loops,1) * rho_MMB;
314 rho_BBM_E = 1/loops * j(1,loops,1) * rho_BBM;
315
316 sim_var_pl = 1/loops * sim_var_pl;
317 sim_var_eql = 1/loops * sim_var_eql;
318 sim_var_mmb = 1/loops * sim_var_mmb;
319 sim_var_anova = 1/loops * sim_var_anova;
320 sim_var_peq = 1/loops * sim_var_peq;
321 sim_var_k = 1/loops * sim_var_k;
322 sim_var_wep = 1/loops * sim_var_wep;
323 sim_var_bbm = 1/loops * sim_var_bbm;
324
325 print pi;
326 print rho;
327 print rho_WEP_V;
328 print rho_WEP_E;
```

Anhang

```
329 print rho_K_E;
330 print rho_PEQ_E;
331 print rho_ANO_E;
332 print rho_EQL_E;
333 print rho_PL_E;
334 print rho_ML_E;
335 print rho_MMB_E;
336 print rho_BBM_E;
337 print sim_var_pl;
338 print sim_var_eql;
339 print sim_var_mmb;
340 print sim_var_anova;
341 print sim_var_peq;
342 print sim_var_k;
343 print sim_var_wep;
344 print sim_var_bbm;
345
346   create y_i from y;                               /* Matrix in Datensatz schreiben
      */
347   append from y;
348   close y_i;
349   create rho_Matrix from rho_Matrix;                 /* Matrix in
      Datensatz schreiben */
350   append from rho_Matrix;
351   close rho_Matrix;
352 quit;
353
354
355 libname a 'R:\AD\Forschung\Programme\Rho_Schaetzer_betavert';
356
357 data a.y_i;
358 set y_i;
359 run;
360
361 data a.rho_Matrix;
362 set rho_Matrix;
363 run;
```

SAS Quellcode für Prozedur ML (Teil-Simulation Beta-Binomial-Modell)

```
1   /*ML*/
2
3   start = 1;
4   y_i = j(k,1,0);                                  /*Anzahl der Einsen in Klumpen i*/
5   do i = 1 to k;
6     hilf = 0;
7     do j = start to start + n_i[i] - 1;
8       hilf = hilf + klumpen[j,2];                 /*Zähle Einsen in Klumpen i*/
9     end;
10  y_i[i] = hilf;
```

Anhang

```
11      start = start + n_i[i];          /*Inkrement*/
12      end;
13
14      do j = 1 to nrow(v_rho);
15      l = 0;
16
17      do i=1 to k;
18      l_hilf = 0;
19      if y_i[i]=0 then                /*Fallunterscheidung da
20      sonst Div. durch Null*/
21      l_hilf = log(v_rho[j]*(1-pi) + (1-v_rho[j]) * (1-pi)**n_i[i]); /*
22      vgl. Formel in Zou*/
23      else if y_i[i]=n_i[i] then
24      l_hilf = log(v_rho[j] * pi + (1-v_rho[j]) * pi ** n_i[i]);
25      else l_hilf = log( comb(n_i[i],y_i[i]) * (1-v_rho[j]) * (pi**y_i[i])
26      * (1-pi)**(n_i[i]-y_i[i]) ); /*Vorsicht: bei v_rho=1 -> log(0)*/
27      /
28      l = l + l_hilf;
29      end;
30      v_ML[j] = l;
31      end;
32
33      /*Finde Rho*/
34      ML_max = max(v_ML);              /*finde max. Likelihoodbeitrag v. PL*/
35      ML = v_rho||v_ML;                /*Vektoren nebeneinanderschreiben*/
36      do i=1 to nrow(v_rho);          /*finde entspr. rho zum max.
37      Likelihoodbeitrag*/
38      if ML[i,2] = ML_max then rho_ML[M] = ML[i,1];
39      end;
```

SAS Quellcode für Prozedur EQL (Teil-Simulation Beta-Binomial-Modell)

```
1      /*EQL*/
2
3      start = 1;
4      y_i = j(k,1,0);                 /*Anzahl der Einsen in Klumpen i*/
5      do i = 1 to k;
6      hilf = 0;
7      do j = start to start + n_i[i] - 1;
8      hilf = hilf + klumpen[j,2];     /*Zähle Einsen in Klumpen i*/
9      end;
10     y_i[i] = hilf;
11     start = start + n_i[i];         /*Inkrement*/
12     end;
13
14     do j = 1 to nrow(v_rho);        /*Schleife für Likelihoodwerte zu
15     entspr. Rho*/
16     l = 0;
17
18     do i = 1 to k;
```

Anhang

