Flexible Mischungsmodelle zur Modellierung von Unsicherheit in ordinalen Regressionsmodellen

Masterarbeit

Cynthia Huber

Betreuung:
Prof. Dr. Gerhard Tutz
Micha Schneider

3. März 2016

Ludwig- Maximilians- Universität München
Institut für Statistik

Abstract

Umfragen bestehen häufig aus Fragen bei welchen die Befragten dazu aufgefordert werden ein Produkt, eine Dienstleistung oder ihre Zufriedenheit auf einer ordinalen Skala zu beurteilen. Einer solchen Entscheidung geht ein psychologischer menschlicher Entscheidungsprozess voraus. Dieser Entscheidungsprozess wird in den CUB- und CUP-Modellen, welche die ordinale Beurteilung als Responsevariable verwenden, als eine Mischung einer Präferenz und Unsicherheit angenommen. Die Präferenz bezieht sich auf eine wohlüberlegte Wahl einer der Kategorien der ordinalen Skala. Die Unentschlossenheit der Befragten wird dabei mit Unsicherheit bezeichnet. Die erste Komponente , die Präferenz, wird im CUB-Modell mit einer Binomialverteilung modelliert. Im CUP-Modell hingegen ist jedes mögliche ordinale Modell für die Modellierung erlaubt. Die zweite Komponente entspricht bei beiden Modellen der diskreten Gleichverteilung über die Ausprägungen der ordinalen Responsevariablen. Der Grund für die Wahl der diskreten Gleichverteilung ist die Einfachheit, die damit einhergeht.
Die vorliegende Arbeit befasst sich mit der Verwendung einer flexibleren diskreten Verteilung zur Modellierung der zweiten Komponente: eine spezielle Form der Betabinomialverteilung. Das so entstehende Modell wird mit BETAMIX-Modell bezeichnet. Mit der Betabinomialverteilung ist es möglich eine Tendenz zu den mittleren oder den extremen Responsekategorien zu modellieren.
Diese Tendenz kann entweder für alle Individuen als gleich oder durch das Einbinden von Kovariablen Individuen spezifisch betrachtet werden. Werden keine Variablen mit dem Parameter der für das BETAMIX-Modell verwendeten Betabinomialverteilung verknüpft, so kann dieser Parameter über das Abtasten eines Gitters mit festen Parameterwerten gewählt werden. Statt des Abtastens ist jedoch auch die Möglichkeit der Schätzung des Parameters gegeben. Die Simulationsstudie zeigt, dass sich geschätzte und gewählte Werte für den Parameter der Betabinomialverteilung nicht stark unterscheiden. Außerdem zeigen die Simulationen, dass die Schätzungen aller Parameter des BETAMIX-Modells gut sind. Da für die Wahl des Parameters eine Vielzahl von Modellen angepasst werden muss um anschließend das Beste darunter zu wählen, kann dies sehr zeitaufwendig werden. Aus diesen Grund ist die Schätzung des Parameters der Betabinomialverteilung im BETAMIX-Modell zu bevorzugen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

In vielen Umfragen werden Befragte aufgefordert ein Produkt oder eine Dienstleistung auf einer ordinalen Skala zu bewerten. Auch die Beurteilungen der eigenen Zufriedenheit oder der Stärke des empfundenen Schmerzes erfolgen meist auf einer solchen Skala. Eine solche Beurteilung muss nicht immer wohldurchdacht sein. Umstände wie Müdigkeit, die Zeit für die Umfrage oder auch ein Mangel an Informationen zu dem befragten Thema können zu einer Unsicherheit bezüglich der finalen Beurteilung führen (vgl. Iannario und Piccolo, 2012).
Das von Iannario und Piccolo (2012) vorgestellte CUB-Modell und das flexiblere CUP-Modell, welches von Tutz et al. (2014) beschrieben wird, berücksichtigen die Mischung aus bedachter Antwort (Präferenz) und Unsicherheit im Entscheidungsprozess.
CUB- und CUP-Modelle unterscheiden sich in der Modellierung der ersten Mischkomponente, der Komponente für die Präferenz. Das CUB-Modell nutzt für die erste Komponente eine verschobene Binomialverteilung. Das CUP-Modell hingegen modelliert die Präferenz mit einem beliebigen ordinalen Regressionsmodell. Die zweite Komponente, die Unsicherheitskomponente, wird sowohl im CUB- als auch im CUP-Modell mit einer diskreten Gleichverteilung modelliert. Für mehr Flexibilität wird diese Verteilungsannahme der Unsicherheitskomponente in den hier betrachteten BETAMIX-Modellen durch eine beschränkte Betabinomialverteilung ersetzt. Dies ermöglicht, dass die Unsicherheit auch eine Tendenz zu den mittleren oder extremen Kategorien aufweisen kann.
In mehreren Studien wurde bereits gezeigt, dass eine solche Tendenz, ein Responsestyle, von Kovariablen abhängig sein kann. Unterschiede im Responsestyle können beispielsweise zwischen unterschiedlichen Nationen (Clarke (2000), Van Herk et al. (2004)) oder Bildungsniveaus (Meisenberg und Williams (2008)) auftreten. Die Möglichkeit einen Responsestyle in den BETAMIX-Modellen zu modellieren ist ebenfalls gegeben. Dafür kann der Parameter der Betabinomialverteilung, welcher die Form der Verteilung bestimmt, mit Kovariablen verknüpft werden.
In Kapitel 2 werden Modelle für die Regression mit einer ordinalen, abhängigen Variable vorgestellt. Dabei werden zunächst Modelle, wie beispielsweise das kumulative Logit-Modell, betrachtet, die keine Unsicherheit in den Beurteilungen berücksichtigen. Im Anschluss werden die von Iannario und Piccolo (2012) und Tutz et al. (2014) vorgestellten CUB- und CUP-Modelle definiert. Das flexiblere BETAMIX-Modell wird in Unterabschnitt 2.2.3 vorgestellt. Weiterhin beinhaltet Kapitel 2 auch einen Abschnitt zu

Kriterien, die dem Modellvergleich dienen.
Der EM-Algorithmus, welcher der Parameterschätzung der Mischmodelle CUB, CUP und BETAMIX dient, wird in Kapitel 3 beschrieben.
Für die Beurteilung der mit dem BETAMIX-Modell geschätzten Parameter wird eine Simulationsstudie durchgeführt. Die Beschreibung der unterschiedlichen Simulationsszenarien und der Ergebnisse findet sich in Kapitel 4 wieder. Das letzte Kapitel, Kapitel 5, beinhaltet die Vorstellung einiger Anwendungen des BETAMIX-Modells.

2 Modelle für eine ordinale abhängige Variable

Ordinale Variablen können laut Anderson (1984) in zwei Gruppen aufgeteilt werden: in gruppiert-stetige Variablen („grouped continuous“) und in ordinale Beurteilungen („assessed ordered“). Kategorisierte, prinzipiell stetig messbare Größen werden gruppiert stetig genannt. Ein Beispiel hierfür ist die Dauer der Arbeitslosigkeit, die in Tagen und somit stetig erfasst werden kann. Häufig findet jedoch die kategorisierte Form dieser Variablen mit den Kategorien „Kurzzeitarbeitslosigkeit“, „Mittelfristigearbeitslosigkeit“ und „Langzeitarbeitslosigkeit“ Anwendung. Die zweite Gruppe der ordinalen Variablen sind ordinale Beurteilungen. Diese Variablen versuchen den Zustimmungsgrad zu unterschiedlichen Sachverhalten auf einer gegebenen Skala zu messen. Mit vier Kategorien lässt sich beispielsweise der Grad der eigenen Befindlichkeit erfassen. Eine Möglichkeit der Bezeichnung dieser vier Kategorien ist „ausgezeichnet“,„gut“, „mittelmäßig“ und „schlecht“. (vgl. Tutz, 2012, S.241)

Für die Regression mit einer ordinal skalierten Zielgröße sind spezielle Modelle nötig. Im Folgenden wird die Realisation des i-ten Individuums der ordinal skalierten Variable Y mit y_i bezeichnet . Die geordneten Kategorien der Variable Y seien $1, \ldots, k$. In Regressionsmodellen mit einer metrischen, abhängigen Variable werden im Gegensatz zu ordinalen Regressionsmodellen stärkere Annahmen getroffen. Neben der stärkeren Annahme des Skalenniveaus in Modellen mit metrischem Response, wird in solchen Modellen auch von einer symmetrischen Verteilung der Variable Y_i gegeben $\mathbf{x}_i$ ausgegangen. Dies ist in ordinalen Modellen im Allgemeinen nicht der Fall. Dabei beschreibt $\mathbf{x}_i$ einen Kovariablenvektor des Individuums i. (vlg. Tutz, 2000, S.208)

Kategorial-nominale Modelle, welche die ordinale Struktur der Zielgröße ignorieren, können angewandt werden, gehen jedoch mit einem Informationsverlust einher. Außerdem können die Parameter unter Umständen in einem kategorial-nominalen Regressionsmodell bei einer hohen Anzahl von Kategorien nicht geschätzt werden. Grund hierfür kann eine hohe Anzahl an zu schätzenden Parametern sein.

2.1 Modelle ohne Mischverteilungsansatz

Durch einen Split der ordinalen Variable Y bei r und $r+1$ entstehen die gruppierten Kategorien $\{1,\dots,r\}$ und $\{r+1,\dots,k\}$, aus welchen eine binäre Variable y_r definiert werden kann:

$$y_r = \begin{cases} 1 & \text{für } Y \leq r \\ 0 & \text{für } Y > r \end{cases}$$

Wird y_r als abhängige Variable betrachtet, so kann für diese ein binäres Modell $P(y_r = 1|\mathbf{x}) = F(\mathbf{x}^T\beta_r)$ verwendet werden. Durch den genutzten Split ergibt sich $P(y_r = 1|\mathbf{x}) = P(Y \leq r|\mathbf{x})$. Dies entspricht dem kumulativen Ansatz. In Unterabschnitt 2.1.1 wird das von McCullagh (1980) vorgestellte kumulative Modell motiviert und definiert.
Ein weiterer Ansatz zur Modellierung eines ordinalen Responses stellt das sequentielle Modell dar. Dabei folgt der Übergang der Kategorie r zu $r+1$ gegeben der Kategorie r oder höher einem binären Modell. Das binäre Modell unterscheidet zwischen $Y = r$ und $Y > r$ gegeben $Y \geq r$. Der daraus folgende sequentielle Ansatz lautet

$$P(Y = r|Y \geq r, \mathbf{x}) = F(\mathbf{x}^T\boldsymbol{\beta}_r) \quad r = 1,\dots,k-1.$$

Das Adjacent Categories Modell bietet eine weitere Möglichkeit zur Modellierung einer ordinalen Zielgröße. Gegeben zweier benachbarten Kategorien unterscheidet das binäre Modell zwischen diesen beiden Kategorien:

$$P(Y = r|Y \in \{r, r+1\}) = F(\mathbf{x}^T\boldsymbol{\beta}_r) \quad r = 1,\dots,k-1.$$

In Unterabschnitt 2.1.2 wird das Adjacent Categories Modell genauer erläutert. (vgl. Tutz, 2012, S.242f)

2.1.1 Kumulatives Modell

Das kumulative Modell wird durch eine metrische, latente Variable $\tilde{Y}_i$ motiviert. Die ordinale Variable Y_i sei die kategorisierte Version dieser latenten Variable $\tilde{Y}_i$. Es wird

angenommen, dass $\tilde{Y}_i$ aus dem Regressionsmodell

$$\tilde{Y}_i = -\mathbf{x}_i^T \boldsymbol{\gamma} + \epsilon_i$$

resultiert. Dabei sei ϵ_i eine Störgröße mit metrischer Verteilungsfunktion F und $E(\epsilon_i) = 0$. Die Stärke des Zusammenhangs der Kovariablen $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ und $\tilde{Y}_i$ wird durch die Parameter $\boldsymbol{\gamma}^T = (\gamma_1, \ldots, \gamma_p)$ bestimmt. Dabei enthält der Variablenvektor $\mathbf{x}_i$ keine Konstante (vgl. Tutz, 2000, S.209f).

Die latente Variable $\tilde{Y}_i$ und die beobachtete ordinale Variable Y_i sind wie folgt verknüpft:

$$Y_i = r \Longleftrightarrow \gamma_{0,r-1} < \tilde{Y}_i \leq \gamma_{0r} \quad .$$

Dabei seien $-\infty = \gamma_{00} < \gamma_{01} < \ldots < \gamma_{0k} = \infty$ Schwellenwerte auf der latenten Skala.
Aus diesen Annahmen folgt das kumulative Modell:

$$P(Y_i \leq r|\mathbf{x}_i) = P(-\mathbf{x}_i^T \boldsymbol{\gamma} + \epsilon_i \leq \gamma_{0r}) = P(\epsilon_i \leq \gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma}) = F(\gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma}) \tag{2.1}$$

Interpretationen der latenten Variablen $\tilde{Y}_i$, welche dem beobachteten Prozess zugrunde liegt, können von Interesse sein. Die latente Variable, die der Klassifizierung der Dauer von Arbeitslosigkeit zugrunde liegt, kann beispielsweise als individuelle Attraktivität auf dem Arbeitsmarkt interpretiert werden. Bei Ratings kann die latente Variable $\tilde{Y}_i$ als nicht direkt beobachtbare Einstellung zu der gefragten Aussage betrachtet werden (vgl. Tutz, 2000, S.210). Auch wenn oftmals eine Interpretation der latenten Variable vorhanden ist, kann das kumulative Modell ohne Bezug auf diese Variable genutzt werden. $\tilde{Y}_i$ dient lediglich zur Motivation des Modells (vgl. Tutz, 2012, S.243).
Das meist genutzte Modell aus der Klasse der kumulativen Modelle ist das kumulative Logit-Modell, auch „Proportional Odds Model“ genannt. Dieses Modell verwendet für die Verteilungsfunktion $F(\cdot)$ die logistische Funktion $F(u) = \exp(u)/(1 + \exp(u))$. Daraus ergibt sich das kumulative Logit-Modell:

$$\log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}\right) = \gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma} \quad , r = 1, \ldots, k-1 \tag{2.2}$$

2.1.2 Adjacent Categories Modell

Ein weiteres Modell für eine ordinale Responsevariable ist das Adjacent Categories Modell, welches benachbarte Kategorien $\{r, r+1\}$ betrachtet. Das Adjacent Categories Modell hat die Form

$$P(Y_i = r | Y_i \in \{r, r+1\}, \mathbf{x}_i) = F(\gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma}) \quad , r = 1, \ldots, k-1$$

Bei Verwendung der logistischen Verteilung ergibt sich das Adjacent Categories Logit Modell

$$\log\left(\frac{P(Y_i = r+1|\mathbf{x}_i)}{P(Y_i = r|\mathbf{x}_i)}\right) = \gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma} \quad , r = 1, \ldots, k-1$$

Das Modell nimmt an, dass die Logits $\log\left(\frac{P(Y_i=r+1|\mathbf{x}_i)}{P(Y_i=r|\mathbf{x}_i)}\right)$ durch einen kategorienspezifischen Intercept γ_{0r} und einen linearen Effekt der erklärenden Variablen $\mathbf{x}_i^T \boldsymbol{\gamma}$ definiert sind. Im Gegensatz zum Multinomialen Logit Modell ist der Koeffizient des Adjacent Categories Logit Modells $\boldsymbol{\gamma}$ nicht von der Kategorie r abhängig und modelliert somit indirekt die ordinale Struktur der Kategorien. Bei einem Zuwachs der Variablen x_j um eine Einheit und bei Konstanthaltung der anderen Variablen, kann $\exp(\gamma_j)$ als Odds Ratio, welcher die Chance der Kategorie $r+1$ und r vergleicht, interpretiert werden. (vgl. Tutz und Berger, 2015)

$$\exp(\gamma_j) = \frac{P(Y_i = r+1|x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip}) / P(Y_i = r|x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip})}{P(Y_i = r+1|x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}) / P(Y_i = r|x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})}$$

2.2 Modelle mit Mischverteilungsansatz

2.2.1 Das CUB- Modell

Für die Erklärung ordinal skalierter Umfrageantworten wurde das von Iannario und Piccolo (2012) beschriebene CUB-Modell entwickelt. Die Wahl einer Kategorie aus einer geordneten Liste mit k Kategorien sei laut Iannario und Piccolo (2012) eine komplexe Entscheidung, welche sich aus der Meinung des Subjekts zum Objekt und einer Unschärfe

bezüglich der endgültigen Antwort zusammensetzt. Im Folgenden werden diese unbeobachtbaren Komponenten als Präferenz bzw. Empfindung („feeling“) und Unsicherheit („uncertainty“) bezeichnet. In konkreten Anwendungen kann es sein, dass sich andere Bezeichnungen besser eignen.
Die Komponente Empfindung resultiert aus unterschiedlichen Faktoren wie beispielsweise Alter, Geschlecht und Ausbildung (vgl. Iannario und Piccolo, 2012, S.234). Die Präferenz bei k Kategorien kann als Zählprozess aufgefasst werden. Wird als Träger $\{1, \ldots, k\}$ gewählt, kann die Präferenzkomponente mit der Shifted-Binomial Verteilung, welche durch den Parameter ξ definiert wird, dargestellt werden. Die Wahrscheinlichkeitsfunktion der Shifted-Binomial- Verteilung lautet

$$b_r(\xi) = \binom{k-1}{r-1} \xi^{k-r} (1-\xi)^{r-1}, \quad r = 1, \ldots, k.$$

Die Unsicherheit, welche als Bestandteil der Entscheidungsfindung betrachtet wird, kann unterschiedliche Ursachen haben. So können beispielsweise die Zeit für die Beantwortung, die Beteiligung an der Problemstellung, die Bereitschaft das Ergebnis zu verfälschen, die Verständlichkeit der Fragestellung, Müdigkeit und Langeweile Gründe für diese Unsicherheit sein. Die einfachste Lösung für die Parametrisierung der Unsicherheit ist die Annahme einer Gleichverteilung.
Im CUB-Modell wird das Antwortverhalten einer Person modelliert. Dabei wird keine Annahme über die Existenz von zwei Gruppen, eine sichere und eine unsichere Gruppe, getroffen. Stattdessen wird angenommen, dass jede Person eine Wahrscheinlichkeit π für Empfindung und eine Wahrscheinlichkeit $1 - \pi$ für Unsicherheit hat. Somit wird die Antwortfindung im CUB-Modell als Mischung zweier Komponenten aufgefasst.
Im Folgenden wird der ordinale Response des Individuums i ($i = 1, \ldots, n$) mit $R_i \in \{1, \ldots, k\}$ bezeichnet. Die erklärenden Variablen des Individuums i erhalten die Bezeichnungen $\mathbf{z}_i$ und $\mathbf{x}_i$.
Das von Iannario und Piccolo (2012) beschriebene CUB-Modell ist für $k > 3$ durch zwei Komponenten definiert:

- stochastische Komponente:

$$P(R_i = r)|\mathbf{z}_i, \mathbf{x}_i) = \pi_i \binom{k-1}{r-1} \xi_i^{k-r} (1-\xi_i)^{r-1} + (1-\pi_i)\frac{1}{k} \quad \text{, für } i = 1, 2 \ldots, n$$

- systematische Komponente:

$$\pi_i = \frac{1}{1+\exp(-\mathbf{z}_i^T\boldsymbol{\beta})} \qquad \xi_i = \frac{1}{1+\exp(-\mathbf{x}_i^T\boldsymbol{\gamma})} \quad \text{, für } i = 1, 2 \ldots, n$$

Die systematische Komponente verbindet die Parameter π und ξ mit den Kovariablen $(\mathbf{z}_i^T, \mathbf{x}_i^T)$. Neben der vorgeschlagenen Linkfunktion sind auch andere $\mathbb{R} \to [0,1]$ Funktionen möglich. Die Kovariablen in den zwei Komponenten können die Gleichen sein, sich überschneiden oder komplett unterschiedlich sein.

2.2.2 Das CUP-Modell

Eine flexiblere Modellierung als das CUB-Modell bietet das CUP-Modell. Dabei wird das Binomial-Modell, welches für die Empfindungskomponente verwendet wird, durch flexiblere ordinale Modelle ersetzt. Die in der Komponente der Unsicherheit verwendete Gleichverteilung wird auch im CUP-Modell beibehalten (vgl. Tutz et al., 2014). Die allgemeine Form des CUP-Modells ist

$$P(R_i = r|\mathbf{x}_i) = \pi_i P_M(Y_i = r|\mathbf{x}_i) + (1-\pi_i)P_U(U_i = r). \tag{2.3}$$

Die Variablen Y_i und U_i sind die unbeobachtbaren, zufälligen Variablen der Präferenz- und Unsicherheitskomponente, welche die Werte $\{1, \ldots, k\}$ annehmen können. Die Verteilung der Variable Y_i ist durch $P_M(Y_i = r|\mathbf{x}_i)$, ein beliebiges ordinales Regressionsmodell M, bestimmt, wie beispielsweise das kumulative Logit Modell:

$$\log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}\right) = \gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma} \quad r = 1, \ldots, k-1.$$

Weitere Modelle, welche zur Modellierung der latenten Variablen Y_i genutzt werden können, sind in Abschnitt 2.1 beschriebenen. Die Verteilung der latenten Variablen U_i entspricht einer diskreten Gleichverteilung. Somit gilt $P_U(U_i = r) = 1/k$.
Das CUB-Modell ist durch die Verwendung der Binomialverteilung ein Speziallfall des CUP-Modells. Im Gegensatz zur strikt unimodalen Binomialverteilung ermöglichen das kumulative und das Adjacent Categories Modell eine flexiblere Modellierung, da diese alle Verteilungsformen durch die Intercepts $\gamma_{01}, \ldots, \gamma_{0k}$ erlauben (vgl. Tutz et al., 2014, S.7).