```
18     l_hilf = 0;
19     if (y_i[i] = 0 then                                /*Fallunterscheidung
        da sonst Div. durch Null*/
20     D = 2 * ((n_i[i] - y_i[i]) * log10( (n_i[i] - y_i[i]) / (n_i[i] - n_
        i[i]*pi))); /*vgl. Formel in Zou*/
21     else if (y_i[i] = (n_i[i]) then
22     D = 2 * (y_i[i] * log10(y_i[i] / (n_i[i] * pi)));
23     else D = 2 * (y_i[i] * log10(y_i[i] / (n_i[i] * pi)) + (n_i[i] - y_i
        [i]) * log10( (n_i[i] - y_i[i]) / (n_i[i] - n_i[i]*pi)));
24     l_hilf = -0.5 * (log10(1 + (n_i[i] - 1) * v_rho[j]) + D / (1 + (n_i[
        i] - 1) * v_rho[j]));
25     l = l + l_hilf;
26     end;
27     *print l;
28     v_EQL[j] = l;
29     end;
30
31     /*Finde Rho*/
32     EQL_max = max(v_EQL);                                /*finde max. Likelihoodbeitrag v.
        EQL*/
33     EQL = v_rho||v_EQL;                                /*Vektoren nebeneinanderschreiben*
        /
34     do i=1 to nrow(v_rho);                                /*finde entspr. rho zum max.
        Likelihoodbeitrag*/
35         if eql[i,2] = eql_max then rho_EQL[M] = eql[i,1];
36     end;
37
38     sim_var_eql = sim_var_eql + (rho - rho_eql[m]) ** 2;    /*
        Fehlerquadrate*/
```

SAS Quellcode für Prozedur PL (Teil-Simulation Beta-Binomial-Modell)

```
1     /*PL*/
2
3     /*siehe Ridout*/
4     pi_sim_gesamt = j(N,1,1) * klumpen[,2] / N;
5
6     start = 1;
7     y_i = j(k,1,0);                                /*Anzahl der Einsen in Klumpen i*/
8     do i = 1 to k;
9         hilf = 0;
10        do j = start to start + n_i[i] - 1;
11            hilf = hilf + klumpen[j,2];                /*Zähle Einsen in Klumpen i*/
12        end;
13        y_i[i] = hilf;
14        start = start + n_i[i];                        /*Inkrement*/
15    end;
16
17    do j=1 to nrow(v_rho);                                /*Schleife für Likelihoodwerte zu
        entspr. Rho*/
```

Anhang

```
18      l = 0;                                /*Likelihood*/
19
20      do i = 1 to k;
21          l_hilf = 0;
22          l_hilf = -0.5 * (log10(1 + (n_i[i] - 1) * v_rho[j]) + ((y_i[i] - n_i
23              [i] * pi_sim_gesamt)**2) / (n_i[i] * pi_sim_gesamt * (1 - pi_sim
24              _gesamt) * (1 + (n_i[i] - 1) * v_rho[j])));
25          l = l + l_hilf;
26      end;
27
28      /*Finde Rho*/
29      PL_max = max(v_PL);                    /*finde max. Likelihoodbeitrag v. PL*/
30      PL = v_rho||v_PL;                      /*Vektoren nebeneinanderschreiben*/
31      do i=1 to nrow(v_rho);                 /*finde entspr. rho zum max.
32          Likelihoodbeitrag*/
33          if PL[i,2] = PL_max then rho_PL[M] = PL[i,1];
34      end;
35      sim_var_pl = sim_var_pl + (rho - rho_pl[M]) ** 2;    /*Fehlerquadrate*/
```

SAS Quellcode für Prozedur DBB (Teil-Simulation Beta-Binomial-Modell)

```
1      /*DBB*/
2
3      start = 1;
4      y_i = j(k,1,0);                        /*Anzahl der Einsen in Klumpen i*/
5      do i = 1 to k;
6          hilf = 0;
7          do j = start to start + n_i[i] - 1;
8              hilf = hilf + klumpen[j,2];    /*Zähle Einsen in Klumpen i*/
9          end;
10         y_i[i] = hilf;
11         start = start + n_i[i];           /*Inkrement*/
12     end;
13
14     pi_sim_gesamt = j(N,1,1) * klumpen[,2] / N;    /*Mittelwert der pi (
15         Klumpenstruktur nicht beachtet)*/
16
17     /*95% credible intervall*/
18     ogrenze = round(k * 0.975);
19     ugrenze = round(k * 0.025);
20
21     call sortndx(idpi, pi_i, 1);
22     pi_i_sort = pi_i[idpi, ];              /*pi_i_sort ist die sortierte
23         Variante von pi_i*/
24
25     aopt = pi_i_sort[ugrenze];
26     bopt = pi_i_sort[ogrenze];
```

Anhang

```
25 dmin = 100;
26
27 do alpha_hilf = 0.2 to 100 by 0.1;           /*siehe auch
      bayesianischesverteilungsmodell.R und Vortrag Dortmund*/
28   beta_hilf = alpha_hilf * (1-pi_sim_gesamt) / pi_sim_gesamt; /*hier das
      pi aus den simulierten Daten geschätzt*/
29   a = betainv(0.025,alpha_hilf,beta_hilf);
30   b = betainv(0.975,alpha_hilf,beta_hilf);
31   d = (a-aopt)**2 + (b-bopt)**2;
32
33   if (d<dmin) then do;
34     dmin = d;
35     alphamin = alpha_hilf;
36     betamin = beta_hilf;
37   end;
38 end;
39 rho_BBM[M] = 1 / (1+alphamin + betamin);
40
41 sim_var_bbm = sim_var_bbm + (rho - rho_bbm[m]) ** 2;      /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur WEP (Teil-Simulation Beta-Binomial-Modell)

```
1   /*WEP*/
2
3 /*Weighted Empirical Pairwise*/
4   F = 0;           /*Anzahl aller möglichen 2er Kombinationen
      (Paaren)*/
5   start = 1;      /*Startzeile*/
6   E = 0;         /*Erwartungswert E(X_ij,X_ij')*/
7   do i = 1 to k;
8     hilf = 0;    /*Anzahl der Einsen in Klumpen i*/
9     do j = start to start + n_i[i] - 1;
10      hilf = hilf + klumpen[j,2]; /*Zähle Einsen in Klumpen i*/
11    end;
12    if (hilf = 1) | (hilf = 0)
13      then hilf2 = 0;
14      else hilf2 = comb(hilf,2); /*Anzahl aller möglichen 2er
      Kombinationen (Erfolg) in Klumpen i*/
15    E = E + hilf2;
16    hilf3 = comb(n_i[i],2); /*Anzahl aller möglichen 2er
      Kombinationen (Paaren)*/
17    F = F + hilf3;
18    *rho_klumpen = ((hilf2/hilf3) - pi**2) / (pi*(1-pi));
19    *print rho_klumpen;
20    start = start + n_i[i]; /*Inkrement*/
21  end;
22  rho_WEP[M,1] = ((E/F) - pi**2) / (pi*(1-pi)); /*Schätzer WEP*/
23
```

Anhang

```
24  sim_var_wep = sim_var_wep + (rho - rho_wep[m]) ** 2;          /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur K (Teil-Simulation Beta-Binomial-Modell)

```
1  /*K*/
2
3  start = 1;
4  y_i = j(k,1,0);          /*Anzahl der Einsen in Klumpen i*/
5  do i = 1 to k;
6      hilf = 0;
7      do j = start to start + n_i[i] - 1;
8          hilf = hilf + klumpen[j,2];          /*Zähle Einsen in Klumpen i*/
9      end;
10     start = start + n_i[i];          /*Inkrement*/
11     y_i[i] = hilf;
12 end;
13 hilf2 = 0;
14 do i = 1 to k;
15     hilf2 = hilf2 + (y_i[i]**2 + (n_i[i] - y_i[i])**2 - n_i[i]) / (n_i[i]
16         * (n_i[i] - 1));
17 end;
18 rho_K[M,1] = (1/k * hilf2 - pi**2 - (1-pi)**2) / (2 * pi - 2*pi**2);
19
20 sim_var_k = sim_var_k + (rho - rho_k[m,1]) ** 2;          /*Fehlerquadrate*/
```

SAS Quellcode für Prozedur PEQ (Teil-Simulation Beta-Binomial-Modell)

```
1  /*PEQ*/
2
3  start = 1;
4  y_i = j(k,1,0);          /*Anzahl der Einsen in Klumpen i*/
5  do i = 1 to k;
6      hilf = 0;
7      do j = start to start + n_i[i] - 1;
8          hilf = hilf + klumpen[j,2];          /*Zähle Einsen in Klumpen i*/
9      end;
10     start = start + n_i[i];          /*Inkrement*/
11     y_i[i] = hilf;
12 end;
13 hilf2 = 0;
14 do i = 1 to k;
15     hilf2 = hilf2 + y_i[i] * (y_i[i] - 1);
16 end;
17 hilf3 = 0;
18 do i = 1 to k;
19     hilf3 = hilf3 + y_i[i] * (n_i[i] - 1);
20 end;
21 hilf4 = 0;
22 do i = 1 to k;
```

Anhang