Somit besteht auch im CUP-Modell, welches die ordinalen Modelle mit einer Unsicherheitskomponente verbindet, eine größere Flexibilität als im CUB-Modell.
Im CUB-Modell ist eine der systematischen Komponenten $logit(\xi_i) = \mathbf{x}_i^T\boldsymbol{\gamma}$. Im kumulativen und im Adjacent Categories Modell ist der Effekt der erklärenden Variablen im linearen Prädiktor $\eta_{ir} = \gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma}$ enthalten. Die Modellierung der Wahrscheinlichkeit π_i durch Kovariablen erfolgt wie im CUB-Modell mit

$$logit(\pi_i) = \mathbf{z}_i^T\boldsymbol{\beta}.$$

Dabei darf der Kovariablenvektor $\mathbf{z}_i$ dem Vektor $\mathbf{x}_i$ entsprechen. Um zu beschreiben, welches Modell für die Präferenz verwendet wird, erfolgt die Verwendung der Schreibweise aus Tutz et al. (2014). Dabei bezeichnet CUP(c) das CUP-Modell mit dem kumulierten Logit-Modell während CUP(a) das Modell mit dem Adjacent Categories Modell in der Präferenzkomponente abgekürzt.

2.2.3 Das BETAMIX()- Modell

Der angeführte Grund für die Verwendung der Gleichverteilung in der Unischerheitskomponente des CUP-Modells ist die Einfachheit (vgl. Tutz et al., 2014). Um dem bereits flexiblen CUP-Modell mehr Flexibilität zu verleihen kann die Gleichverteilung durch eine andere diskrete Verteilung ersetzt werden. Eine Möglichkeit bietet hierzu die Betabinomialverteilung.

Betabinomialverteilung

Für die betabinomialverteilte Variable Y mit dem Träger $\{1, \ldots, k\}$, $(Y \sim BetaBin(k, \alpha, \beta))$, ergibt sich die Dichte der verschobenen Betabinomialverteilung

$$f(y) = \begin{cases} \binom{k-1}{y-1} \frac{B(\alpha+y-1, \beta+k-y)}{B(\alpha,\beta)} & , y \in 1, \ldots, k \\ 0 & , \text{sonst,} \end{cases}$$

mit $\alpha, \beta > 0$ und mit der Betafunktion $B(\alpha, \beta)$, die wie folgt definiert ist

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

Mit $\mu = \alpha/(\alpha+\beta)$ und $\delta = 1/(\alpha+\beta+1)$ erhält man für den Erwartungswert und für die Varianz

$$E(Y) = (k-1)\mu \qquad Var(Y) = (k-1)\mu(1-\mu)[1+(k-2)\delta].$$

Aufgrund der Form der Betabinomialverteilung kann eine Tendenz zu den mittleren oder extremen Kategorien ermöglicht werden. Dafür wird für die Betabinomialverteilung $\mu = 0.5$ gewählt. Bei dieser Wahl ergibt sich für die restlichen Parameter

$$\alpha = \beta, \quad \delta = 1/(2\alpha+1)$$

$$E(Y) = (k-1)/2$$

$$Var(Y) = [(k-1)/4]\frac{2\alpha+k-1}{2\alpha+1}.$$

Für den extremen Fall $\alpha = 0$ ergibt sich

$$Var(Y) = \frac{(k-1)^2}{4},$$

was einer Zweipunktverteilung auf 0 und $k-1$ entspricht. Für die Verwendung der Parameter $0 < \alpha < 1$ hat die Wahrscheinlichkeitsverteilung der Betabinomialverteilung die Form eines „U"s.Die Wahrscheinlichkeit für das Beobachten der Randkategorien ist verglichen mit den mittleren Kategorien höher. In Abbildung 2.1 erkennt man, dass die Kurve mit steigendem α immer flacher wird. Für den Fall $\alpha = 1$ ist die Wahrscheinlichkeit für jede Kategorie gleich groß. Die Betabinomialverteilung mit $\alpha = 1$ und $\mu = 0.5$ entspricht somit der diskreten Gleichverteilung.

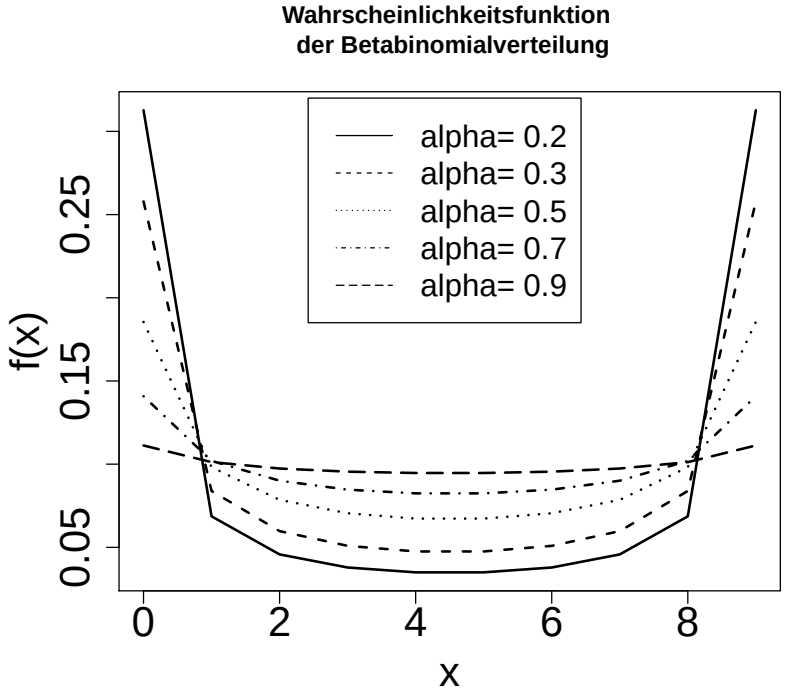

Abbildung 2.1: Wahrscheinlichkeitsfunktion der Betabinomialverteilung für α Werte zwischen 0 und 1

In Abbildung 2.2 sind die Wahrscheinlichkeitsfunktionen für fünf α- Werte > 1 dargestellt. Hier lässt sich erkennen, dass die Wahrscheinlichkeiten für die mittleren Kategorien größer sind als die der Randkategorien.

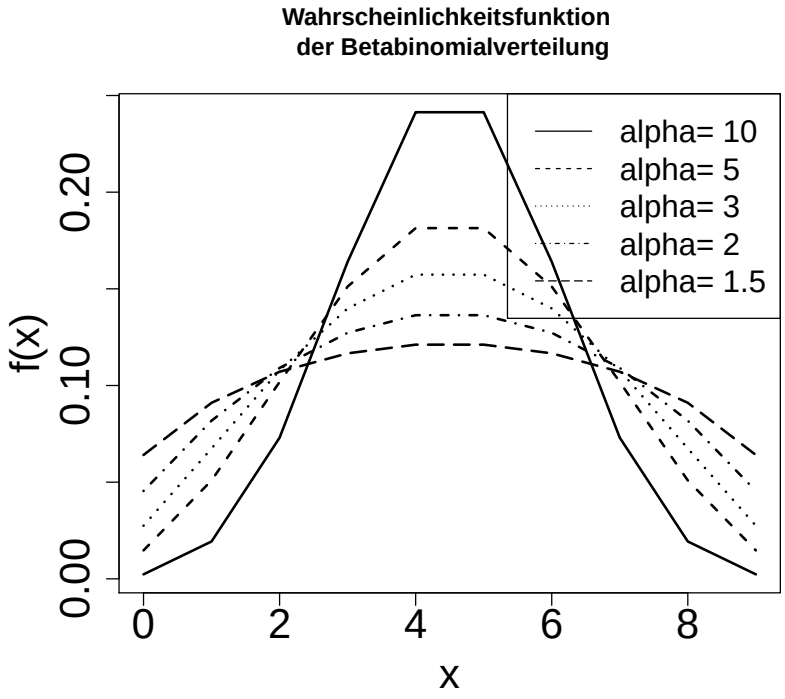

Abbildung 2.2: Wahrscheinlichkeitsfunktion der Betabinomialverteilung für α Werte größer 1

Für den Fall $\alpha \to \infty$ ist die Varianz definiert durch

$$Var(Y) = \frac{k-1}{4}$$

und man erhält somit als Grenzverteilung die Binomialverteilung.

BETMIX()-Modellgleichung

Das BETAMIX-Modell verwendet statt der Gleichverteilung die Betabinomialverteilung in der Unsicherheitskomponente. Damit ergibt sich folgende Modellgleichung:

$$P(R_i = r|\mathbf{x}_i) = \pi_i P_M(Y_i = r|\mathbf{x}_i) + (1-\pi_i)P_B(B_i = r). \tag{2.4}$$

Wie auch in den vorherigen Kapiteln steht R_i für den beobachteten Response. Die Variablen Y_i und B_i sind die unbeobachteten Zufallsvariablen, welche die Werte $\{1, \ldots, k\}$ annehmen

können.
Die Verteilung der Zufallsvariable Y_i wird durch $P_M(Y_i = r|\mathbf{x}_i)$ bestimmt und kann jedes ordinale Regressionsmodell sein. Die Verteilung der Variable B_i hingegen wird als Betabinomialverteilung festgelegt ($B_i \sim BetaBin(k, \alpha, \alpha)$). Aufgrund der Annahme $\mu = 0.5$ hat die Betabinomialverteilung nur einen Parameter α.

$$P_B(B_i = r) = \begin{cases} \binom{k-1}{r-1} \frac{B(\alpha+r-1,\alpha+k-r)}{B(\alpha,\alpha)} & , r \in 1, \ldots, k \\ 0 & , \text{sonst,} \end{cases}$$

Bei der Wahl des kumulativen Modells für die Mischkomponenten P_M wird das BETAMIX-Modell mit BETAMIX(c) bezeichnet. Das kumulative Modell hat die allgemeine Form (vgl. Unterabschnitt 2.1.1)

$$\log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}\right) = \gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma} \quad r = 1, \ldots, k-1.$$

Weiterhin besteht im BETAMIX-Modell die Möglichkeit die Mischwahrscheinlichkeit π_i in Abhängigkeit von Kovariablen $\mathbf{z}_i$ zu schätzen. Dafür wird der Logit-Link vorgeschlagen:

$$logit(\pi_i) = \mathbf{z}_i^T\boldsymbol{\beta}.$$

Neben der Anpassung eines BETAMIX-Modells mit festem α gibt es auch die Möglichkeit diesen Parameter mit Kovariablen durch den log- Link zu verknüpfen und somit das α zu schätzen:

$$\log(\alpha_i) = \mathbf{w}_i^T\boldsymbol{\alpha}.$$

Dabei wird mit $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$ der Koeffizientenvektor bezeichnet. Mit α_j wird somit der Effekt der j-ten Variablen auf den Parameter α, der die Betabinomialverteilung definiert, bestimmt. Mit $\mathbf{w}_i$ wird der Kovariablenvektor des Indivduums i bezeichnet. Wie in Abbildung 2.1 und Abbildung 2.2 zu sehen, bestimmt der α- Wert die Form der Wahrscheinlichkeitsform. Die Koeffizienten $\boldsymbol{\alpha}$ können somit im Zusammenhang mit einem Responsestyle interpretiert werden.
Aufgrund der hohen Anzahl an Koeffizienten im Falle der Verknüpfung von π und α mit Kovariablen wird nur einer dieser Parameter mit Kovariablen verbunden. Für den jeweils anderen Parameter wird ein Intercept-Modell verwendet.

2.3 Modellvergleich

Im folgenden Abschnitt werden Kriterien für den Vergleich von CUB-, CUP- und BETAMIX-Modellen aufgeführt. Im Allgemeinen sind diese Modelle nicht genestet, weshalb die Verwendung des Likelihood-Quotienten-Tests nicht möglich ist. Selbst bei dem Vergleich von einem kumulativen Mischmodell und einem reinen kumulativen Modell ist die Annahme der approximativen χ^2- Verteilung für die Teststatistik problematisch, da man am Rand des Parameterraums testet.
Bei dem Vergleich von Mischmodellen ist die Verwendung der Informationskriterien AIC und BIC üblich. AIC und BIC werden durch

$$AIC = -2l(\hat{\boldsymbol{\theta}}) + 2m \text{ und } BIC = -2l(\hat{\boldsymbol{\theta}}) + m\log(n)$$

definiert. Dabei ist n die Anzahl an Beobachtungen, m die Anzahl der Modellparameter und $l(\hat{\boldsymbol{\theta}})$ der Wert der log-Likelihoodfunktion für den geschätzten Wert des Parametervektors $\boldsymbol{\theta}$. Da die Informationskriterien zum Modellvergleich ohne genauere Begründung verwendet werden, sollten auch andere Möglichkeiten in Betracht gezogen werden.
Eine Alternative bietet die Betrachtung der Prognosegüte der Modelle. In Tutz et al. (2014) werden der log-Score und der Ranked-Score als Maße der Prognosegüte definiert.
Für den log- Score wird die Devianz benötigt, welche ein Maß der Diskrepanz zwischen Daten und Fit ist. Für einen multinomial-verteilten Response gibt es zwei Darstellungsformen der Devianz. Bei einer Gruppierung der Beobachtungen nach festen Werten der Kovariablen, erhält man die Verteilung $\mathbf{r}_i = (r_{i1}, \ldots, \mathbf{r}_{ik}) \sim M(n_i, \mathbf{p}_i)$ mit $i = 1, \ldots, N$. Dabei steht N für die Anzahl der unterschiedlichen Werte der Kovariablen. Die Anzahl der Beobachtungen des i-ten Werts der Kovariablen wird mit n_i und die wahren Wahrscheinlichkeiten werden mit $\mathbf{p}_i^T = (p_{i1}, \ldots, p_{ik})$ bezeichnet. Die Schätzung dieser Wahrscheinlichkeiten sind ohne die Annahme eines Modells die relativen Häufigkeiten $(f_{i1}, \ldots, f_{ik})$. Mithilfe dieser Bezeichnungen lässt sich die Devianz für die Multinomialverteilung mit

$$D = 2\sum_{i=1}^{N} n_i \sum_{r=1}^{k} f_{ir} \log\left(\frac{f_{ir}}{\hat{p}_{ir}}\right)$$

berechnen. Die zweite Darstellungsform der Devianz verwendet die einzelnen Beobachtungen $\mathbf{r}_i \sim M(1, \mathbf{p}_i), \quad i = 1, \ldots, n$. Für die Devianz wird somit die Gleichung

$$D = 2 \sum_{i=1}^{n} \sum_{l=1}^{k} r_{il} \log \left(\frac{r_{il}}{\hat{p}_{il}} \right) = -2 \sum_{i=1}^{n} \log(\hat{p}_{iR_i})$$

erhalten. Dabei bezeichnet $R_i \in \{1, \ldots, k\}$ die Beobachtung in den Kategorien. Für die Verwendung der Devianz als Maß für die Prognosegüte wird der Datensatz in einen Trainings- und Testdatensatz aufgeteilt. Das Modell wird mit dem Trainingsdatensatz gefittet. Die Berechnung der Devianz erfolgt auf dem Testdatensatz mit n_V Beobachtungen. Der log- Score ist die gemittelte Devianz

$$D/n_V = 2 \sum_{i=1}^{n_V} \log(\hat{p}_{iR_i^V}).$$

R_i^V ist der Response im Testdatensatz und $\hat{p}_{il}$ die mit den Kovariablenwerten des Testdatensatzes $\mathbf{x}_i^{(V)}$ geschätzte Wahrscheinlichkeit für Kategorie l.
Bei einem ordinalen Response kann man für Maße von dem in Gneiting und Raftery (2007) diskutierten „ranked probability score“ Ansatz Gebrauch machen. Dieser Ansatz verwendet im Gegensatz zum logScore die gesamte prädiktive Verteilung.
Mit $\hat{p}_i(r) = \hat{p}_{i1} + \ldots + \hat{p}_{ir}$ als geschätzte kumulative Wahrscheinlichkeit für den Wert $\mathbf{x}_i^{(V)}$ und $I(\cdot)$ als Indikatorfunktion ergibt sich für einen kategorialen Response der gemittelte Wert

$$L_{RPS}/n_V = \sum_{i=1}^{n_V} \sum_{r} (\hat{p}_i(r) - I(R_i \leq r))^2 / n_V$$

Dieses Maß berücksichtigt die Abweichung der gesamten geschätzten Verteilung von den beobachteten Werten. (vgl. Tutz et al., 2014, S.15f)

3 Parameterschätzung

Im Folgenden wird der Algorithmus für die Parameterschätzung in allgemeinen Mischmodellen mit m Mischkomponenten beschrieben. So kann dieser für alle in Abschnitt 2.2 beschrieben Modelle und für eventuelle Erweiterungen verwendet werden. Die Wahrscheinlichkeitsfunktion der Beobachtung R gegeben $\mathbf{x}$ wird im Allgemeinen durch

$$f(r|\mathbf{x}) = \sum_{j=1}^{m} \pi_j f_j(r|\mathbf{x}, \boldsymbol{\gamma}_j) \tag{3.1}$$

definiert. Für π_j muss $\sum_{j=1}^{m} \pi_j = 1$ und $0 \leq \pi_j \leq 1$ gelten. Die Dichten werden mit $f_j(r|\mathbf{x}, \boldsymbol{\gamma}_j)$ und die Mischverhältnisse der j-ten Komponente mit π_j bezeichnet. In den CUB-, CUP- und BETAMIX-Modellen sind jeweils nur zwei Mischkomponenten enthalten ($m = 2$). Bei der Verwendung der Gleichverteilung oder der Betabinomialverteilung mit festgelegtem Parameter α in der Unsicherheitskomponente wird statt $\boldsymbol{\gamma}_1$ und $\boldsymbol{\gamma}_2$ nur $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}$ benötigt. Wird statt der Fixierung des Parameters α in der Unsicherheitskomponente des BETAMIX-Modells eine Schätzung des Parameters vorgenommen, so entspricht $\boldsymbol{\gamma}_2 = \boldsymbol{\alpha}$. Für die Parameterschätzung in den betrachteten Modellen kann eine Maximierung der log-Likelihood

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \pi_j f_j(r_i|\mathbf{x}, \boldsymbol{\gamma}_j) \right)$$

erfolgen. Dabei beinhaltet der Parametervektor $\boldsymbol{\theta}$ alle Parameter und n sei die Anzahl der Beobachtungen. Da eine direkte Maximierung sehr zeitintensiv sein kann und der Erhalt stabiler Schätzungen zudem von der Wahl der Startwerte abhängt, bietet der Expectation-Maximation-Algorithmus (kurz: EM-Algorithmus) eine Alternative. Der EM-Algorithmus findet vor allem bei der Schätzung im Falle fehlender Daten Anwendung. In Mischmodellen ist die Zugehörigkeit einer Antwort r_i zu einer Mischkomponenten j unbekannt. Diese Zugehörigkeit wird im weiteren Verlauf mit der binären Variable

$$v_{ij} = \begin{cases} 1, & \text{Beobachtung } r_i \text{ stammt aus der } j\text{-ten Mischkomponenten} \\ 0, & \text{sonst} \end{cases}$$

bezeichnet. Der Vektor $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{im}) = (\ldots, 0, 1, 0, \ldots)$, welcher nur an der Position j eine Eins stehen hat, zeigt an, dass die Beobachtung r_i aus der j-ten Komponente stammt.

Für die Beobachtung r_i ergibt sich

$$f(r_i|v_{ij} = 1, \mathbf{x}_i, \theta) = f_j(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_j) = \prod_{l=1}^{m} f_l(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_l)^{v_{il}}.$$

Für die gesamte Dichte ergibt sich aufgrund der Multinomialverteilung von $\mathbf{v}_i^T$ mit dem Wahrscheinlichkeitsvektor $\boldsymbol{\pi}^T = (\pi_1, \ldots, \pi_m)$:

$$f(r_i, \mathbf{v}_i|\mathbf{x}_i, \boldsymbol{\theta}) = f(r_i|\mathbf{v}_i, \mathbf{x}_i, \boldsymbol{\theta}) f(\mathbf{v}_i|\boldsymbol{\theta}) = \prod_{j=1}^{m} f_j(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_j)^{v_{ij}} \prod_{j=1}^{m} \pi_j^{v_{ij}}.$$

Aus der Gesamtdichte ergibt sich die log-Likelihood

$$l_G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(f(r_i, \mathbf{v}_i|\mathbf{x}_i, \boldsymbol{\theta})) = \sum_{i=1}^{n} \sum_{j=1}^{m} v_{ij}(\log(\pi_j)) + \log(f_j(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_j)). \tag{3.2}$$

Für die Schätzung der Parameter in den Mischmodellen werden im EM-Algorithmus ein Expectation- und ein Maximierungs-Schritt durchgeführt. Im Expectation-Schritt erfolgt die Berechnung der bedingten Erwartung der Gesamtlikelihood, gegeben der beobachteten Daten $\mathbf{r}$ und der aktuellen Schätzung $\boldsymbol{\theta}^{(s)}$

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = E(l_G(\boldsymbol{\theta})|\mathbf{r}, \boldsymbol{\theta}^{(s)}).$$

Es genügt jedoch $E(v_{ij}|\mathbf{y}, \boldsymbol{\theta}^{(s)})$ zu berechnen, da die Gesamtlikelihood in den unbeobachteten Daten v_{ij} linear ist. Mithilfe des Bayes-Theorems lassen sich die Gewichte $\hat{v}_{ij}^s$ berechnen:

$$
\begin{aligned}
E(v_{ij}|\mathbf{y}, \boldsymbol{\theta}^{(s)}) &= f(v_{ij} = 1|r_i, \mathbf{x}_i, \boldsymbol{\theta}^{(s)}) \\
&= \frac{f(r_i|v_{ij} = 1, \mathbf{x}_i, \boldsymbol{\theta}^{(s)}) f(v_{ij} = 1|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})}{f(r_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})} \\
&= \frac{\pi_j f_j(r_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})}{f(r_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})} \\
&= \frac{\pi_j f_j(r_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})}{\sum_{l=1}^{m} \pi_l f_l(r_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})} \\
&= \hat{v}_{ij}^{s}
\end{aligned}
$$

Im Anschluss erfolgt ein Update der Parameter, indem Folgendes in der s-ten Iteration maximiert wird:

$$
M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}^{(s)} \log(\pi_j)}_{M_1} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}^{(s)} \log(f_j(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_j))}_{M_2}.
$$

Statt der Maximierung $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ können auch M_1 und M_2 maximiert werden. Dadurch werden die neuen Schätzungen

$$
\pi_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_{ij}^{(s)} \quad \text{und} \quad \boldsymbol{\gamma}_j^{(s+1)} = argmax_{\boldsymbol{\gamma}_j} \sum_{i=1}^{n} \hat{v}_{ij}^{(s)} \log(f_j(r_i|\mathbf{x}_i, \boldsymbol{\gamma}_j))
$$

erhalten.
Die in Kapitel 2 beschriebenen Modelle haben alle einen ordinalen Response. Die log-Likelihood wird in diesen Fällen durch das Summieren über die k Kategorien erhalten und entspricht somit nicht mehr genau Gleichung 3.2.
Im Folgenden wird der EM-Algorithmus für Mischmodelle mit kategorialem Response beschrieben:

- Input: Startwerte für den Parametervektor $\boldsymbol{\theta}^{(0)}$
- s- te Iteration:
  - Expectation-Schritt (E-Schritt):
    Berechnung der bedingten Erwartung $E(v_{ij}|\mathbf{x}_i, \boldsymbol{\theta}^{(s)}) = \hat{v}_{ij}^{s}$ mithilfe des Bayes-

Theorems:

$$\hat{z}_{ij}^{s} = \frac{\pi_j^{(s)} \sum_{r=1}^{k} f_j(r|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})}{\sum_{l}^{m} \pi_l \sum_{r=1}^{k} f_l(r|\mathbf{x}_i, \boldsymbol{\theta}^{(s)})}$$

– Maximization-Schritt (M-Schritt):
Update des Parametervektors $\boldsymbol{\theta}$, indem $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ maximiert wird.

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{v}_{ij}^{(s)} \left(\log(\pi_j)) + \sum_{r=1}^{k} \log(f_j(r|\mathbf{x}_i, \boldsymbol{\gamma}_j) \right)$$

Somit werden die neuen Schätzungen

$$\pi_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_{ij}^{(s)} \text{ und } \boldsymbol{\gamma}_j^{(s+1)} = argmax_{\boldsymbol{\gamma}_j} \sum_{i=1}^{n} \hat{v}_{ij}^{(s)} \sum_{r=1}^{k} \log(f_j(r|\mathbf{x}_i, \boldsymbol{\gamma}_j)$$

erhalten. Die Schätzung des Parametervektors $\boldsymbol{\gamma}_j^{(s+1)}$ kann mit bekannten Optimierungsalgorithmen erfolgen.

- Wiederholung der E- und M-Schritte bis sich die Parameterschätzung oder die log-Likelihood im Vergleich zum vorherigen Iterationsschritt kaum noch ändert.
Die Genauigkeit dieser Änderung kann zuvor festgelegt werden.

Bei der Hinzunahme von Kovariablen für die Spezifizierung der Wahrscheinlichkeit, dass Beobachtung i zu der j-ten Mischkomponente gehört, wird π_j durch

$$\pi_{ij} = \frac{1}{1 + \exp(-\mathbf{z}_i^T \boldsymbol{\beta}_j)}$$

ersetzt. Im Gegensatz zum Modell in Gleichung 3.1, welches für alle Beobachtungen ein konstantes π_j annimmt, erhält man durch die Hinzunahme der Kovariablen Wahrscheinlichkeiten, die von den individuellen Eigenschaften der Individuen abhängt.(vgl. Tutz et al., 2014, S.22f)

4 Simulationsstudie

Für die Simulationsstudie werden Datensätze gemäß dem BETAMIX(c)-Modell simuliert (vgl. Unterabschnitt 2.2.3).

Die allgemeine Form des kumulativen Modells, welches Bestandteil des BETAMIX(c)-Modells ist, muss für das Simulieren der Daten umgestellt werden. Der Grund hierfür ist die Modellformel des BETAMIX(c)-Modells:

$$P(R_i = r|\mathbf{x}_i) = \pi_i P_M(Y_i = r|\mathbf{x}_i) + (1 - \pi_i)P_B(B_i = r),$$

welche die Wahrscheinlichkeit $P_M(Y_i = r|\mathbf{x}_i)$ beinhaltet. Das kumulative Modell wird jedoch meist über die logarithmierten Chancen definiert. Um daraus $P(Y_i = r|\mathbf{x}_i)$ berechnen zu können sind folgende Umformungen nötig: Die Modellgleichung aus Gleichung 2.2 lässt sich auch wie folgt formulieren:

$$P(Y \leq r|\mathbf{x}_i^T) = \frac{\exp(\gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma})}$$

Die Wahrscheinlichkeiten $P(Y_i = r|\mathbf{x}_i)$ mit $r = 1, \ldots, k$ erhält man durch folgende Berechnungen:

$$P(Y_i = 1|\mathbf{x}_i) = P(Y_i \leq 1|\mathbf{x}_i)$$

$$P(Y_i = j|\mathbf{x}_i) = P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i) \text{ für } j = 2, \ldots, k-1$$

$$P(Y_i = k|\mathbf{x}_i) = 1 - P(Y_i \leq k-1|\mathbf{x}_i).$$

In den unterschiedlichen Szenarien der Simulationen werden die Parameter π und α, die Beobachtungsanzahl n, die Anzahl der Kategorien des ordinalen Response k, die Verteilungen der Kovariablen und die Koeffizientenvektoren variiert.
Mithilfe dieser simulierten Datensätze wird geprüft wie gut das BETAMIX(c)-Modell die Modellparameter schätzt. Für die Schätzung der Parameter mit dem BETAMIX-Modell werden die durch Micha Schneider zur Verfügung gestellten Funktionen, die dem elektronischen Anhang zu entnehmen sind, verwendet. Außerdem wird geprüft, ob die Wahl des Parameters α durch das Abtasten eines Gitters (α_G) oder durch die Schätzung $\hat{\alpha}_S$ im BETAMIX- Modell zu einem besseren Ergebnis führt. Das „bessere Ergebnis“ liefert das

Modell, dessen α eine geringere Abweichung zum wahren Wert α aufweist.
Die folgenden Abschnitte geben die zur Simulation der Daten verwendeten Parameter an und stellen die Ergebnisse vor.
In den Szenarien 1 bis 4 wird der Parameter α des BETAMIX(c)-Modells mit keinen Kovariablen verknüpft. Der Koeffizientenvektor $\boldsymbol{\alpha}$ hat somit nur einen Eintrag: α_0. In diesen ersten vier Szenarien wird statt dem Koeffizient α_0 der Wert $\exp(\alpha_0)$ betrachtet, da aus $\exp(\alpha_0)$ Rückschlüsse auf die Form der Betabinomialverteilung getroffen werden können. Aufgrund der im BETAMIX-Modell verwendeten Linkfunktion für den Parameter α wird $\exp(\alpha_0)$ mit in diesen Szenarien mit α bezeichnet.
Die Szenarien 5 und 6 verwenden Kovariablen zur Bestimmung des Parameters α der Betabinomialverteilung. Jeder Eintrag des zur Datengenerierung genutzten Koeffizientenvektors $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \ldots)$ wird dabei einzeln angegeben und betrachtet.
In Abschnitt 4.7 wird ein BETAMIX(c)-Modell mit einer Parametrisierung von α mit Kovariablen als Daten generierender Prozess gewählt. Dabei liegt das Interesse auf den Parameterschätzungen, wenn die Responsestyle-Komponente ignoriert wird.

4.1 Szenario 1

Für das erste Simulationsszenario wird eine stetig gleichverteilte Kovariable X_1 auf dem Intervall $[-3, 3]$ genutzt. Die ordinale Responsevariable hat vier Kategorien. Als wahres Modell wird das BETAMIX(c)-Modell gewählt. Es wird eine konstante Mischwahrscheinlichkeit $\pi = 0.7$ für die Simulation verwendet.
Es werden 30 Datensätze mit folgenden Koeffizienten simuliert:

- $\gamma_{01} = -3.5$
- $\gamma_{02} = -2.25$
- $\gamma_{03} = -1.35$
- $\gamma_1 = 0.912$.

Der Parameter α und die Beobachtungsanzahl n werden variiert. Es werden jeweils 30 Datensätze mit $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$ erstellt. Dies entspricht $\alpha_0 = \log(0.5)$, $\alpha_0 = \log(1)$ und $\alpha_0 = \log(3)$. Als Anzahl an Beobachtungen werden die Werte $n = 400$ und $n = 1000$ betrachtet. Im Folgenden wird zur Verkürzung des genauen Szenarios nachkommende Schreibweise verwendet: Szenario 1 $(\alpha; n)$. Damit wird die Zugehörigkeit der Ergebnisse

genau festgelegt.
Für die simulierten Datensätze wird das BETAMIX(c)-Modell angepasst. Als Kovariablenvektor **x** des BETAMIX(c)-Modells wird die Variable X_1 verwendet. Die Mischwahrscheinlichkeit und der Parameter α werden als konstant geschätzt. Wird der Parameter α durch eine Konstante geschätzt, so erfolgt dies mit dem Modell $\alpha = \exp(\hat{\alpha}_0)$. Die im Folgenden mit $\hat{\alpha}$ bezeichneten Werte entsprechen den Werten $\exp(\hat{\alpha}_0)$.

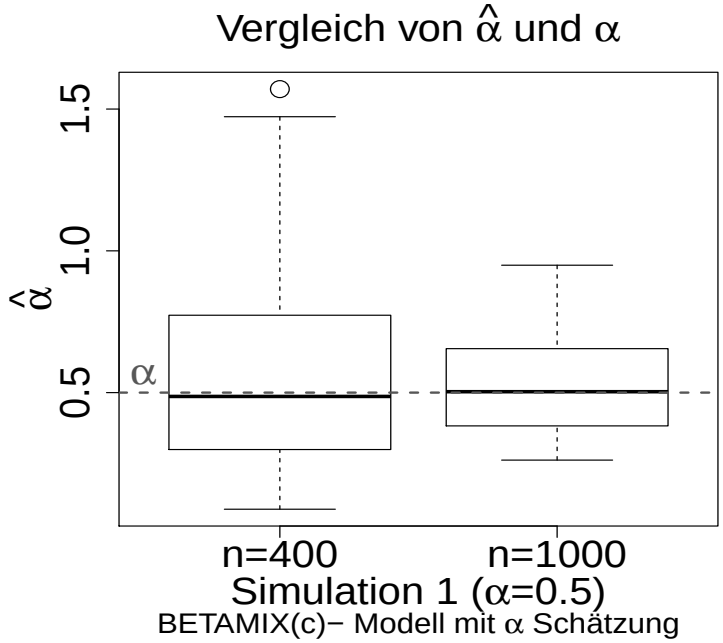

Abbildung 4.1: Szenario 1 ($\alpha = 0.5$): Boxplot des Parameters $\hat{\alpha}$

In Abbildung 4.1 sind die 30 geschätzten Werte für den Parameter α, welche sich durch die Schätzung $\hat{\alpha}_0$ ergeben, als Boxplot für $n = 400$ (links) und $n = 1000$ (rechts) dargestellt. Der wahre Wert $\alpha = 0.5$ wird durch die waagrechte graue Linie dargestellt.
Bei 400 Beobachtungen streuen die Schätzungen stärker als bei 1000 Beobachtungen. Die Werte $\exp(\alpha_0)$ liegen bei $n = 400$ im Intervall $([0.09; 1.57])$. Das Intervall, in welchem die Schätzungen für 1000 Beobachtungen liegen, ist hingegen kleiner $[0.26; 0.95]$.
Bei 1000 Beobachtungen entspricht der Median der $\hat{\alpha}$'s fast dem wahren Wert. Außerdem scheint die Verteilung der $\hat{\alpha}$-Schätzungen in diesem Fall symmetrisch zu sein. Abbildung 4.1 macht deutlich, dass die Schätzung des Parameters α bei den Datensätzen mit mehr Beobachtungen ein besseres Ergebnis liefert. Trotzdem lässt sich auch bei 400 Beobachtungen erkennen, dass der Median fast dem wahren Wert entspricht.

Der Interquartilsabstand bei 400 Beobachtungen ist zwar größer als bei 1000 Beobachtungen und die Verteilung der $\hat{\alpha}$ scheint eher linkssteil als symmetrisch, trotzdem ist die Abweichung des oberen und untere Quartils vom wahren Wert $\alpha = 0.5$ nicht allzu groß.

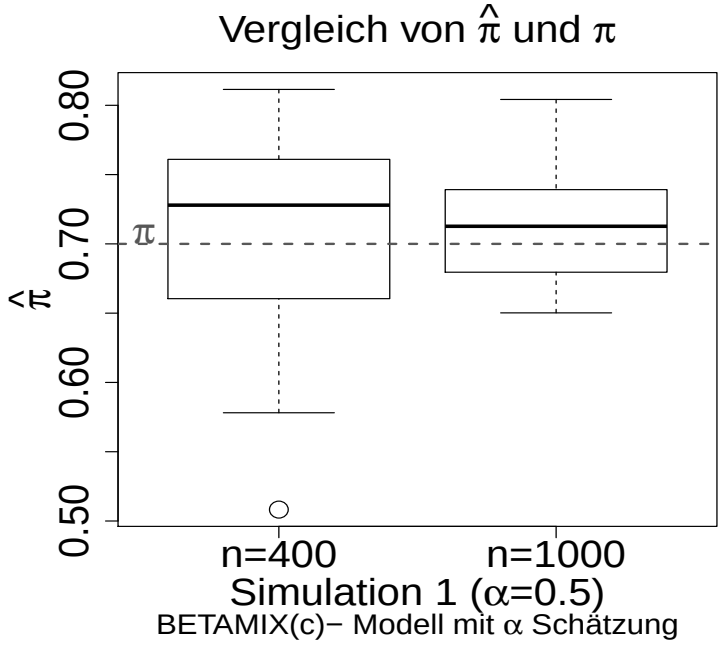

Abbildung 4.2: Szenario 1: Boxplot des Parameters $\hat{\pi}$

Der Median der geschätzten Mischwahrscheinlichkeiten $\hat{\pi}$ des BETAMIX(c)-Modells mit α-Schätzung weicht sowohl für $n = 400$ als auch für $n = 1000$ nicht stark vom wahren Wert $\pi = 0.7$ ab (vgl. Abbildung 4.2). Der Abbildung nach zu urteilen scheint die kleinere Beobachtungsanzahl kleinere Werte für $\hat{\pi}$ zu liefern als die größere Beobachtungsanzahl. Für $n = 400$ liegen ein viertel der kleinsten geschätzten Wahrscheinlichkeiten $\hat{\pi}$ zwischen 0.51 und 0.66. Bei 1000 Beobachtungen liegen die 25% der kleinsten geschätzten Wahrscheinlichkeiten im Intervall $[0.65; 0.68]$.
Das BETAMIX(c)-Modell mit einem Interceptmodell für α scheint die Koeffizienten γ für 1000 Beobachtungen gut zu schätzen (vgl. Abbildung 4.3 und Abbildung 4.4). Der Koeffizient γ_1 scheint nach Abbildung 4.4 zu urteilen leicht unterschätzt zu werden. Der betrachtete Wertebereich ist jedoch sehr klein.
Auch bei 400 Beobachtungen streuen die Koeffizientenschätzungen um die wahren Werte, wobei die Variabilität stärker ist als bei 1000 Beobachtungen (vgl. Abbildung 4.5 und

Abbildung 4.6).

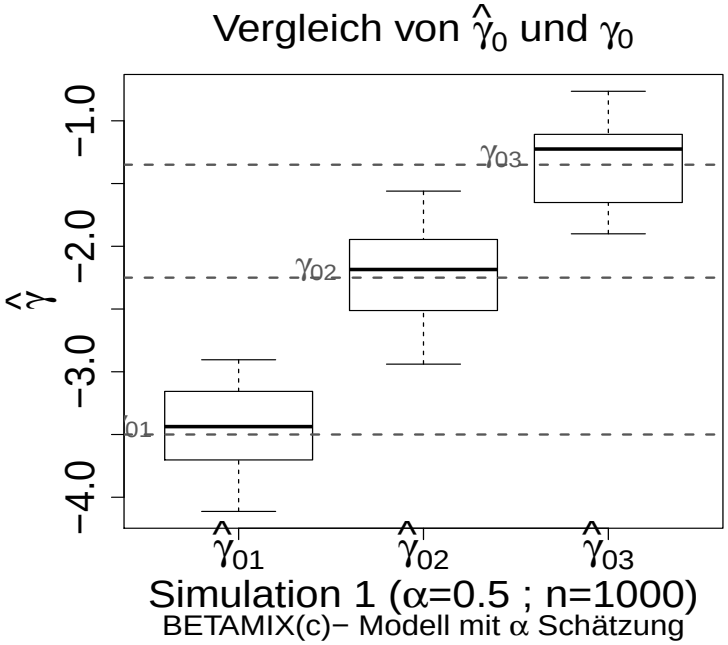

Abbildung 4.3: Szenario 1 ($\alpha = 0.5; n = 1000$): Boxplot des Koeffizientenvektors $\hat{\gamma}_0$

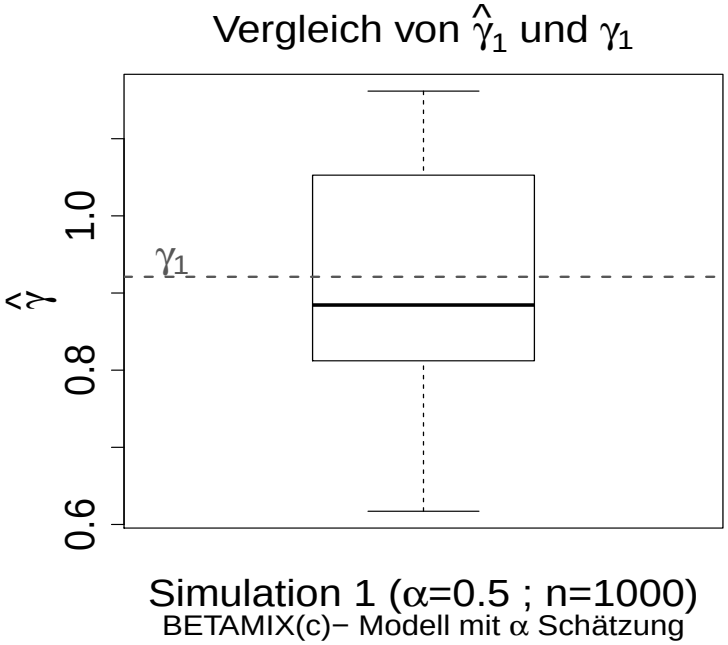

Abbildung 4.4: Szenario 1 ($\alpha = 0.5; n = 1000$): Boxplot des Koeffizienten $\hat{\gamma}_1$

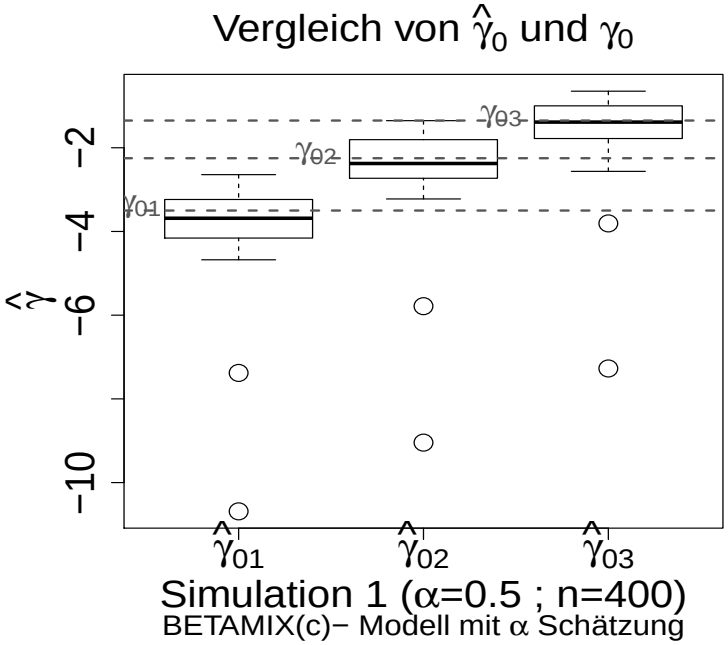

Abbildung 4.5: Szenario 1 ($\alpha = 0.5; n = 400$): Boxplot des Koeffizientenvektors $\hat{\gamma}_0$

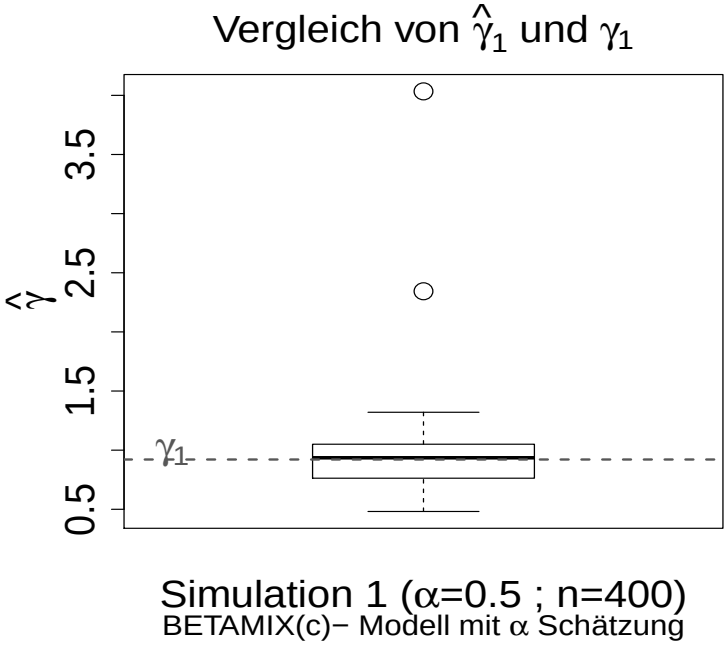

Abbildung 4.6: Szenario 1 ($\alpha = 0.5; n = 400$): Boxplot des Koeffizienten $\hat{\gamma}_1$

Nachstehen erfolgt die Betrachtung der Parameterschätzungen für die Datensätze, die den wahren Wert $\alpha = \exp(\alpha_0) = 1$ zur Datengenerierung verwenden. Es erfolgt wie zuvor eine Anpassung des BETAMIX(c)-Modells, welches die Variable X_1 für die Präferenzkomponente verwendet. Die Wahrscheinlichkeit π und der Parameter α der Betabinomialverteilung werden auch für diese Datensätze ohne den Einfluss von Kovariablen geschätzt.