```
23   hilf4 = hilf4 + n_i[i] * (n_i[i] - 1);
24   end;
25   hilf5 = hilf3 / hilf4;
26   rho_PEQ[M,1] = 1/(hilf5 * (1 - hilf5)) * (hilf2 / hilf4 - hilf5**2);
27
28   sim_var_peq = sim_var_peq + (rho - rho_peq[m,1]) ** 2;      /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur ANOVA (Teil-Simulation Beta-Binomial-Modell)

```
1   /*ANOVA*/
2
3   start = 1;
4   y_i = j(k,1,0);      /*Anzahl der Einsen in Klumpen i*/
5   do i = 1 to k;
6       hilf = 0;
7       do j = start to start + n_i[i] - 1;
8           hilf = hilf + klumpen[j,2];      /*Zähle Einsen in Klumpen i*/
9       end;
10      y_i[i] = hilf;
11      start = start + n_i[i];      /*Inkrement*/
12  end;
13  hilf = 0;
14  do i = 1 to k;
15      hilf = hilf + ((y_i[i]**2) / n_i[i]);
16  end;
17  sqa = (1 / (k-1)) * (hilf - (1/N * ((j(k,1,1) ` * y_i)**2)));
      /*s. Ridout*/
18  sqr = (1 / (N-k)) * ((j(k,1,1) ` * y_i) - hilf);
19  n_0 = (1 / (k-1)) * (N - 1/N * n_i ` * n_i) ;
20  rho_ANO[M] = (sqa - sqr) / (sqa + (n_0 -1) * sqr);
21
22  sim_var_anova = sim_var_anova + (rho - rho_ano[m]) ** 2;      /*
      Fehlerquadrate*/
```

SAS Quellcode für Prozedur MMB (Teil-Simulation Beta-Binomial-Modell)

```
1   /*MMB*/
2
3   pi_sim_gesamt = j(N,1,1) ` * klumpen[,2] / N;      /*Mittelwert der pi (
      Klumpenstruktur nicht beachtet)*/
4
5   start = 1;
6   y_i = j(k,1,0);      /*Anzahl der Einsen in Klumpen i*/
7   do i = 1 to k;
8       hilf = 0;
9       do j = start to start + n_i[i] - 1;
10          hilf = hilf + klumpen[j,2];      /*Zähle Einsen in Klumpen i*/
11      end;
12      y_i[i] = hilf;
```

Anhang

```
13     start = start + n_i[i];           /*Inkrement*/
14     end;
15
16     do j=1 to nrow(v_rho);           /*Schleife für Likelihoodwerte zu
17         entspr. Rho*/
18         l = 0;                       /*"Likelihood"*/
19
20         do i = 1 to k;               /*Schleife für Summenbildung*/
21             l_hilf = 0;
22             l_hilf = ((y_i[i] - n_i[i] * pi_sim_gesamt)**2) / (n_i[i] * pi_sim_
23                 gesamt * (1 - pi_sim_gesamt) * (1 + (n_i[i] + 1) * v_rho[j]));
24             l = l + l_hilf;
25         end;
26         v_MMB[j] = l;
27     end;
28
29     /*Finde Rho*/
30     **rho_MMB[M] = 0;
31     MMB_diff = (v_MMB - k);
32     MMB_min = min(abs(MMB_diff));
33     *print MMB_min; *print MMB_diff;
34     MMB = v_rho||MMB_diff;           /*Vektoren
35     nebeneinanderschreiben*/
36
37     do i = 1 to nrow(v_rho);
38         if MMB[i,2] = MMB_min then rho_MMB[M] = MMB[i,1];
39     end;
40
41     sim_var_mmb = sim_var_mmb + (rho - rho_mmb[m]) ** 2;           /*
42     Fehlerquadrate*/
```

A.3.2 Konfidenzintervall für die Intra-Klumpen-Korrelation

SAS Quellcode für Kapitel 5.5 (am Beispiel INSOL)

```
1     /*Konfidenzintervall für rho in einem Beta-Bin. Modell (vgl. Lui 1996)
2     */
3
4
5     libname a 'R:\AD\Forschung\Daten';
6
7
8     data n;                           /*Initialisierung n-Vektor = Anzahl Individuen
9         in Cluster i=1..k */
10    input n_i;
11    cards;
12    7915
13    12363
```