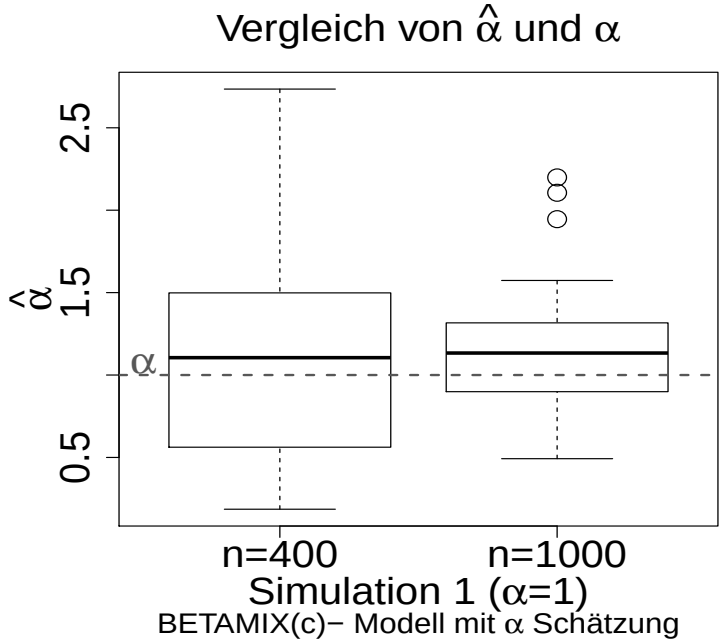

Abbildung 4.7: Szenario 1($\alpha = 1$): Boxplot des Parameters $\hat{\alpha}$

Abbildung 4.7 zeigt die geschätzten Parameter $\hat{\alpha}$ für die zwei unterschiedlichen Beobachtungsanzahlen, welche durch die Schätzungen mit dem BETAMIX(c)-Modell erhalten werden. Zwei der dreißig geschätzten $\hat{\alpha}$'s der Datensätze mit 400 Beobachtungen haben Werte größer als 10^5. Aus diesem Grund stellt der Boxplot in Abbildung 4.7 für $n = 400$ nur 28 geschätzte Werte dar. Für $n = 1000$ basiert der Boxplot auf dreißig Werten. Der Median beträgt für 400 Beobachtungen $\hat{\alpha}_{med} = 1.2$. Für die Datensätze mit 1000 Beobachtungen sind 50% der kleinsten geschätzten $\hat{\alpha}$ kleiner als 1.13. Die Interquartilsabstände der zwei betrachteten Boxplots unterscheiden sich stark voneinander. Im Vergleich zu den Schätzungen für die Datensätze mit $\alpha = 0.5$ scheinen hier die geschätzten Werte stärker von dem wahren Wert abzuweichen, was in Abbildung 4.9 nochmals aufgegriffen und mittels der mittleren quadratischen Abweichung genauer betrachtet wird.

In Abbildung 4.8 werden die geschätzten Werte $\hat{\alpha} = \exp(\alpha_0)$ betrachtet, die mittels der Anpassung des BETAMIX(c)-Modells für die Datensätze mit $\alpha = 3$ erhalten werden. In Abbildung 4.8 werden aufgrund einiger Ausreißer nur Werte kleiner als zehn betrachtet. Die zwei Boxplots stellen neunzehn Werte dar.

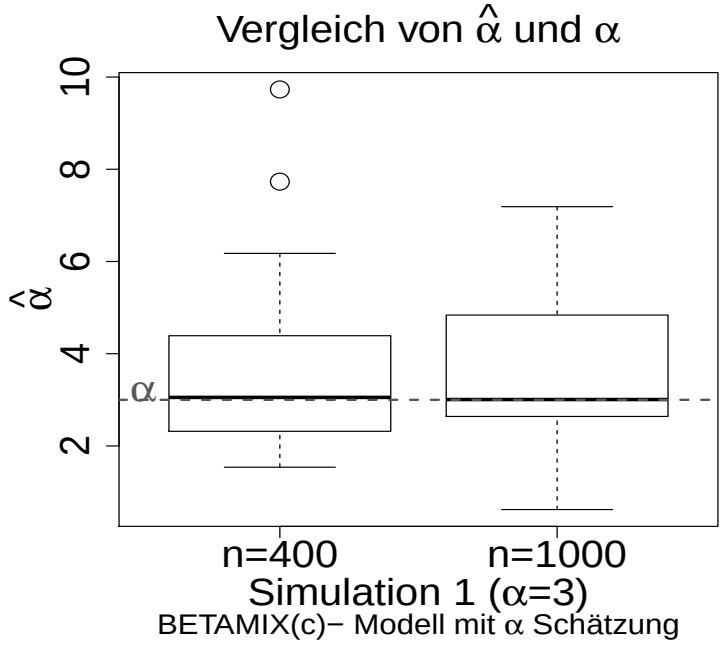

Abbildung 4.8: Szenario 3: Boxplot des Parameters $\hat{\alpha}$

Der Median $\hat{\alpha}_{med}$ dieser neunzehn Werte liegt sowohl für $n = 400$ als auch für $n = 1000$ annähernd bei 3, dem wahren Wert für α. Im Gegensatz zu den vorherigen Fällen mit $\alpha = 0.5$ und $\alpha = 1$ ist der Interquartilsabstand für $n = 1000$ größer als für $n = 400$.

In Abbildung 4.9 sind die mittleren quadratischen Abweichungen (MSE) der geschätzten $\hat{\alpha}$ von den wahren α-Werten dargestellt. Die $\hat{\alpha}$ stammen aus der Anpassung des BETAMIX(c)-Modells an die simulierten Datensätze des beschriebenen Szenarios mit 1000 Beobachtungen. In den drei Boxplots der Abbildung 4.9 sind nur 23 MSE- Werte dargestellt, da MSE-Werte größer als zehn nicht zur Erstellung der Boxplots aufgenommen wurden.

Die Berechnung der mittleren quadratischen Abweichung des Parameters α erfolgt durch folgende Berechnung:

$$MSE(\alpha) = (\hat{\alpha} - \alpha)^2.$$

Es ist zu erkennen, dass die Streuung der MSE-Werte bei $\alpha = 1$ und $\alpha = 3$ größer ist als bei $\alpha = 0.5$, was jedoch bei größeren wahren Werten nicht überrascht. Für den Wert $\alpha = 3$ liegt das untere Quartil bei 0.04 und das obere Quartil bei 1.61. Für $\alpha = 0.5$ liegen hingegen 50% der MSE-Werte zwischen 0.004 und 0.07.

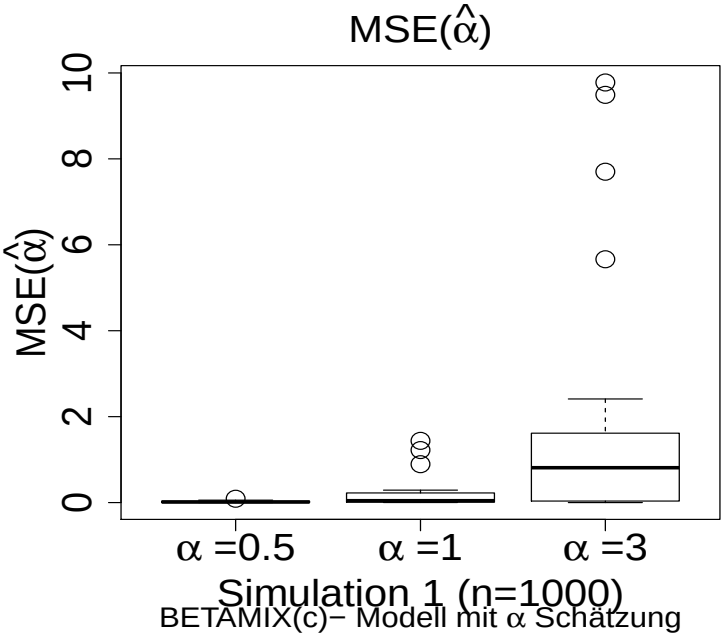

Abbildung 4.9: $MSE(\hat{\alpha})$ des Szenarios 1 ($n = 1000$) mit $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$

Weitere Abbildungen zu Szenario 1, wie beispielsweise die Schätzungen des Koeffizientenvektors und der Wahrscheinlichkeit π sind im Anhang zu finden (siehe Abbildungen A.1 bisA.6).

Neben der Möglichkeit der Schätzung des Parameters α im BETAMIX(c)-Modell, besteht auch die Möglichkeit der Wahl des Parameters α über das Abtasten eines Gitters. Dazu werden für Szenario 1 ($\alpha = 0.5$) folgende Werte für das α-Gitter vorgege-

ben: $0.1, 0.12, 0.14, \ldots, 0.98, 1, 1.1, 1.2, 1.3$. Für jeden dieser Werte wird ein BETAMIX(c)-Modell angepasst, in welchem das α fest ist. Außerdem wird nur der kumulative Teil des BETAMIX(c)-Modells mit Kovariablen parametrisiert. Die dafür verwendete Kovariable ist X_1. Die Wahrscheinlichkeit π wird als konstanter Wert geschätzt. Für die Wahl eines der α- Werte über das Gitter wird das AIC-Kriterium verwendet.
Das α, welches über das Gitter gewählt wird, erhält die Bezeichnung α_G. Der mit dem BETAMIX(c)-Modell zur Basis e exponierte Koeffizient $\hat{\alpha}_0$ wird mit $\hat{\alpha}_S$ bezeichnet. Somit gilt $\hat{\alpha}_S = \exp(\hat{\alpha}_0)$.

Abbildung 4.10 zeigt die geschätzten und die gewählten α-Werte für die Datensätze mit 1000 Beobachtungen. Außerdem ist der wahre Wert $\alpha = 0.5$ als roter Punkt in der Grafik kenntlich gemacht. Wie der Abbildung zu entnehmen ist, liegen die Werte auf der 1. Winkelhalbierenden. Somit liefern sowohl die Wahl über das Gitter als auch die Schätzung ein ähnliches Ergebnis für das α.

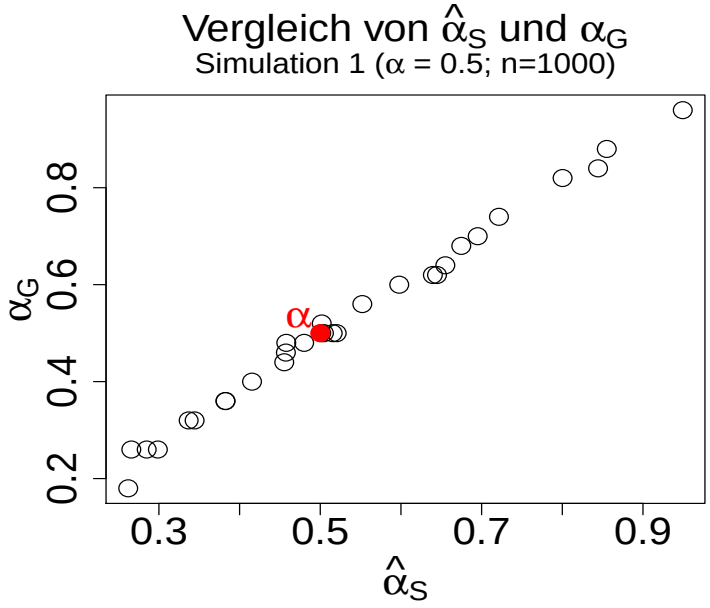

Abbildung 4.10: Szenario 1 ($\alpha = 0.5; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Auch für 400 Beobachtungen scheinen sich die gewählten α_G und die geschätzten $\hat{\alpha}_S$ nicht stark voneinander zu unterscheiden (vgl. Abbildung 4.11).

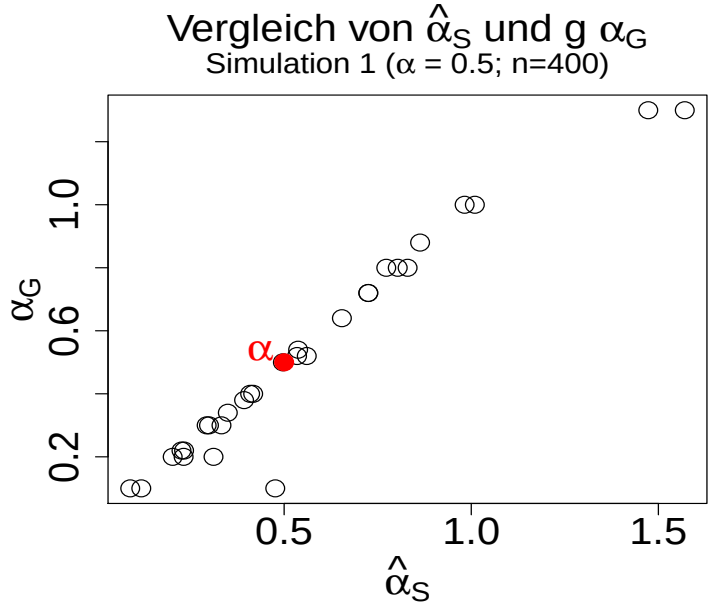

Abbildung 4.11: Szenario 1 ($\alpha = 0.5; n = 400$): Vergleich der geschätzten α_S und der gewählten α_G

4.2 Szenario 2

Das wahre Modell der simulierten Daten ist auch hier das BETAMIX(c)-Modell. Es werden fünf Kovariablen verwendet. Drei Kovariablen stammen aus der stetigen Gleichverteilung auf dem Intervall $[-3, 3]$. Die zwei weiteren Kovariablen sind binär. Folgende Aufzählung bietet eine Übersicht zu den Parametern, welche für die Simulation des aus zehn Kategorien bestehenden Response verwendet wurden.

- Anzahl der simulierten Datensätze $S = 30$
- Kategorienanzahl $k = 10$
- Mischwahrscheinlichkeit $\pi = 0.7$
- X_1 und X_2: zwei binäre Kovariablen ($X_1, X_2 \sim B(1, 0.7)$)
- drei auf dem Intervall $[-3, 3]$ stetig gleichverteilte Kovariablen X_3, X_4 und X_5

- Koeffizientenvektor:

$$\gamma = (\gamma_{01}, \ldots, \gamma_{09}, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (-4.5, -4, -3.5, -2.25, -1.95, 0.912, 1.25, 2, 2.5, 1.2, 1.5, -1.25, -2, 1)$$

Die Koeffizienten der binären Kovariablen sind γ_1 und γ_2. Szenario 2 verwendet die zuvor angeführten Parametereinstellungen für die Beobachtungszahlen $n = 400$ und $n = 1000$, sowie für die Werte $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$.
Wie in Szenario 1 erfolgt die Angabe der Beobachtungsanzahl und des Wertes für α nach der Szenario Nummer, beispielsweise Szenario 2 ($\alpha = 0.5; n = 400$).
Die geschätzten Werte für $\hat{\alpha}$, welche durch die Verwendung des BETAMIX(c)-Modells für die Daten aus Szenario 2 ($\alpha = 0.5$; $n = 1000$) und Szenario 2 ($\alpha = 0.5$; $n = 400$) erhalten werden, sind in Abbildung 4.12 dargestellt.
Für $n = 1000$ und $n = 400$ sind 75% der $\hat{\alpha}$ kleiner als 0.58. Außerdem weicht der Median $\hat{\alpha}_{med}$ für beide Beobachtungszahlen nur um einen kleinen Wert von $\alpha = 0.5$ ab. Der Median liegt für beide Beobachtungszahlen bei $\hat{\alpha}_{med} = 0.51$. Die Interquartilsabstände der $\hat{\alpha}$ für $n = 1000$ und $n = 400$ unterscheiden sich kaum.

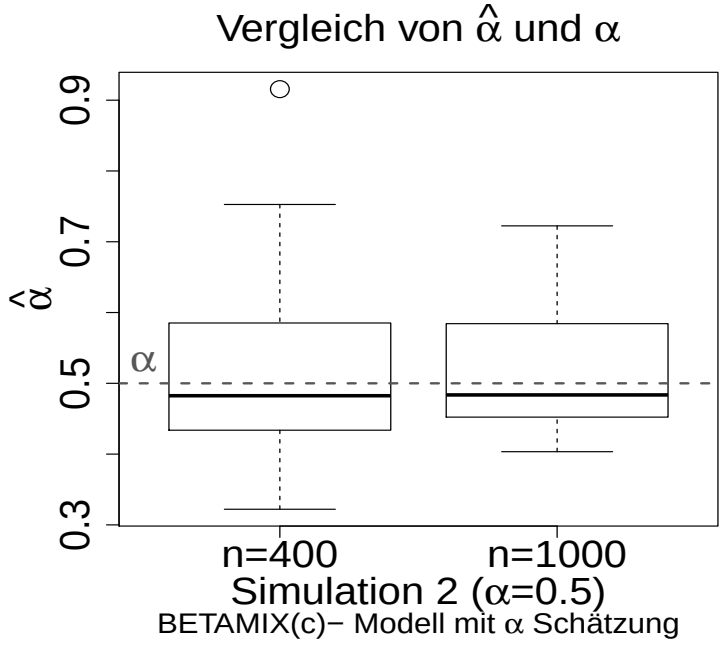

Abbildung 4.12: Szenario 2 ($\alpha = 0.5$): Boxplot des Parameters $\hat{\alpha}$

Die Verteilungen der $\hat{\pi}$ für 1000 und 400 Beobachtungen scheinen sehr symmetrisch um den wahren Wert $\pi = 0.7$ zu sein (siehe Abbildung 4.13). Der Interquartilsabstand der geschätzten Wahrscheinlichkeiten $\hat{\pi}$, welcher bei den Datensätzen mit 400 Beobachtungen erhalten wird, ist nur um 0.2 größer als der für 1000 Beobachtungen. Somit scheint die Wahrscheinlichkeit π auch für eine geringere Beobachtungsanzahl gut geschätzt werden zu können.

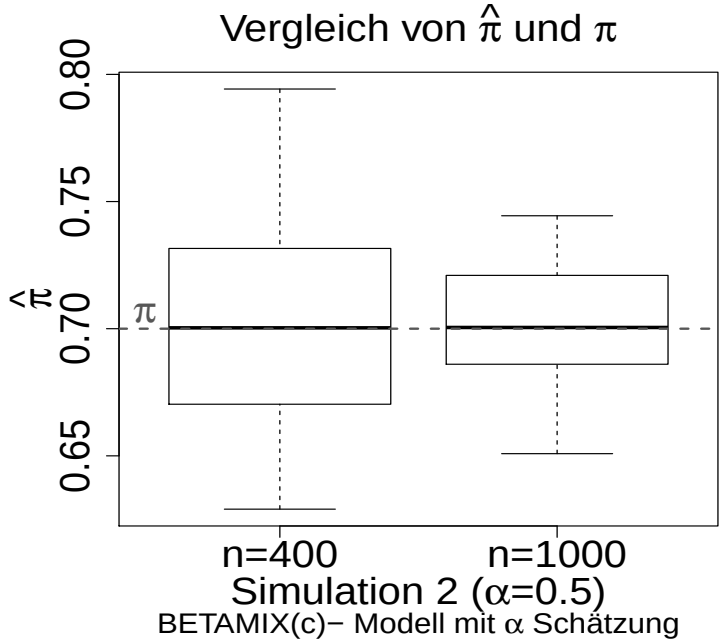

Abbildung 4.13: Szenario 2: Boxplot des Parameters $\hat{\pi}$

In Abbildung 4.14, Abbildung 4.15 und Abbildung 4.16 sind die Schätzung der Intercepts, welche mit dem BETAMIX(c)- Modell erhalten werden, dargestellt. Die Mediane der geschätzten Koeffizienten $\hat{\gamma}_{01}, \ldots, \hat{\gamma}_{09}$ entsprechen annähernd den wahren Werten. Die Schätzungen aller Intercepts scheinen jedoch tendenziell eher kleiner als die wahren Werte zu sein.