Anhang

```
13 5625
14 2899
15 966
16 11050
17 1353
18 5484
19 5736
20 ;
21 run;
22
23 data y;                                /*Initialisierung n-Vektor = Anzahl Individuen
      in Cluster i=1..k */
24 input y_i;
25 cards;
26 180
27 121
28 75
29 49
30 8
31 107
32 1
33 11
34 48
35 ;
36 run;
37
38
39 Proc IML;
40
41 use n;                                  /*n-Vektor übernehmen in IML*/
42 show datasets;
43 show contents;
44 read all into n_i;
45
46 use y;                                  /*y-Vektor übernehmen in IML*/
47 show datasets;
48 show contents;
49 read all into y_i;
50
51 /*****
      *****/
52 alpha = 4;                              /*nach beta binomial model schätzung (siehe
      *insol.r)*/
53 beta = 365;
54 /*****
      *****/
55
56 k = nrow(n_i);                          /*k = Anzahl Cluster*/
57 N = j(k,1,1) `*n_i;
58 pi_exakt = alpha / (alpha + beta);      /*determiniertes Pi*/
```

Anhang

```
59 T_exakt = alpha + beta;
60 Y = y_i`*j(k,1,1);          /*Summe aller Erfolge*/
61 pi = Y/N;
62
63 print pi pi_exakt;
64
65 E_Y_i_bed = j(k,1,1);      /*Initialisierung Vektor */
66 E_Y_i2_bed = j(k,1,1);
67 E_Y_i3_bed = j(k,1,1);
68 E_Y_i4_bed = j(k,1,1);
69 var_Y_i_bed = j(k,1,1);
70 var_Y_i2_bed = j(k,1,1);
71 cov_YY2_bed = j(k,1,1);
72 var_B_i_bed = j(k,1,1);
73 var_W_i_bed = j(k,1,1);
74 cov_BW_i_bed = j(k,1,1);
75
76 /*****
77      *****/
78 /*Sum of Squares*/
79 BMS = 0;                    /*Between bei unbalancierten Cluster*/
80 do i = 1 to k;
81   BMS = BMS + n_i[i] * (Y_i[i]/n_i[i] - Y/N)**2 / (k-1);
82 end;
83 WMS = 0;                    /*Within bei unbalancierten Cluster*/
84 do i = 1 to k;
85   WMS = WMS + (Y_i[i] * (1-Y_i[i]/n_i[i])**2 + (n_i[i]-Y_i[i]) * (Y_i[i]/n_i[i])**2) / (N-k);
86 end;
87
88 /*ein möglicher Punktschätzer für rho (ANOVA)*/
89 n_0 = (N - 1/N * (n_i` * n_i)) / (k-1);
90 T_anova = K * WMS / (BMS - WMS);          /* S 424 Mitte*/
91 rho_anova = (BMS - WMS)/(BMS + (n_0 - 1)*WMS);
92
93 /*bedingte Erwartungswerte + höhere Momente*/
94 T = 369; /*T_anova;          /*Wahl des Punktschätzers bzw.
95   geschätzer Wert alpha 4+ beta 365*/
96 E_Y_i_bed = pi * n_i;
97 print rho_anova t_anova t;
98
99 do i = 1 to k;
100  E_Y_i2_bed[i] = (n_i[i] * pi + n_i[i] * pi) * ((n_i[i] - 1) * (pi*T+1)/(T+1)) + n_i[i] * pi;
101
102 do i = 1 to k;
103  E_Y_i3_bed[i] = (n_i[i] * pi + n_i[i] * pi) * ((n_i[i] - 1) * (pi*T+1)/(T+1)) * ((n_i[i] - 2) * (pi*T+2)/(T+2)) + 3 * E_Y_i2_bed[i] ;
```