Für die binäre Variablen X_1 und X_2 scheinen die Schätzungen der zugehörigen Koeffizienten stärker von den wahren Werten abzuweichen als es bei den stetigen Variablen der Fall zu sein scheint (vgl. Abbildung 4.17). Der Koeffizient γ_1 scheint unterschätzt und γ_2 überschätzt zu werden. Aufgrund der Größenordnung lässt sich jedoch sagen, dass die Schätzungen aller Koeffizienten gut sind.

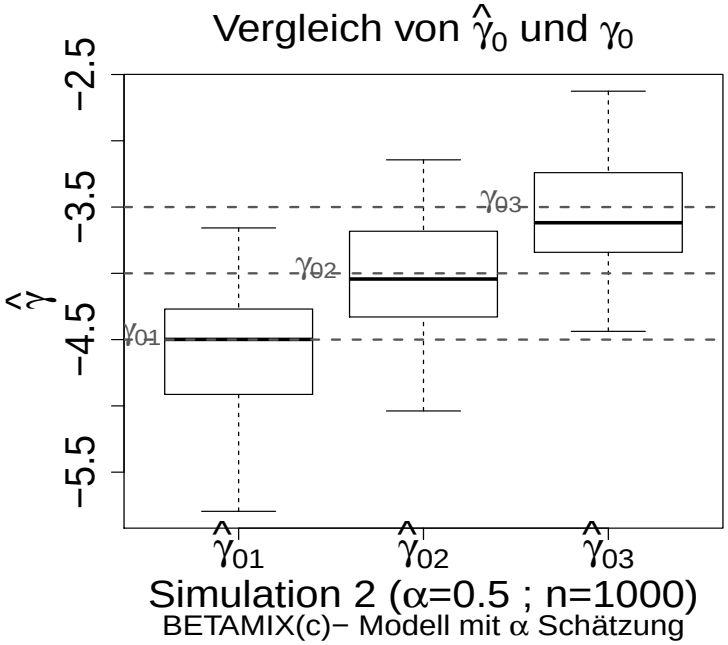

Abbildung 4.14: Szenario 2 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\hat{\gamma}_{01}, \hat{\gamma}_{02}, \hat{\gamma}_{03}$

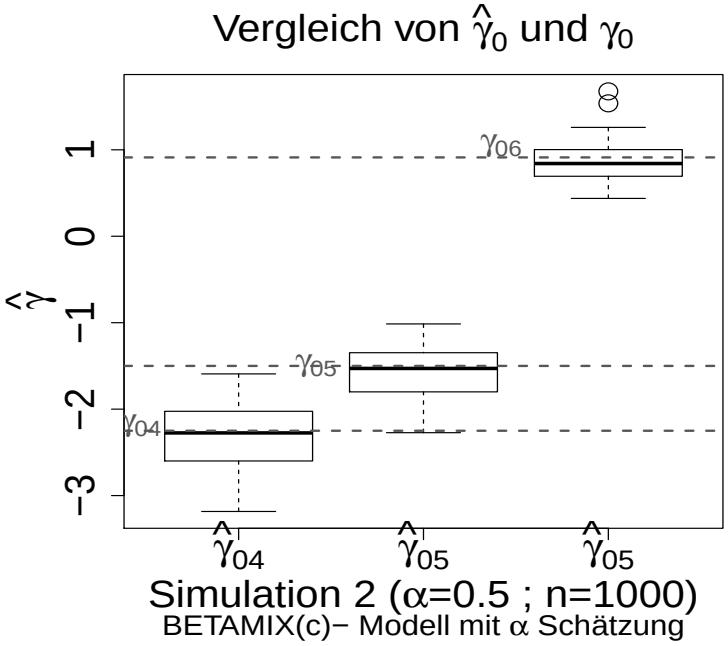

Abbildung 4.15: Szenario 2 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\hat{\gamma}_{04}, \hat{\gamma}_{05}, \hat{\gamma}_{06}$

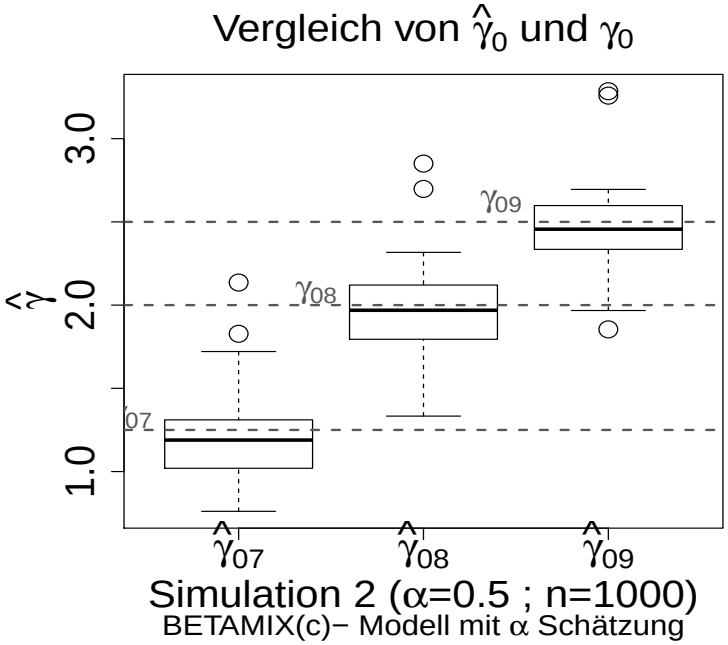

Abbildung 4.16: Szenario 2 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\hat{\gamma}_{07}, \hat{\gamma}_{08}, \hat{\gamma}_{09}$

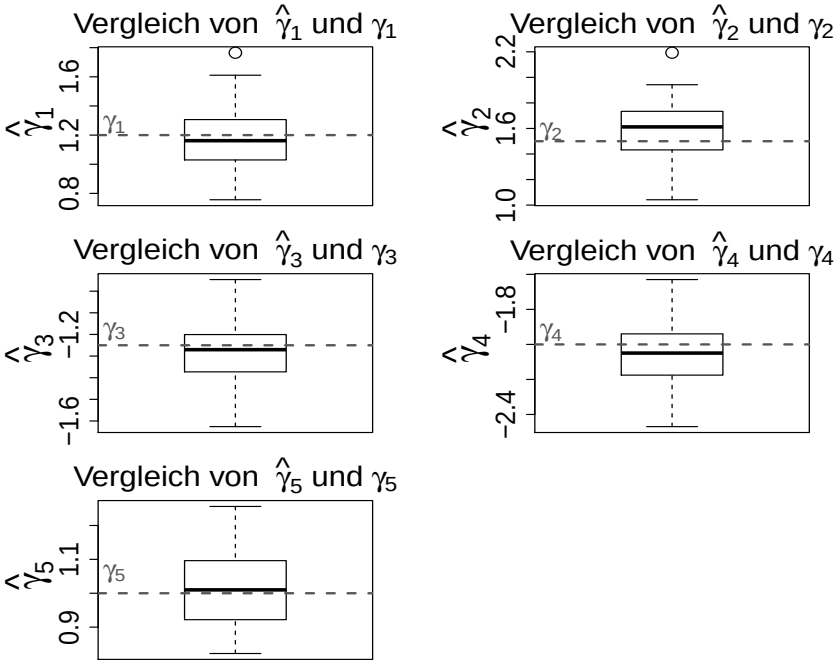

Abbildung 4.17: Szenario 2 ($\alpha = 0.5; n = 1000$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

Abbildung 4.18 vergleicht die α, welche durch die Wahl über das Abtasten eines Gitters und durch Schätzung erhalten werden. Für die Wahl des Parameters α wird, wie in Szenario 1, für jeden der folgenden achtundzwanzig Werte $0.1, 0.2, \ldots, 1.9, 2, 2.5, 3, 3.5, 4, 5, 10, 15, 20$ das BETAMIX(c)-Modell mit den Kovariablen $X_1, \ldots, X_5$ als Kovariablen für den kumulativen Teil des Mischmodells angepasst. Die Strukturkomponente der Mischwahrscheinlichkeit π entspricht einem Interceptmodell.
Anschließend werden die AIC-Werte dieser achtundzwanzig Modelle verglichen. Das α des Modells, welches den kleinsten AIC- Wert aufweist, wird in Abbildung 4.18 für den Wert α_G genutzt. Das geschätzte α stammt aus dem BETAMIX(c)-Modell, welches das Interceptmodell als Strukturkomponente der Mischwahrscheinlichkeit und des Parameters α verwendet. Die Variablen $X_1, \ldots, X_5$ dienen dem kumulativen Teil dieses BETAMIX(c)-Modells als Kovariablen.

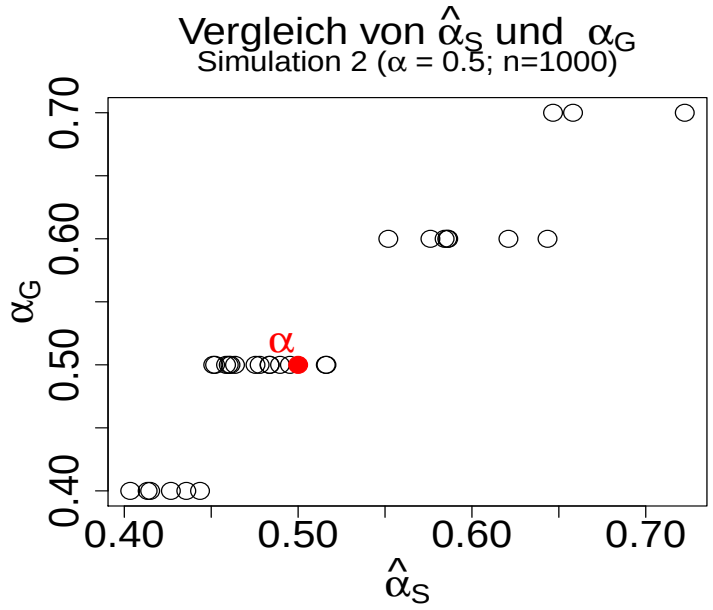

Abbildung 4.18: Szenario 2 ($\alpha = 1; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Der Abbildung nach zu urteilen scheint bis auf wenige Ausnahmen der gewählte Wert α_G dem auf eine Nachkommastelle (was dem gewählten Gitter zuzuschreiben ist) geschätzten $\hat{\alpha}_S$ zu entsprechen.
Bei den Datensätzen mit 400 Beobachtungen liegen sowohl Werte für $\hat{\alpha}_S$ als auch für α_G

außerhalb des Intervalls $[0.4, 0.75]$. Bei den Datensätzen mit 1000 Beobachtungen liegen alle Werte für $\hat{\alpha}_S$ und α_G innerhalb des zuvor angeführten Intervalls.
Letztlich ist jedoch, wie auch in Abbildung 4.18, erkennbar, dass sich geschätzte und gewählte α kaum unterscheiden.

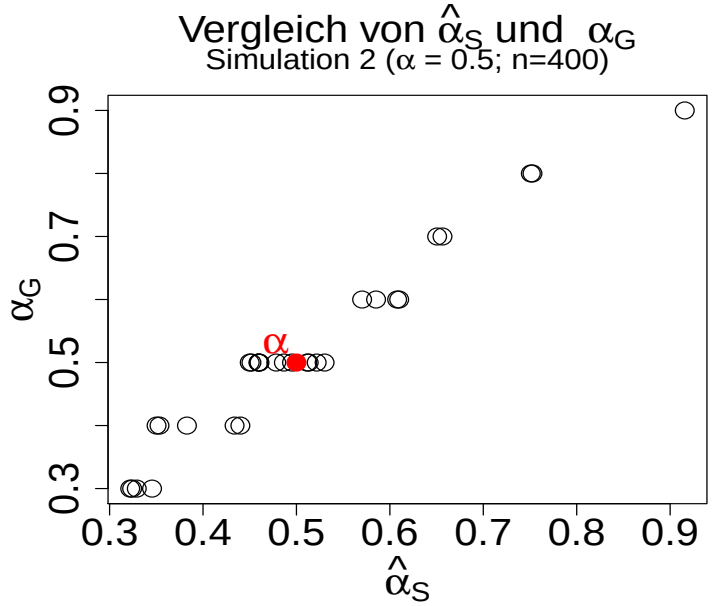

Abbildung 4.19: Szenario 2 ($\alpha = 0.5; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Die geschätzten Werte $\hat{\alpha}$ für die 30 Datensätze des Szenarios 2 ($\alpha = 1; n = 1000$) und des Szenarios 2 ($\alpha = 1; n = 400$) sind in Abbildung 4.20 dargestellt. Bei 400 Beobachtungen liegen die Hälfte der geschätzten $\hat{\alpha}$ zwischen 0.85 und 1.07. Bei den Datensätzen mit 1000 Beobachtungen liegen 50% der geschätzten $\hat{\alpha}$ im Intervall $[0.87; 1.12]$. Bei beiden Beobachtungszahlen streuen die geschätzten Werte um den wahren Wert $\alpha = 1$.

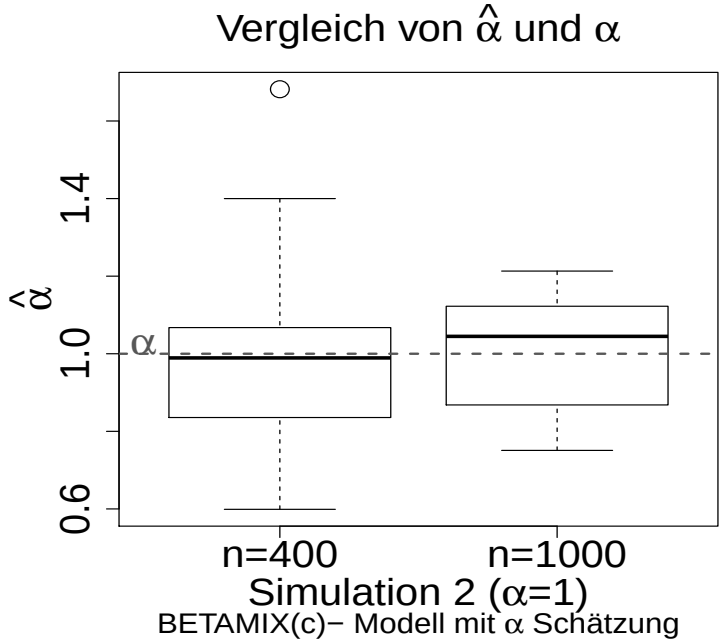

Abbildung 4.20: Szenario 2 ($\alpha = 1$): Boxplot des Parameters $\hat{\alpha}$

Die Schätzungen des Parameters α des Szenarios 2 ($\alpha = 3$) weichen im Gegensatz zu den Schätzungen der Szenarien 2 ($\alpha = 0.5$) und 2 ($\alpha = 1$) stark vom wahren Wert ab. Dies kann man den $MSE(\hat{\alpha})$-Werten entnehmen, welche in Abbildung 4.21 dargestellt sind.

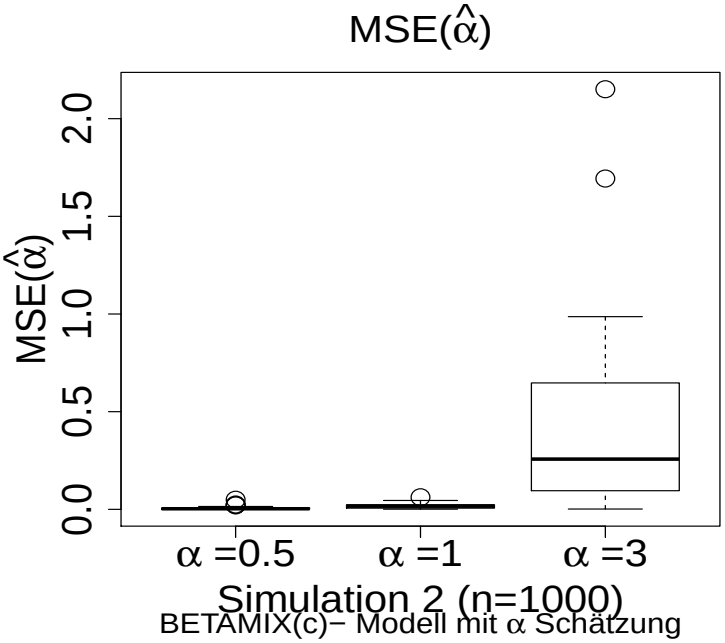

Abbildung 4.21: $MSE(\hat{\alpha})$ des Szenarios 2 ($n = 1000$) mit $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$

Die Schätzungen der Wahrscheinlichkeit $\hat{\pi}$ weichen für das Szenario 2 ($\alpha = 1$) sowohl bei $n = 400$ als auch für $n = 1000$ kaum vom wahren Wert $\pi = 0.7$ ab (vgl. Abbildung 4.22). Dies ist auch bei Szenario 2 ($\alpha = 3; n = 400$) und Szenario 2 ($\alpha = 3; n = 1000$) zu beobachten (siehe Abbildung 4.23).

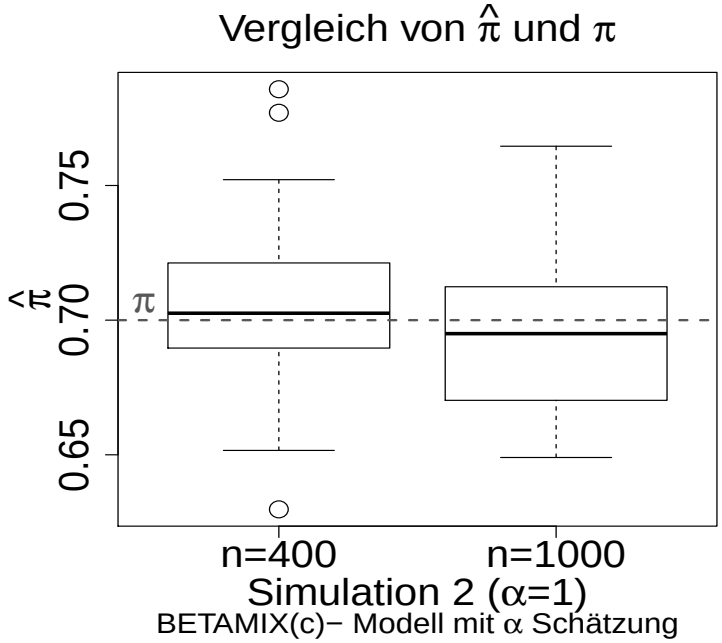

Abbildung 4.22: Szenario 2 ($\alpha = 1$): Boxplot des Parameters $\hat{\pi}$

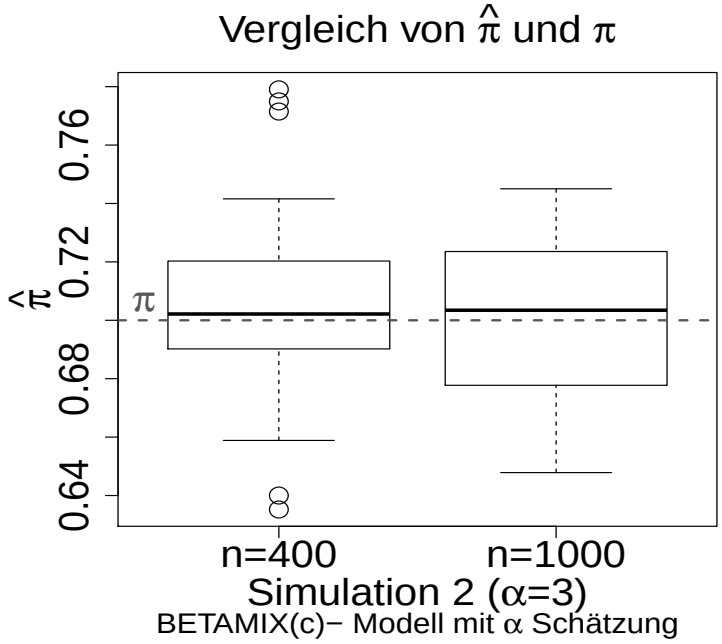

Abbildung 4.23: Szenario 2 ($\alpha = 3$): Boxplot des Parameters $\hat{\pi}$

Wie bei den Datensätzen mit $\alpha = 0.5$ und 1000 Beobachtungen, scheinen sich auch gewählte α_G und geschätzte $\hat{\alpha}_S$ für Szenario 2 ($\alpha = 1; n = 1000$), Szenario 2($\alpha = 1; n = 400$), Szenario 2 ($\alpha = 3; n = 1000$) und Szenario 2 ($\alpha = 3; n = 400$) kaum zu unterscheiden. Die entsprechenden Abbildungen sind im Anhang zu finden (vgl. Abbildung A.11 bis Abbildung A.14).