Anhang

```
104 end;
105
106 do i = 1 to k;
107   E_Y_i4_bed[i] = (n_i[i] * pi + n_i[i] * pi) * ((n_i[i] - 1) * (pi*T+1)/(
      T+1)) * ((n_i[i] - 2) * (pi*T+2)/(T+2)) * ((n_i[i] - 3) * (pi*T+3)/(
      T+3))
108 + 6 * (n_i[i] * pi + n_i[i] * pi) * ((n_i[i] - 1) * (pi*T+1)/(T+1)) * ((n_
      i[i] - 2) * (pi*T+2)/(T+2))
109 + 7 * E_Y_i2_bed[i] ;
110 end;
111
112 var_Y_i_bed = E_Y_i2_bed - E_Y_i_bed # E_Y_i_bed;      /*S. 422 Lui 1996*/
113 var_Y_i2_bed = E_Y_i4_bed - E_Y_i2_bed # E_Y_i2_bed;
114 cov_YY2_bed = E_Y_i3_bed - E_Y_i_bed # E_Y_i2_bed;
115
116 do i = 1 to k;
117   var_B_i_bed[i] = (var_Y_i2_bed[i] - 4*E_Y_i_bed[i]*cov_YY2_bed[i] + 4*(E
      _Y_i_bed[i]**2)*var_Y_i_bed[i]) / n_i[i]**2; /*S. 423 Lui */
118 end;
119
120 do i = 1 to k;
121   var_W_i_bed[i] = 1/(N/k-1)**2 * (var_Y_i_bed[i] - 2*cov_YY2_bed[i]/n_i[i
      ] + var_Y_i2_bed[i]/(n_i[i]**2) );      /*hier mit \mu_c := n_i_quer*/
122 end;
123
124 do i = 1 to k;
125   cov_BW_i_bed[i] = 1/((N/k-1)*n_i[i]) * (cov_YY2_bed[i] - 2*E_Y_i_bed[i]
      * var_Y_i_bed[i] - var_Y_i2_bed[i]/n_i[i] + 2*E_Y_i_bed[i]*cov_YY2_
      bed[i]/n_i[i] );      /*hier mit \mu_c := n_i_quer*/
126 end;
127
128 /*Konfidenzintervall*/
129 BMS_stern = (k-1) / k * BMS;
130
131 A = WMS**2 - probit(.975)**2 * var_W_i_bed`*j(k,1,1) / k**2 ;      /*S.
      424 für unbalancierte Cluster, */
132 B = BMS_stern * WMS - probit(.975)**2 * cov_BW_i_bed`*j(k,1,1) / k**2 ;
133
134 theta_l = (B - sqrt(B**2 - A*B)) / A ;      /*S 423 unten*/
135 theta_u = (B + sqrt(B**2 - A*B)) / A ;
136
137 rho_l = (theta_l - 1) / (theta_l + N/k - 1);      /*hier mit \mu_c := n_i_
      quer*/ /*S. 424 oben*/
138 rho_r = (theta_u - 1) / (theta_u + N/k - 1);
139
140 print rho_l rho_r rho_anova;
141
142
143 quit;
```

Literaturverzeichnis

- Aalen, O. O. und Frigessi, A. (2007). What can Statistics Contribute to a Causal Understanding? *Scandinavian Journal of Statistics*, 34(1):155–168.
- Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., und Campbell, M. J. (2004). Patterns of Intra-Cluster Correlation from Primary Care Research to Inform Study Design and Analysis. *Journal of Clinical Epidemiology*, 57(8):785–794.
- Bleymüller, J. und Weißbach, R. (2015). *Statistik für Wirtschaftswissenschaftler*. Vahlen, München, 17. Auflage.
- Blokland, G. A. M., McMahon, K. L., Thompson, P. M., Hickie, I. B., Martin, N. G., Zubicaray, G. I. d., und Wright, M. J. (2014). Genetic Effects on The Cerebellar Role in Working Memory: Same Brain, Different Genes? *NeuroImage*, 86:392–403.
- Boles, R. E., Scharf, C., Filigno, S. S., Saelens, B. E., und Stark, L. J. (2013). Differences in Home Food and Activity Environments Between Obese and Healthy Weight Families of Preschool Children. *Journal of Nutrition, Education and Behavior*, 45(3):222–231.
- Bortz, J. und Döring, N. (2002). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. Springer-Lehrbuch. Springer, Berlin, 3. Auflage.
- Bundesministerium der Justiz und für Verbraucherschutz (2015). Gesetz über das Kreditwesen (Kreditwesengesetz): KWG.
- Carroll, R. J. und Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.
- Chmura Kraemer, H., Periyakoil, V. S., und Noda, A. (2002). Kappa Coefficients in Medical Research. *Statistics in Medicine*, 21(14):2109–2129.
- Commerzbank (2010). Geschäftsbericht 2010: Herausforderungen annehmen, Ziele erreichen.
- Cox, D. R. und Snell, E. J. (1989). *Analysis of Binary Data*. Chapman and Hall, London, New York, 2. Auflage.

Literaturverzeichnis

- Credit Suisse First Boston (1997). Credit Risk+: A Credit Risk Management Framework: Technical Dokument.
- Crouhy, M., Galai, D., und Mark, R. (2009). *Risk Management*. McGraw-Hill, New York.
- DAJ-Mitgliederversammlung (1998). Deutsche Arbeitsgemeinschaft für Jugendzahn-pflege: Dokumentation der Maßnahmen der Gruppenprophylaxe in Kindergärten und Schulen - Jahresauswertung 1996/97.
- Davidian, M. und Carroll, R. J. (1988). A Note on Extended Quasi-Likelihood. *Journal of the Royal Statistical Society*, (50):74–82.
- Dixon, V., Read, M. J. F., O'Brien, K. D., Worthington, H. V., und Mandall, N. A. (2002). A Randomized Clinical Trial to Compare Three Methods of Orthodontic Space Closure. *Journal of Orthodontics*, 29(1):31–36.
- Donner, A. und Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.
- Duffie, D. und Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement and Manage-ment*. Princeton Series in Finance. Princeton Univ. Press, Princeton, NJ.
- Dupuis, J. A. (1995). Bayesian Estimation of Movement and Survival Probabilities from Capture-Recapture Data. *Biometrika*, (82):761–772.
- Eldridge, S. M., Ukoumunne, O. C., und Carlin, J. B. (2009). The Intra-Cluster Correla-tion Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review*, 77(3):378–394.
- Fieller, E. C. (1954). Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.
- Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edin-burgh, 14. Auflage.
- Fleiss, J. L. und Cuzick, J. (1979). The Reliability of Dichotomous Judgments: Unequal Numbers of Judges per Subject. *Applied Psychological Measurement*, 3(4):537–542.
- Fleiss, J. L., Levin, B., und Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Statistics. J. Wiley, Hoboken, N.J., 3. Auflage.
- Ganninger, M., Häder, S., und Gabler, S. (2007). Design Effects and Interviewer Effects in the European Social Survey: Where Are We Now and Where Do We Want to Go Tomorrow? *Centre for Survey Research and Methodology*.

Literaturverzeichnis

- Gouriéroux, C. und Monfort, A. (1995). *Statistics and Econometric Models*. Cambridge Univ. Press, Cambridge, 1. Auflage.
- Gulliford, M., Adams, G., Ukoumunne, O., Latinovic, R., Chinn, S., und Campbell, M. (2005). Intraclass Correlation Coefficient and Outcome Prevalence are Associated in Clustered Binary Data. *Journal of Clinical Epidemiology*, 58(3):246–251.
- Hartung, J., Elpelt, B., und Klösener, K.-H. (2009). *Statistik: Lehr- und Handbuch der angewandten Statistik*. Oldenbourg, München, 15. Auflage.
- Hesse, C. (2003). *Angewandte Wahrscheinlichkeitstheorie: Eine fundierte Einführung mit über 500 realitätsnahen Beispielen und Aufgaben*. Vieweg, Braunschweig and Wiesbaden, 1. Auflage.
- Jablonski-Momeni, A., Liebegall, F., Stoll, R., Heinzl-Gutenbrunner, M., und Pieper, K. (2013). Performance of a New Fluorescence Camera for Detection of Occlusal Caries in Vitro. *Lasers in Medical Science*, 28(1):101–109.
- Jin, Y., Shi, Y., Zhan, L., Gutman, B. A., Zubicaray, G. I. d., McMahon, K. L., Wright, M. J., Toga, A. W., und Thompson, P. M. (2014). Automatic Clustering of White Matter Fibers in Brain Diffusion MRI with an Application to Genetics. *NeuroImage*, 100:75–90.
- Kish, L. (1965). *Survey Sampling*. Wiley Classics Library. J. Wiley, New York.
- Krämer, W., Schoffer, O., und Tschiersch, L. (2008). *Datenanalyse mit SAS®: Statistische Verfahren und ihre grafischen Aspekte*. Springer, Berlin, 2. Auflage.
- Kremer, A., Weißbach, R., und Liese, F. (2014). Maximum Likelihood Estimation for Left-Censored Survival Times in an Additive Hazard Model. *Journal of Statistical Planning and Inference*, 149:33–45.
- Lancaster, T. (2000). The Incidental Parameter Problem Since 1948. *Journal of Econometrics*, 95(2):391–413.
- Lehmann, E. L. und Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, 2. Auflage.
- Lester, P., Stein, J. A., Saltzman, W., Woodward, K., MacDermid, S. W., Milburn, N., Mogil, C., und Beardslee, W. (2013). Psychological Health of Military Children: Longitudinal Evaluation of a Family-Centered Prevention Program to Enhance Family Resilience. *Military Medicine*, 178(8):838–845.
- Lui, K.-J., Cumberland, W. G., und Kuo, L. (1996). An Interval Estimate for the Intraclass Correlation in Beta-Binomial Sampling. *Biometrics*, 52(2):412.