Für die Szenarien 1 und 2, welche beide nur den kumulativen Teil des BETAMIX(c)-Modells mit Kovariablen parametrisieren, können basierend auf den durchgeführten Simulationen folgende Aussagen getroffen werden: Die Schätzungen des Parameters α scheinen bei einem wahren Wert von $\alpha = 0.5$ oder $\alpha = 1$ besser als bei $\alpha = 3$ zu sein. Bei dem größeren Wert $\alpha = 3$ ist eine größere Variabilität zu beobachten, trotzdem scheinen die Schätzungen gut zu sein. Die Schätzungen des Parameters π und des Koeffizientenvektors γ scheinen für alle drei α- Werte gut zu funktionieren, da nur kleinere Abweichungen zu beobachten sind. Bei den Schätzungen für die Datensätze mit 400 Beobachtungen ist meist eine größere Streuung der Schätzwerte zu verzeichnen. Trotzdem scheinen 400 Beobachtungen für gute Schätzungen auszureichen.
Wählt man den Parameter α des BETAMIX(c)- Modells über ein Gitter, so erhält man einen Wert α_G, welcher ungefähr dem durch das BETAMIX(c)-Modell geschätzte $\hat{\alpha} = \exp(\hat{\alpha}_0)$ entspricht. Da die Wahl des α's über das Abtasten eines Gitters die Anpassung mehrerer BETAMIX(c)-Modelle erfordert und dies unter Umständen sehr zeitintensiv sein kann, ist aufgrund der ähnlichen Werte, die erhalten werden, die Schätzung des Parameters α innerhalb des BETAMIX(c)-Modells zu bevorzugen.

4.3 Szenario 3

Die Daten, die in Szenario 3 simuliert werden, stammen aus einem BETAMIX(c)-Modell. Die Parametrisierung der Mischwahrscheinlichkeit π erfolgt in Szenario 3 mit einer standardnormalverteilten Kovariable Z_1. Für den kumulativen Teil des BETAMIX(c)- Modells wird eine auf dem Intervall $[-3; 3]$ stetig gleichverteilte Kovariable X_1 verwendet.
Die Kategorienanzahl der ordinalen Responsevariablen wird auf sieben festgelegt. Die Beobachtungsanzahl n und der Parameter α werden variiert. Dabei werden als Beobachtungsanzahl die Werte $n = 400$ und $n = 1000$ verwendet. Außerdem werden die Werte

$\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$ betrachtet. Weitere zur Simulation verwendete Werte sind im Folgenden aufgelistet:

- Anzahl der simulierten Datensätze $S = 30$
- Kategorienanzahl $k = 7$
- eine auf dem Intervall $[-3, 3]$ stetig gleichverteilte Kovariablen X_1
- eine normalverteilte Kovariablen Z_1 mit dem Erwartungswert 0 und der Varianz 1
- Koeffizientenvektor γ :

$$\begin{aligned}\gamma &= (\gamma_{01}, \ldots, \gamma_{06}, \gamma_1,) \\ &= (-4.5, -3.25, -2, \\ &\qquad 1.25, 2, 3.1, \\ &\qquad -1.9)\end{aligned}$$

- Koeffizientenvektor $\boldsymbol{\beta}$:

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_0, \beta_1,) \\ &= (1.4, -1.51)\end{aligned}$$

Es wird für die jeweils 30 Datensätze mit gleicher Beobachtungsanzahl und gleichem α ein BETAMIX(c)-Modell angepasst. Für den Kovariablenvektor $\mathbf{x}$, der zu dem kumulativen Teil des Mischmodells gehört, wird die Variable X_1 verwendet. Für den Vektor $\mathbf{z}$, welcher die Wahrscheinlichkeit π parametrisiert, wird die Variable $Z1$ verwendet. Die Schätzung des Parameters α erfolgt ohne Kovariablen. Die Schätzungen von α scheinen zu ähnlichen Ergebnissen wie in Szenario 1 und Szenario 2 zu führen. Die Streuung der $\hat{\alpha}$'s scheint bei den Datensätzen mit $n = 400$ und $\alpha = 3$ am größten zu sein. In allen Fällen scheinen die geschätzten $\hat{\alpha}$'s um den jeweils wahren Wert von α zu streuen (vgl. Abbildung 4.24, Abbildung 4.25 und Abbildung 4.26).

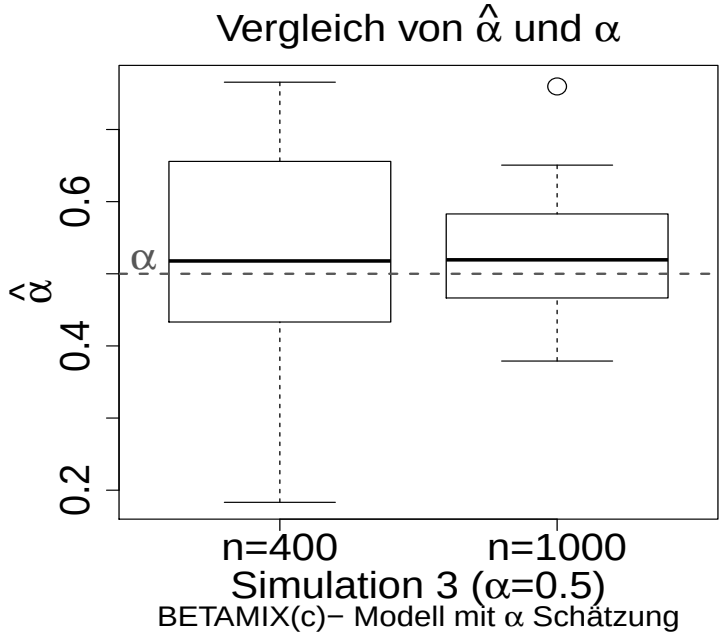

Abbildung 4.24: Szenario 3 ($\alpha = 0.5$): Boxplot des Parameters $\hat{\alpha}$

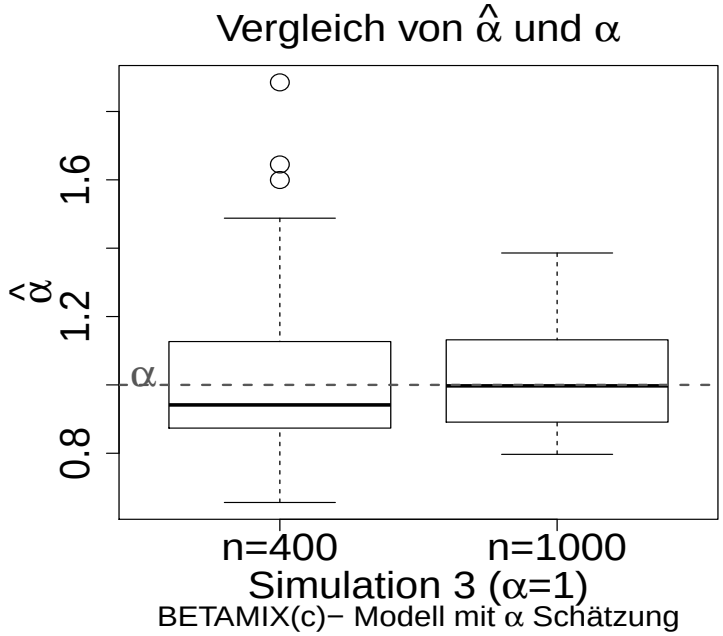

Abbildung 4.25: Szenario 3 ($\alpha = 1$): Boxplot des Parameters $\hat{\alpha}$

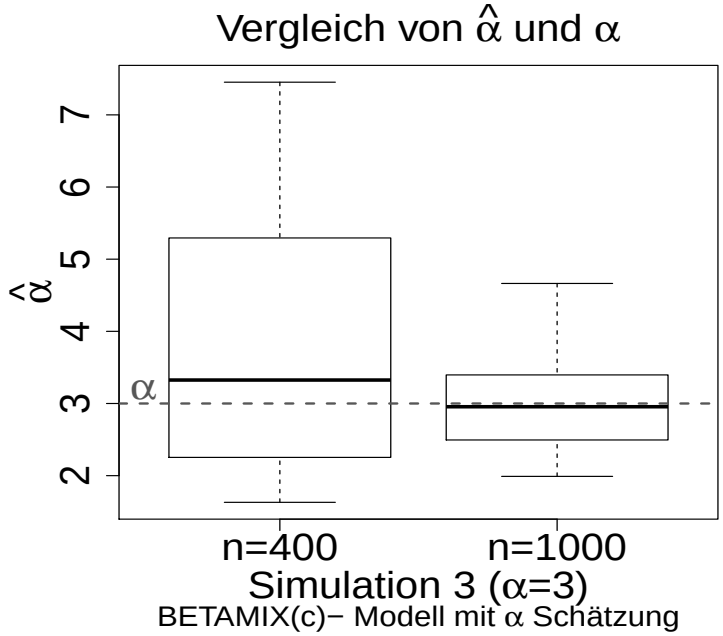

Abbildung 4.26: Szenario 3 ($\alpha = 3$): Boxplot des Parameters $\hat{\alpha}$

In Abbildung 4.27 sind die 30 geschätzten Werte $\hat{\gamma}_1$ des Szenarios 3 ($\alpha = 3$) für 400 und 1000 Beobachtungen dargestellt. Der wahre Wert $\gamma_1 = -1.9$ ist in der Grafik mit einer dunkelgrauen gestrichelten Linie eingezeichnet. Für $n = 400$ scheint $\hat{\gamma}_1$ unterschätzt zu werden, da ungefähr 75% der geschätzten Werte unter dem wahren Wert γ_1 liegen. Bei 1000 Beobachtungen ist zu erkennen, dass der Median fast dem wahren Wert $\gamma_1 = -1.9$ entspricht und dass die Abweichung der Schätzungen vom wahren Wert maximal ± 0.3 beträgt.

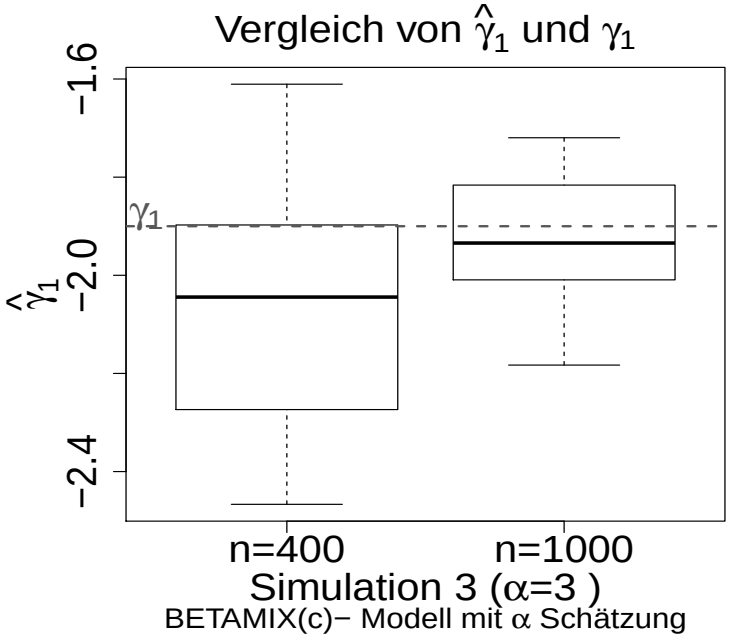

Abbildung 4.27: Szenario 3 ($\alpha = 3$): Boxplot des Parameters $\hat{gamma}_1$

Für den Koeffizientenvektor $\boldsymbol{\beta}$ werden in diesem Szenario die Werte $(1.4, -1.51)$ gewählt. Die Abbildungen 4.28 und 4.29 stellen die Schätzungen für β_0 und β_1 dar, welche die Mischwahrscheinlichkeiten π_i bestimmen. Die Variabilität der β_0 Schätzungen ist bei $n = 400$ etwas größer als bei mehr als doppelt so vielen Beobachtungen. Die geschätzten $\hat{\beta}_0$ Werte sind bei beiden Beobachtungszahlen kaum verzerrt (vgl. Abbildung 4.28). Auch die Schätzung des Koeffizienten $\hat{\beta}_1$ für die Kovariable Z_1 scheint für die zwei betrachteten Beobachtungszahlen kaum vom wahren Wert abzuweichen. Die Verteilung der $\hat{\beta}_1$ ist für $n = 400$ und $\alpha = 3$ rechtssteil. Für $n = 1000$ und $\alpha = 3$ scheint die Verteilung des Koeffizienten $\hat{\beta}_1$ eher linkssteil (vgl. Abbildung 4.29).

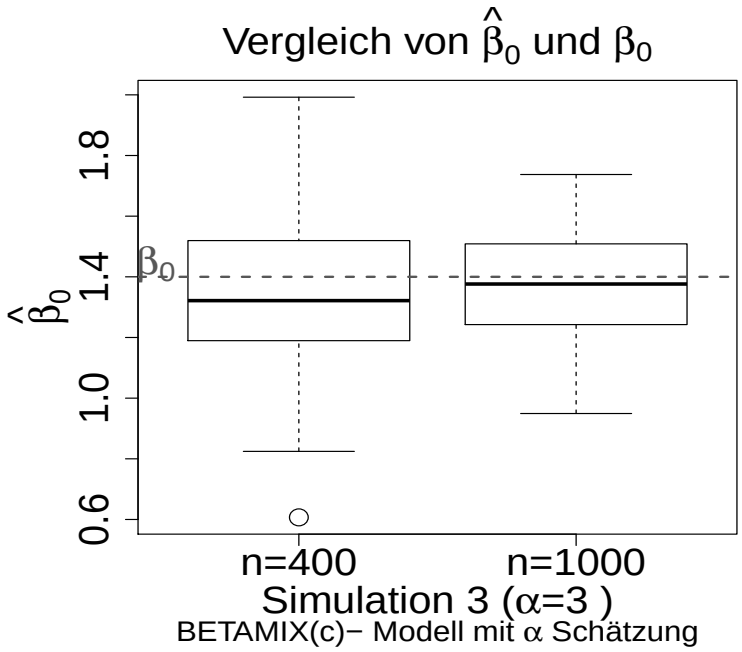

Abbildung 4.28: Szenario 3 ($\alpha = 3$): Boxplot des Parameters $\hat{\beta}_0$

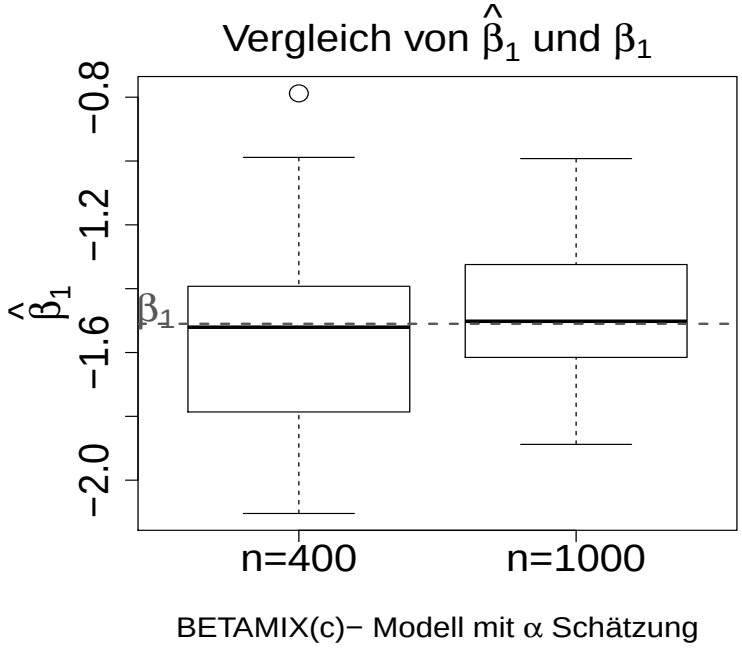

Abbildung 4.29: Szenario 3 ($\alpha = 3$): Boxplot des Parameters $\hat{\beta}_1$

Die Schätzungen der Koeffizienten des Szenarios 3 mit $\alpha = 0.5$ und $\alpha = 1$ scheinen ebenfalls kaum verzerrt zu sein. Die zugehörigen Boxplots sind im Anhang zu finden. Auch die Boxplots der geschätzten Intercepts des Szenarios 3 sind dem Anhang zu entnehmen. Auch bei den simulierten Datensätzen des Szenarios 3 scheint die Wahl des α's über ein Gitter mit vorgegebenen festen Werten für α unnötig, da sich die Werte des gewählten α_G's und des geschätzten $\hat{\alpha}_S = \exp(\hat{\alpha}_0)$ nicht merklich voneinander unterscheiden.
Für die Wahl des α's wird in Szenario 3 ($\alpha = 0.5, n = 1000$) das Gitter
$0.1, 0.12, 0.14, \ldots, 0.98, 1, 1.1, 1.2, 1.3$ genutzt.

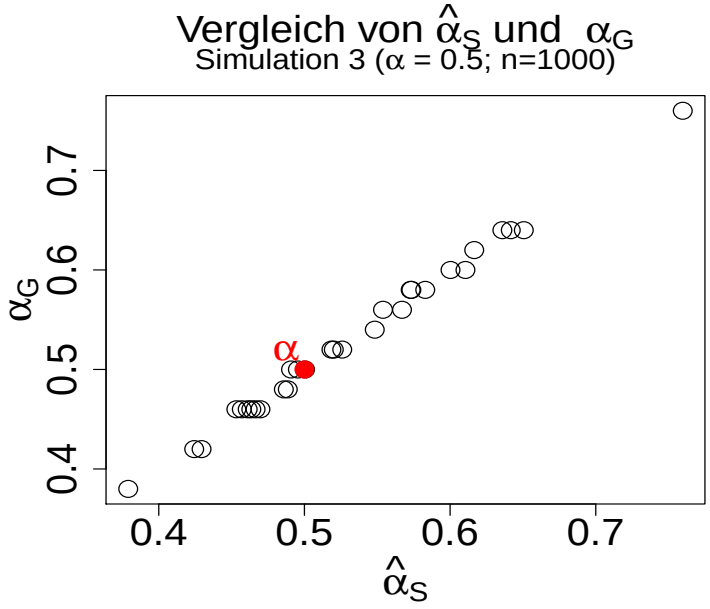

Abbildung 4.30: Szenario 3 ($\alpha = 0.5; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Für Szenario 3 ($\alpha = 1, n = 1000$) werden andere Werte für α vorgegeben, da aufgrund des wahren Wertes $\alpha = 1$ davon auszugehen ist, dass vermehrt Werte nahe des Wertes eins gewählt werden. Das Gitter ist $(0.3, 0.4, 0.5, 0.52, \ldots, 1.48, 1.5, 1.6, 1.7, \ldots, 2.3, 2.4, 3, 4, 5)$. Auch Abbildung 4.31 spricht für die Schätzung des Parameters α, da keine größeren Unterschiede bei den gewählten und geschätzten Werten f+r α zu verzeichnen sind.

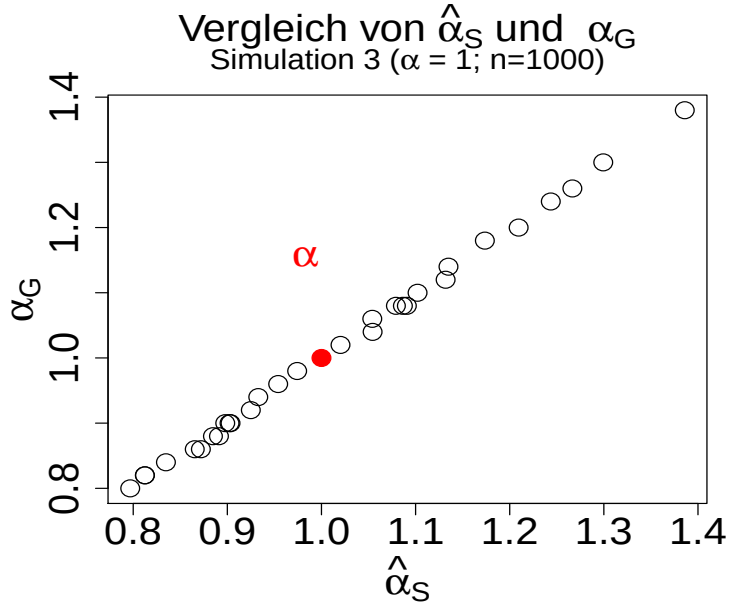

Abbildung 4.31: Szenario 3 ($\alpha = 1; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

4.4 Szenario 4

Es werden wieder zwei Beobachtungsanzahlen $n = 400$ und $n = 1000$ betrachtet. Die Daten werden gemäß dem BETAMIX(c)-Modell generiert. Auch der Parameter α der Betabinomialverteilung wird variiert. Es erfolgen die Generierung von Datensätzen mit den Werten $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$. Die Koeffizienten, die Verteilung und die Anzahl der Kovariablen, sowie die Anzahl der Responsekategorien können der folgenden Aufzählung entnommen werden:

- Kategorienanzahl $k = 7$
- Fünf Kovariablen für die Präferenzkomponente
  - X_1 und X_2: zwei binäre Kovariablen ($X_1, X_2 \sim B(1, 0.7)$)
  - drei auf dem Intervall $[-3, 3]$ stetig gleichverteilte Kovariablen X_3, X_4 und X_5
- Koeffizientenvektor $\gamma = (-4.5, -3.25, -2, 1.25, 2, 3.1, -1.9, -2.8, 1, 2.1, 1.6)$

- Fünf Kovariablen zur Parametrisierung der Mischwahrscheinlichkeit π_i:
  - Z_1,Z_2 und Z_3 sind standardnormalverteilt
  - Z_4 und Z_5: zwei binäre Kovariablen ($Z_4, Z_5 \sim B(1, 0.7)$)
- Koeffizientenvektor $\beta = (0.4, -1.5, -2.4, 1.4, 2.4, 0.8)$

Mit diesen Parametern werden für die unterschiedlichen Beobachtungsanzahlen und Werten von α jeweils 30 Datensätze generiert.
Die im Folgenden gezeigten Ergebnisse stammen aus der Anpassung eines BETAMIX(c)-Modells an die simulierten Datensätze. Dabei werden die Kovariablen $X_1, \ldots, X_5$ für die Präferenzkomponente und $Z_1, \ldots, Z_5$ für die Wahrscheinlichkeit π verwendet. Der Parameter α wird für alle Beobachtungen konstant geschätzt.
Die drei Abbildungen 4.32,4.33 und 4.34 zeigen, dass die Schätzung des Parameters α in den betrachteten Fällen gut funktioniert.

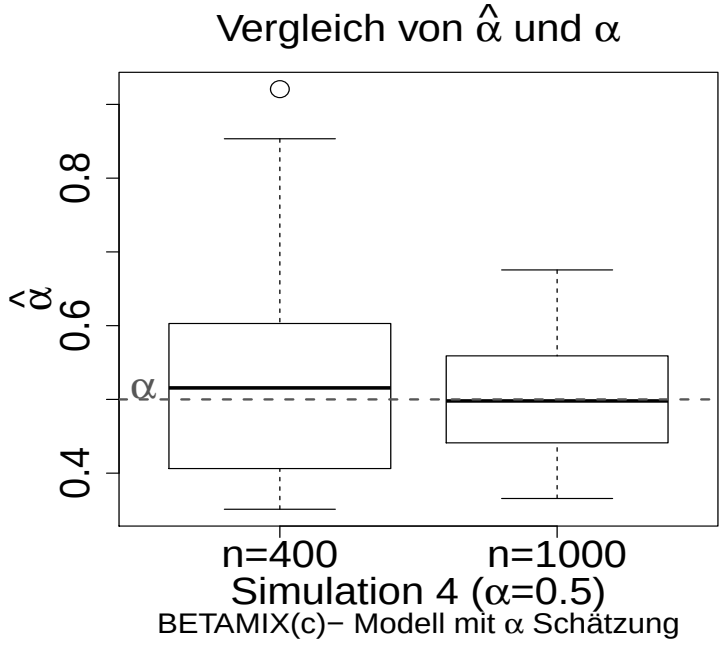

Abbildung 4.32: Szenario 4 ($\alpha = 0.5$): Boxplot des Parameters $\hat{\alpha}$

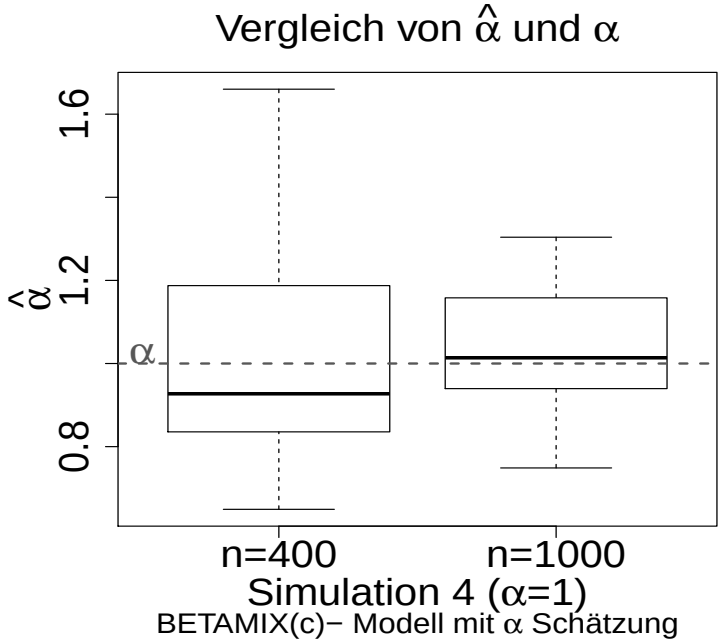

Abbildung 4.33: Szenario 4 ($\alpha = 1$): Boxplot des Parameters $\hat{\alpha}$

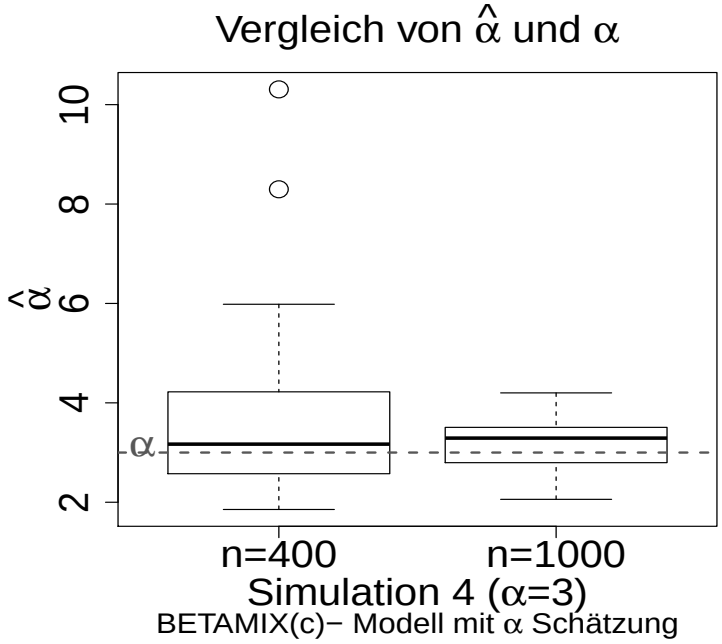

Abbildung 4.34: Szenario 4 ($\alpha = 3$): Boxplot des Parameters $\hat{\alpha}$

Die Mediane der kategoriespezifischen Intercepts $\gamma_{01}, \ldots, \gamma_{06}$ des Szenarios 4 ($\alpha = 1; n = 1000$) entsprechen fast den zur Datengenerierung verwendeten Werten. Nur bei γ_{05} und γ_{06} scheint der Wert des jeweiligen Medians etwas größer als der wahre Wert zu sein (vgl. Abbildungen 4.35 und 4.36). Die Schätzungen der Effekte für die Präferenz sind in Abbildung 4.37 dargestellt. Es ist zu erkennen, dass bei Koeffizient γ_2 die Abweichung des Medians zum wahren Wert etwas größer ist. Die Hälfte der $\hat{\gamma}_2$ ist kleiner als der Wert -2.93. Der wahre Wert für γ_2 liegt bei -2.8. Der Koeffizient γ_2 scheint somit etwas unterschätzt zu werden. Für die anderen Koeffizienten ist nur ein sehr kleiner Unterschied zwischen dem Wert des Medians und dem wahren Wert zu erkennen.

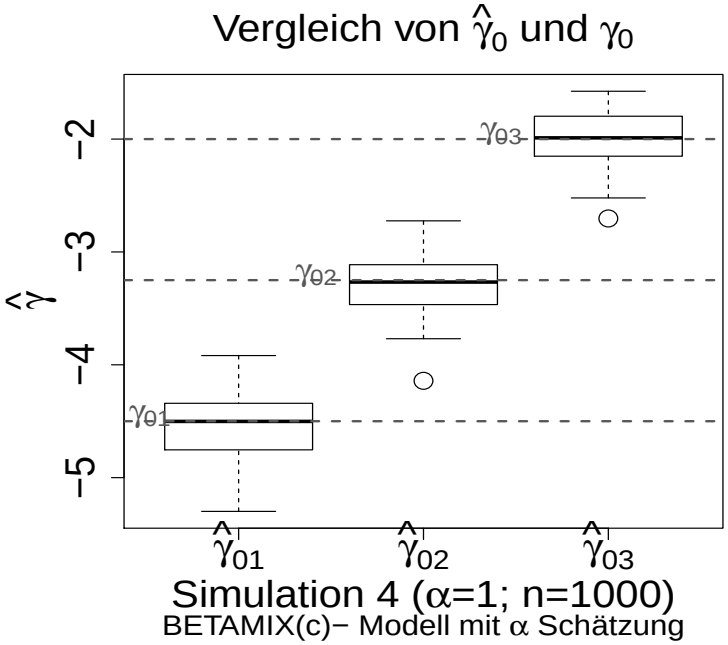

Abbildung 4.35: Szenario 4 ($\alpha = 1; n = 1000$): Boxplot der Intercepts γ_{01},γ_{02}, γ_{03}

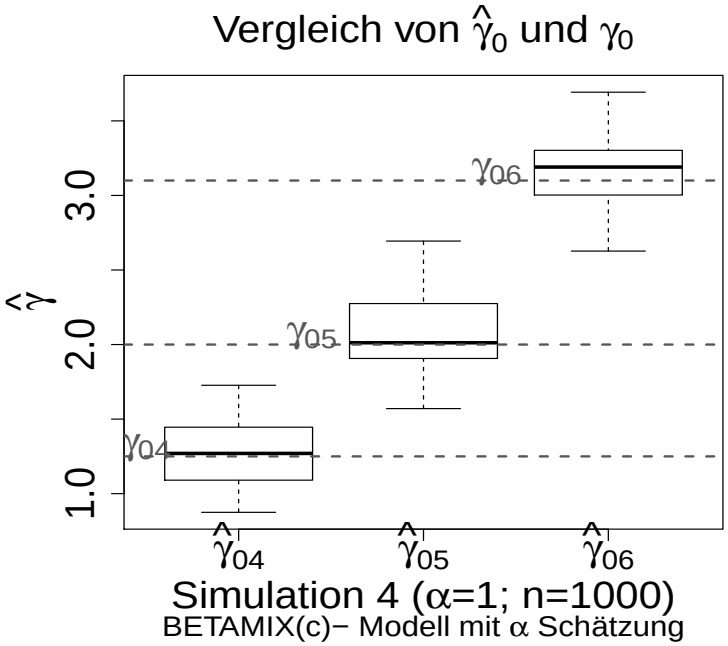

Abbildung 4.36: Szenario 4 ($\alpha = 1$; $n = 1000$): Boxplot der Intercepts γ_{04},γ_{05}, γ_{06}

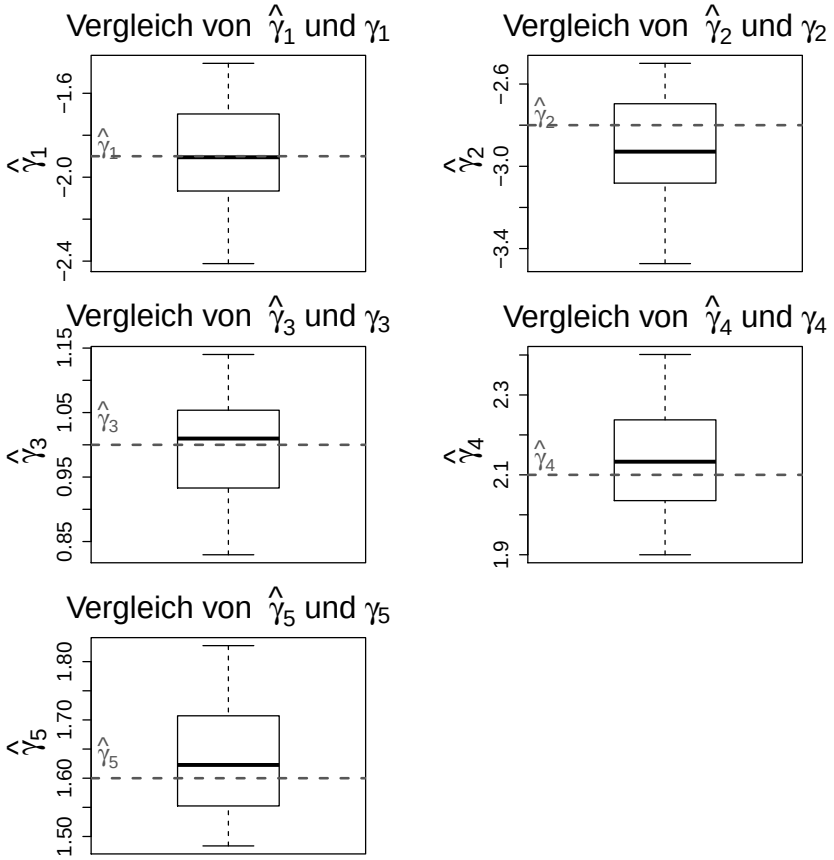

Abbildung 4.37: Szenario 4 ($\alpha = 1$; $n = 1000$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

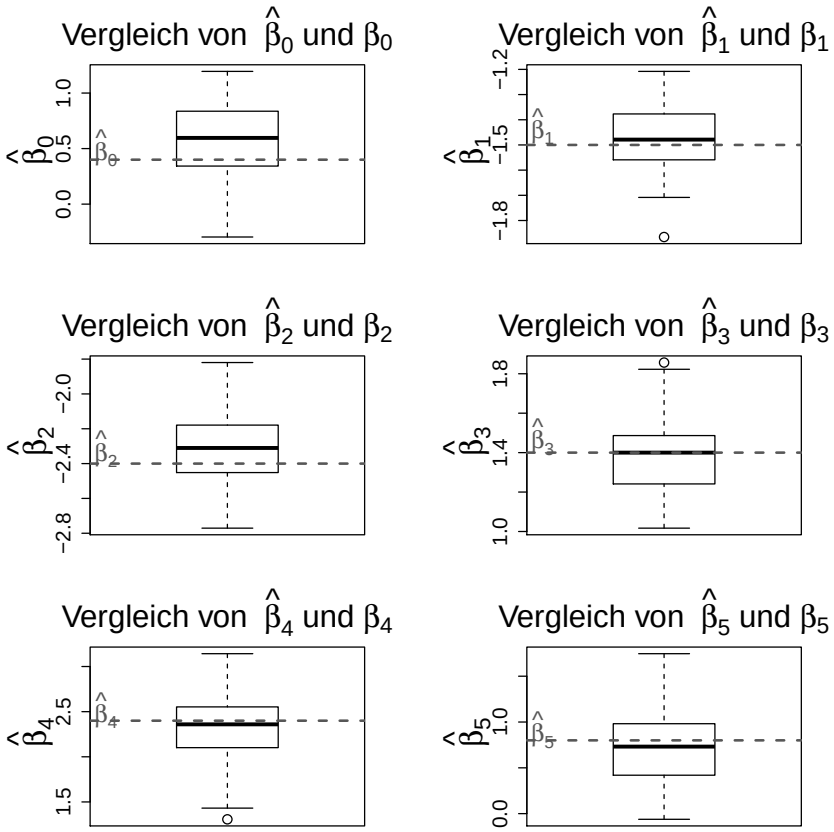

Abbildung 4.38: Szenario 4 ($\alpha = 1$; $n = 1000$): Boxplot der Koeffizienten $\beta_0, \ldots, \beta_5$

In Abbildung 4.38 sind die Schätzungen der Koeffizienten $\beta_0, \ldots, \beta_5$ als Boxplots dargestellt. Auch die Schätzung dieser Koeffizienten funktioniert mit dem BETAMIX(c)-Modell gut. Die geschätzten Werte streuen jeweils um den wahren Wert.
Die geschätzten Koeffizienten bei den Datensätzen mit $\alpha = 0.5$ oder $\alpha = 3$ weichen nicht stark von den zur Datengenerierung verwendeten Koeffizienten ab. Grafiken, die dies zeigen, sind im Anhang zu finden.
Die Wahl über ein Gitter führt auch in diesem Szenario zu ähnlichen Werten des Parameters α_G wie die Schätzung $\hat{\alpha}_S$ (vgl. Abbildung 4.39). Das in verwendete Gitter ist das Folgende: $(0.1, 0.12, \ldots, 0.98, 1, 1.1, 1.2, \ldots, 1.9, 2, 3)$.

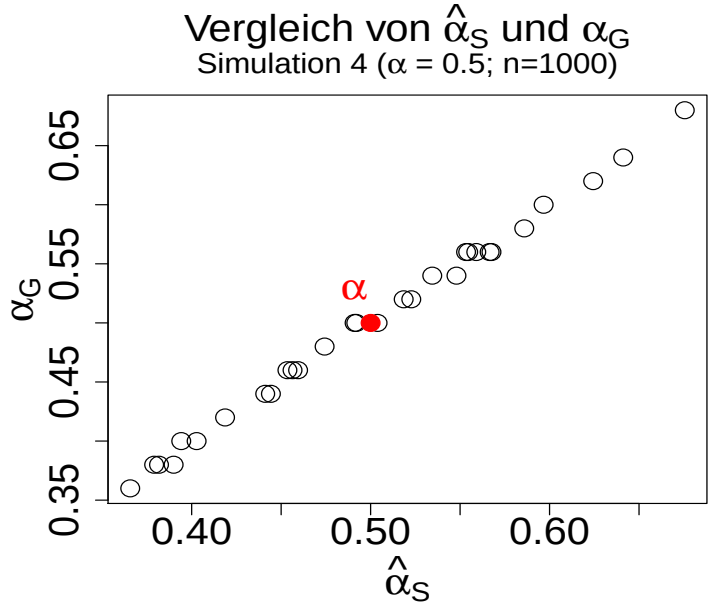

Abbildung 4.39: Szenario 4 ($\alpha = 0.5; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Auch bei den simulierten Datensätzen mit dem Parameter $\alpha = 1$ erhält man bei der Gitter basierten Wahl und der Schätzung kaum voneinander abweichende Werte für α. Das Gitter, welches für die Wahl der α's, die in Abbildung 4.40 dargestellt sind, verwendet wird, ist $(0.3, 0.4, 0.5, 0.52, \ldots, 1.48, 1.5, 1.6, 1.7, \ldots, 2.4, 3, 4, 5)$.

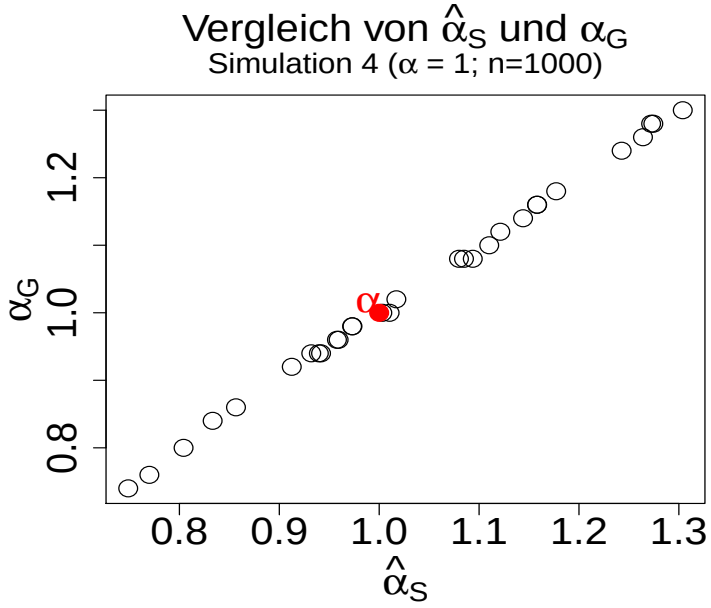

Abbildung 4.40: Szenario 4 ($\alpha = 1; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

4.5 Szenario 5

Die Parametrisierung des Parameters α durch Kovariablen ist eine weitere Möglichkeit des BETAMIX(c)-Modells. Szenario 5 simuliert solche Daten. Dabei wird jeweils eine stetige Variable für die Präferenzkomponente und für die Parametrisierung von α verwendet. Die Variable X, welche für die Präferenzkomponente genutzt wird, ist auf dem Intervall $[-3, 3]$ stetig gleichverteilt. Die Variable W ist standardnormalverteilt. Weitere Parameter, die für dieses Szenario gewählt werden, sind

- Kategorienanzahl $k = 7$
- $\pi = 0.7$
- Beobachtungsanzahl $n = 1000$
- $\boldsymbol{\gamma} = (-4.5, -3.25, -2, 1.25, 2, 3.1, -1.9)$
- $\boldsymbol{\alpha} = (1.4, -1.51)$

Es werden 30 Datensätze mit diesen wahren Werten generiert.

Die Schätzungen der Koeffizienten α_0 und α_1 scheinen den Abbildungen 4.41 und 4.42 nach zu urteilen leicht unterschätzt zu werden. Der betrachtete Wertebereich ist jedoch weder für α_0 noch für α_1 sehr groß. Weiterhin scheinen diese geschätzten Koeffizienten symmetrisch um den jeweiligen Median zu liegen.

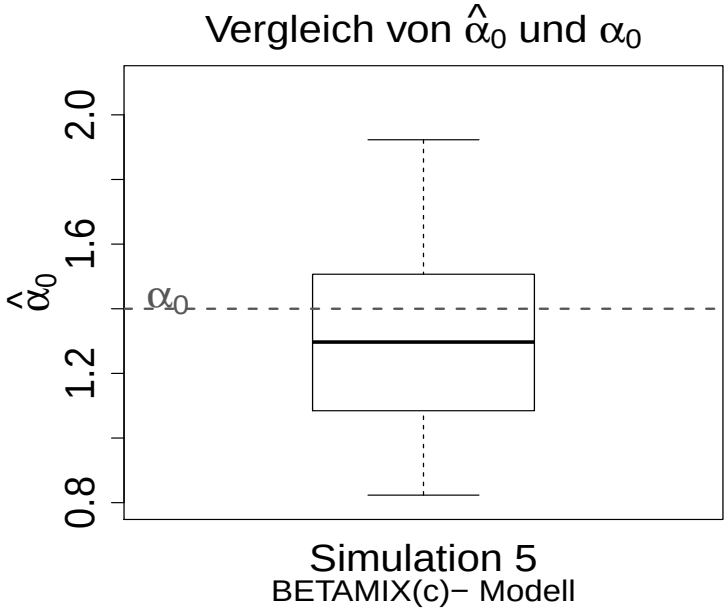

Abbildung 4.41: Szenario 5: Boxplot des Parameters $\hat{\alpha}_0$

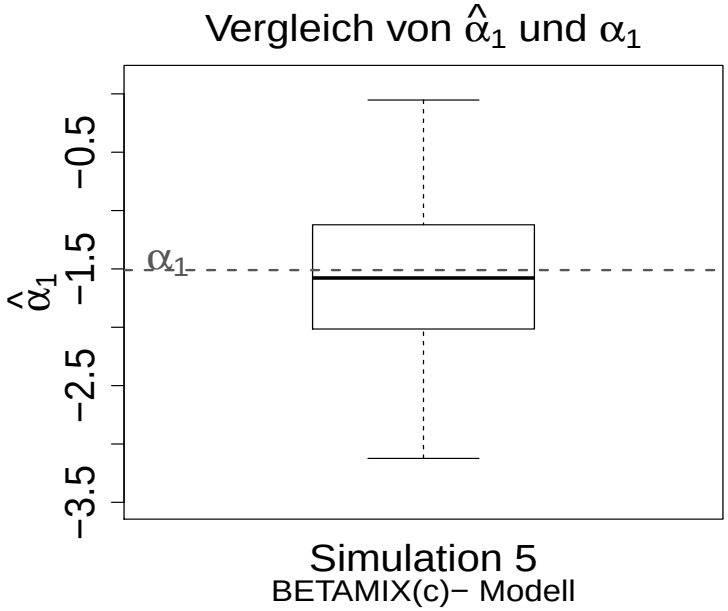

Abbildung 4.42: Szenario 5: Boxplot des Parameters $\hat{\alpha}_1$

Den Abbildungen 4.43 bis 4.45 ist zu entnehmen, dass auch die Schätzungen der Effekte γ aus der Präferenzkomponente gut zu sein scheinen. Sowohl die Mediane der einzelnen Intercepts als auch der Median des Effekts der Variable X_1 weichen nur geringfügig von den jeweiligen wahren Werten ab. Der Koeffizient $\hat{\gamma}_1$ scheint leicht unterschätzt zu werden.