Literaturverzeichnis

- Madsen, R. W. (1993). Generalized Binomial Distributions. *Communications in Statistics - Theory and Methods*, 22(11):3065–3086.
- McCullagh, P. und Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2. Auflage.
- McNeil, A. J., Frey, R., und Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton Univ. Press, Princeton, NJ.
- Merlo, J., Ohlsson, H., Chaix, B., Lichtenstein, P., Kawachi, I., und Subramanian, S. V. (2013). Revisiting Causal Neighborhood Effects on Individual Ischemic Heart Disease Risk: a Quasi-Experimental Multilevel Analysis among Swedish Siblings. *Social Science and Medicine*, 76(1):39–46.
- Pagel, C., Prost, A., Lewycka, S., Das, S., Colbourn, T., Mahapatra, R., Azad, K., Costello, A., und Osrin, D. (2011). Intracluster Correlation Coefficients and Coefficients of Variation for Perinatal Outcomes from five Cluster-Randomised Controlled Trials in Low and Middle-Income Countries: Results and Methodological Implications. *Trials*, 12:151.
- Ridout, M. S., Demetrio, C. G. B., und Firth, D. (1999). Estimating Intraclass Correlation for Binary Data. *Biometrics*, 55(1):137–148.
- Rinne, H. (2008). *Taschenbuch der Statistik*. Harri Deutsch, Frankfurt a. M., 4. Auflage.
- Rosenow, B. und Weißbach, R. (2009). Modelling Correlations in Credit Portfolio Risk. *Journal of Risk Management in Financial Institutions*, 3(1):16–30.
- Snedecor, G. W. und Cochran, W. G. (1989). *Statistical Methods*. Iowa State University Press, Ames, 8. Auflage.
- Statistik-Hessen (2013). <http://www.statistik-hessen.de/infomaterial/index.html>; zuletzt geprüft am 13.08.2015.
- Statistisches Amt MV (2015a). http://www.statistik-mv.de/cms2/STAM_prod/STAM/de/gui/Veroeffentlichungen/index.jsp?para=e-BiboInterTh08&linkid=080202&head=0802; zuletzt geprüft am 31.08.2015.
- Statistisches Amt MV (2015b). http://www.statistik-mv.de/cms2/STAM_prod/STAM/de/gui/Veroeffentlichungen/index.jsp?para=e-BiboInterTh08&linkid=080302&head=0803; zuletzt geprüft am 31.08.2015.
- Stefanescu, C. und Turnbull, B. W. (2003). Likelihood Inference for Exchangeable Binary Data with Varying Cluster Sizes. *Biometrics*, 59(1):18–24.

Literaturverzeichnis

- Tasche, D. (2013). Bayesian Estimation of Probabilities of Default for Low Default Portfolios. *Journal of Risk Management in Financial Institutions*, 6(3):302–326.
- Thompson, D. M., Fernald, D. H., und Mold, J. W. (2012). Intraclass Correlation Coefficients Typical of Cluster-Randomized Studies: Estimates From the Robert Wood Johnson Prescription for Health Projects. *Annals of Family Medicine*, 10(3):235–240.
- Ullah, A. und Giles, David E. A. (2011). *Handbook of Empirical Economics and Finance. Statistics: Textbooks and Monographs*. Chapman and Hall, Boca Raton.
- Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61(3):439–447.
- Weißbach, R. und Herzog, M. (2009). Schätzung des Kariesbefalls 3–5 jähriger Kinder aus einstufigen Clusterstichproben. *Das Gesundheitswesen*, 71(03):121–126.
- Weißbach, R., Herzog, M., und Menzel, G. (2015). Regionaler Anteil kariesfreier Vorschulkinder - eine cluster-randomisierte Studie in Südhessen -. *AStA Wirtschafts- und Sozialstatistisches Archiv*, (9):27–39.
- Wooldridge, J. M. (2003). Cluster-Sample Methods in Applied Econometrics. *American Economic Review*, 93(2):133–138.
- Yamamoto, E. und Yanagimoto, T. (1992). Moment Estimators for the Beta-Binomial Distribution. *Journal of Applied Statistics*, 19(2):273–283.

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehörde zur Erlangung eines akademischen Grades vorgelegt.

Ort, Datum

Unterschrift