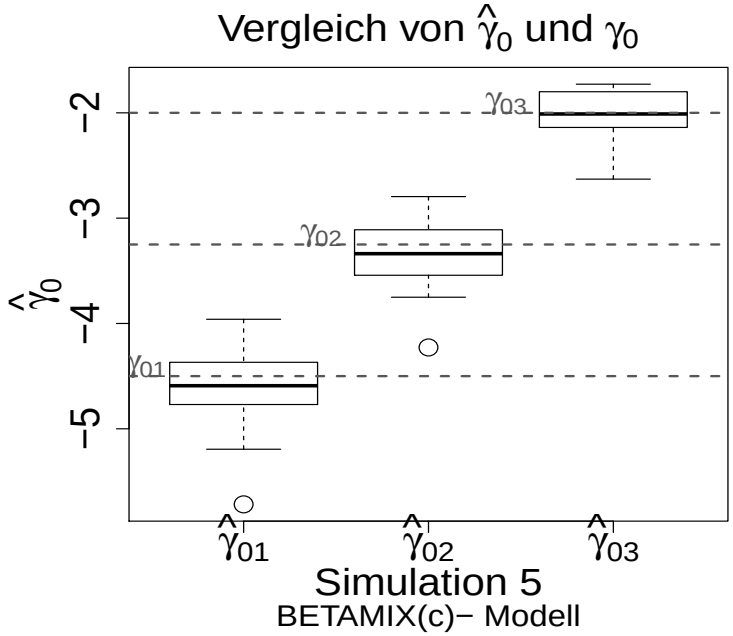

Abbildung 4.43: Szenario 5: Boxplot des Parameters $\hat{\gamma}_0$

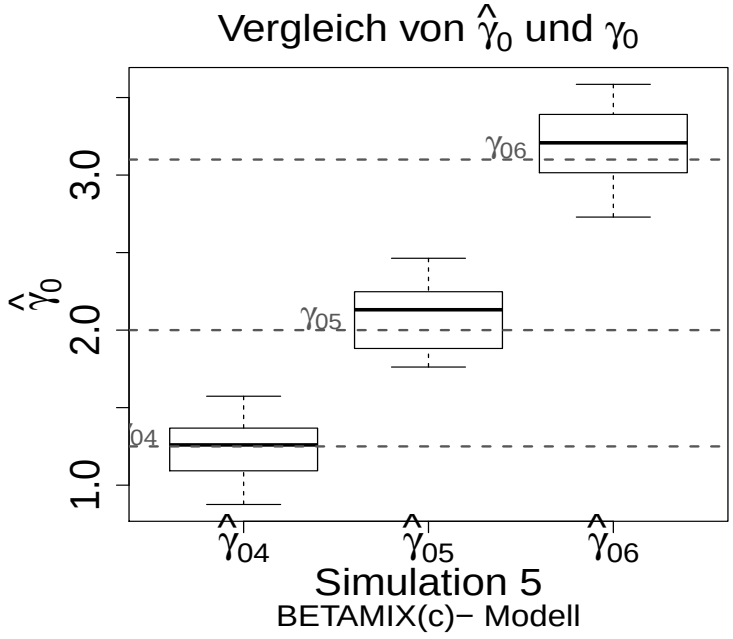

Abbildung 4.44: Szenario 5: Boxplot des Parameters $\hat{\gamma}_0$

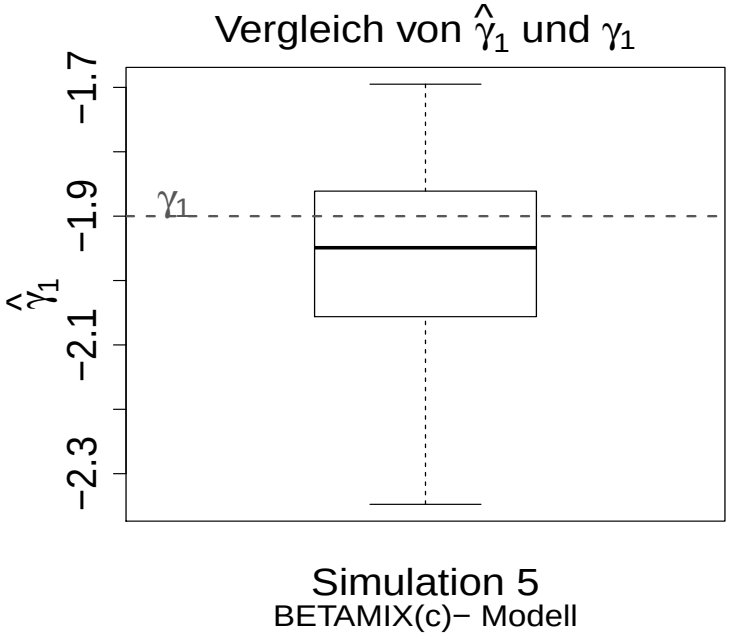

Abbildung 4.45: Szenario 5: Boxplot des Parameters $\hat{\gamma}_1$

Auch die Schätzungen der Mischwahrscheinlichkeit π weichen kaum vom wahren Wert 0.7 ab. Der kleinste Wert der 30 geschätzten $\hat{\pi}$ beträgt 0.65 und der größte Wert liegt bei 0.76 (vgl. Abbildung 4.46). Es treten bei den Schätzungen von π also kaum Ausreißer auf.

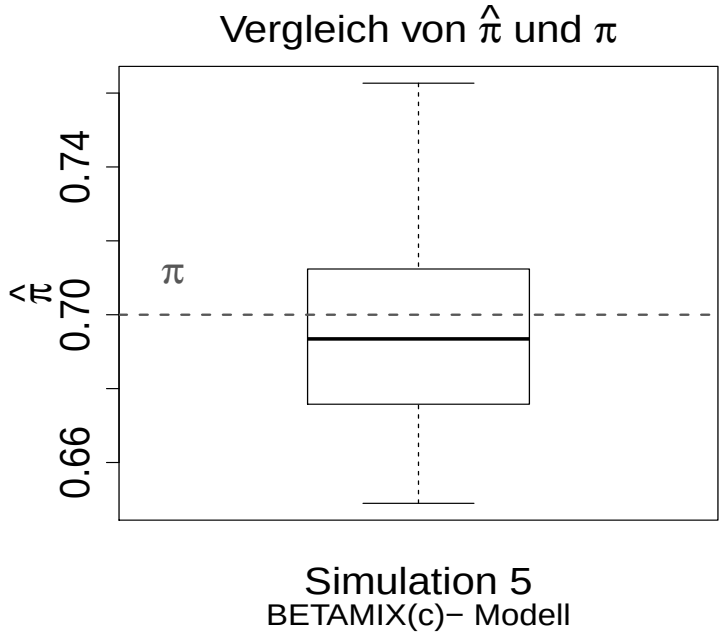

Abbildung 4.46: Szenario 5: Boxplot des Parameters $\hat{\pi}$

4.6 Szenario 6

Das datengenerierende Modell ist auch hier das BETAMIX(c)-Modell. In diesem Szenario werden mehrere Kovariablen für die Vektoren **x** und **w** verwendet. Für die Präferenzkomponente werden fünf stetig auf dem Intervall $[-3, 3]$ gleichverteilte Variablen verwendet. Diese fünf Variablen werden auch für den Parameter α aus der Betabinomialverteilung der Unsicherheitskomponente verwendet. Die wahren Koeffizienten $\boldsymbol{\gamma}$ und $\boldsymbol{\alpha}$ sind unter anderem der folgenden Aufzählung zu entnehmen:

- Kategorienanzahl $k = 7$
- $\pi = 0.7$
- Beobachtungsanzahl $n = 1000$
- $\boldsymbol{\gamma} = (-4.0, -3.25, -2, 1.25, 2, 3.1, -1.9, -2.5, 2, 3.2, 1.2)$
- $\boldsymbol{\alpha} = (1.4, -1.51, -2.25, -1, 2, 1.1)$

Die Anzahl der simulierten Datensätze beträgt für Szenario 6, wie in den zuvor betrachteten Szenarien, $S = 30$.

Die Schätzungen der einzelnen Einträge des Koeffizientevektors $\boldsymbol{\alpha}$ werden in Abbildung 4.47 dargestellt. Bis auf die Schätzungen für α_2 und α_4 scheinen diese symmetrisch um den jeweiligen Median verteilt zu sein. $\hat{\alpha}_2$ weist hier eine eher rechtssteile Verteilung, $\hat{\alpha}_4$ eine linkssteile Verteilung auf. Die Mediane der Schätzungen für $\alpha_0, \ldots, \alpha_5$ entsprechen fast den zur Datengenerierung verwendeten Werten. Der Responsestyle, welcher im Koeffizientenvektor $\boldsymbol{\alpha}$ beinhaltet ist, scheint somit gut geschätzt zu werden.

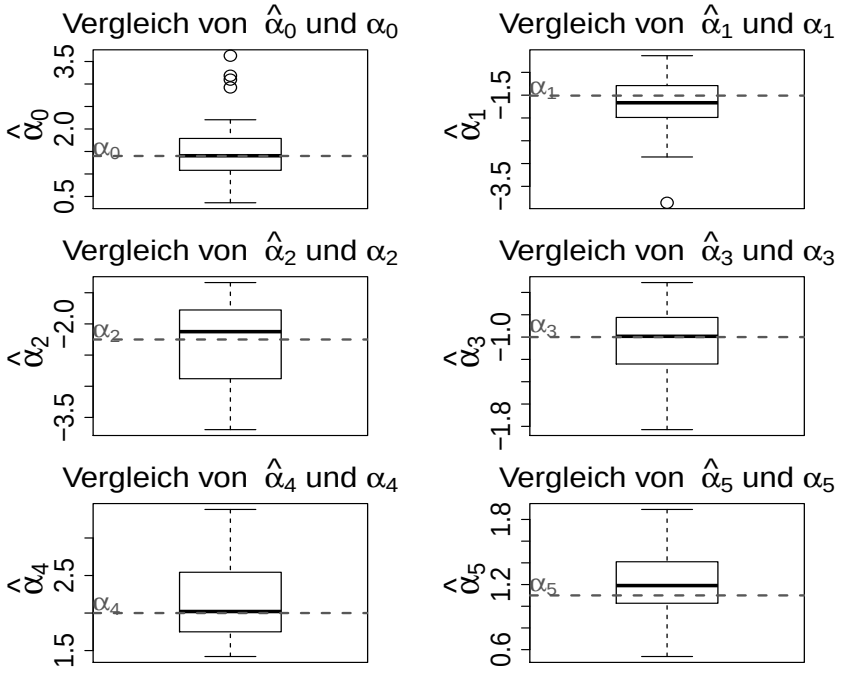

Abbildung 4.47: Szenario 6: Boxplot für die Werte des Vektors $\hat{\boldsymbol{\alpha}}$

Nicht nur die Responsestyle-Effekte sondern auch die Effekte der Präferenz werden mit dem BETAMIX(c)-Modell gut geschätzt. Die Mediane der Intercepts $\gamma_{01}, \ldots, \gamma_{06}$ weisen kaum Abweichungen zu dem jeweils wahren Wert auf. Auch die geschätzten Effekte $\gamma_1, \ldots, \gamma_5$ entsprechen fast den zur Datengenerierung genutzten Werten (vgl. Abbildungen 4.48, 4.49, 4.50 und 4.51).

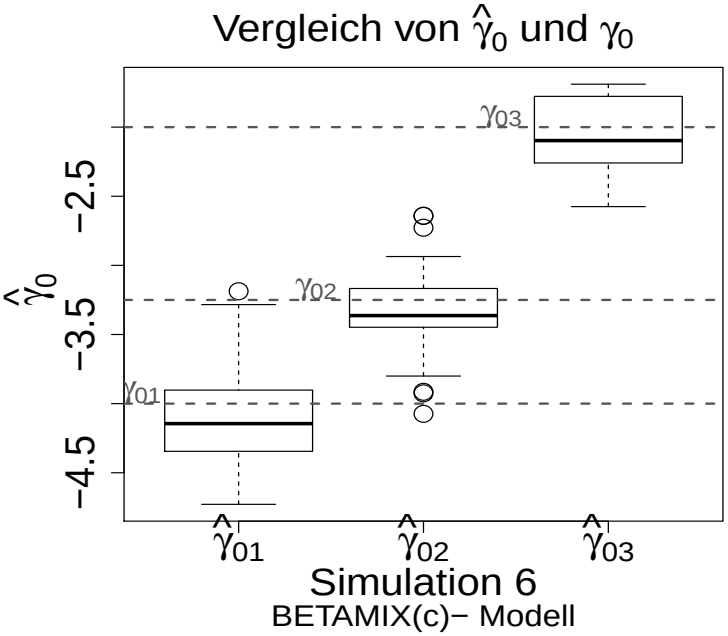

Abbildung 4.48: Szenario 6: Boxplot für die Intercepts $\hat{\gamma}_{01}, \ldots, \hat{\gamma}_{03}$

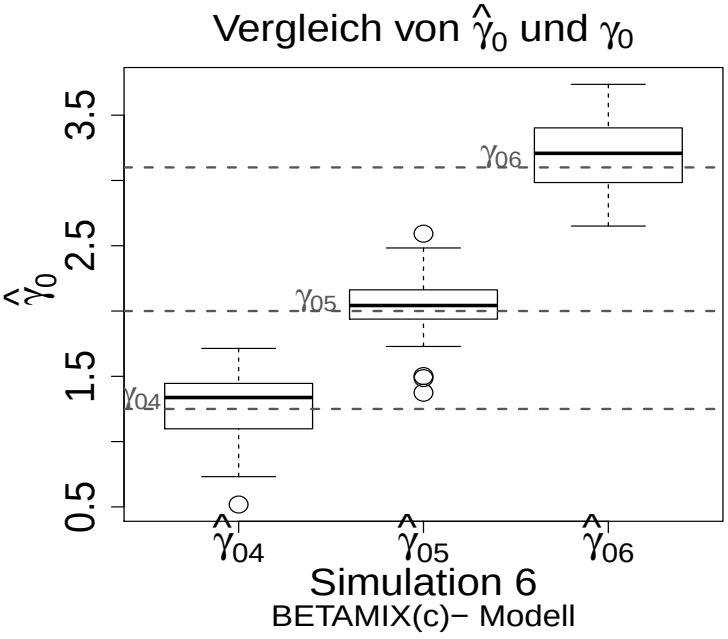

Abbildung 4.49: Szenario 6: Boxplot für die Intercepts $\hat{\gamma}_{04}, \ldots, \hat{\gamma}_{06}$

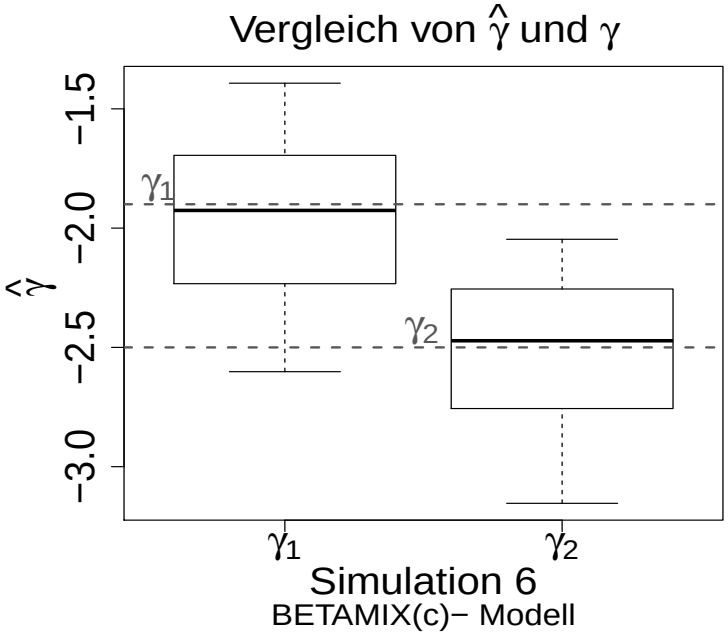

Abbildung 4.50: Szenario 6: Boxplot für die Koeffizienten $\hat{\gamma}_1, \hat{\gamma}_2$

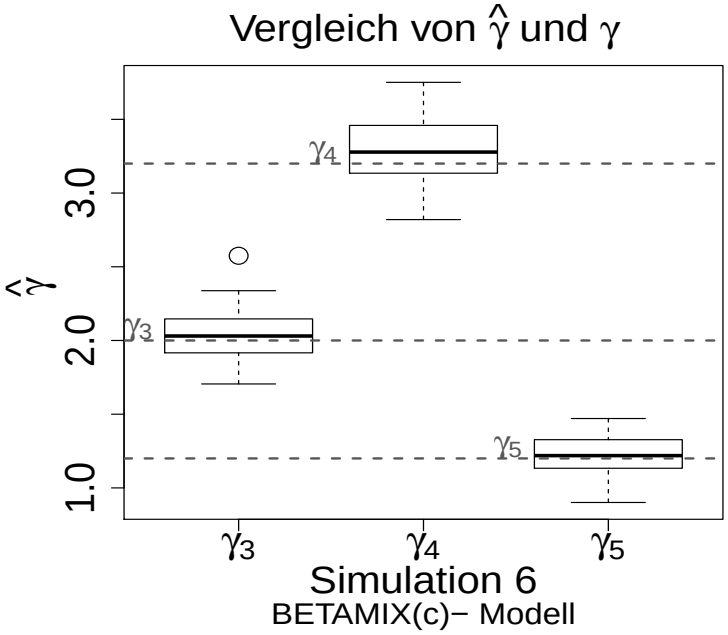

Abbildung 4.51: Szenario 6: Boxplot für die Koeffizienten $\hat{\gamma}_3, \hat{\gamma}_4, \hat{\gamma}_5$

Die Schätzungen der Mischwahrscheinlichkeit π sind in Abbildung 4.52 dargestellt. Wie auch in Szenario 5 und den Szenarien 1 und 2, in welchen die Wahrscheinlichkeit π nicht mit Kovariablen verknüpft wird, ist die Schätzung von π hier sehr gut.

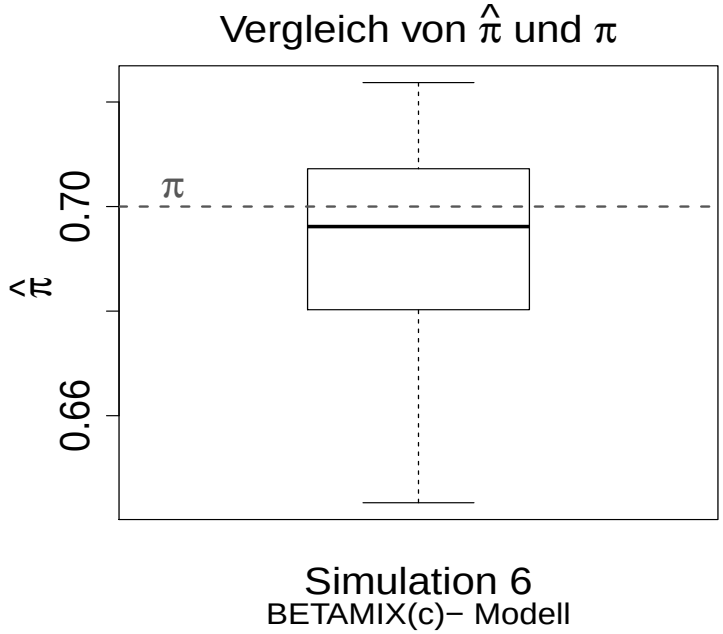

Abbildung 4.52: Szenario 6: Boxplot für die Wahrscheinlichkeit π

4.7 Ignorieren des Responsestyles

Das datengenerierenden Modell ist das BETAMIX(c)-Modell. Neben einer stetigen, auf dem Intervall $[-3;3]$ gleichverteilten, Kovariablen X_1 für die Präferenzkomponente wird eine standardnormalverteilte Variable W_1 für das Modell $\log(\alpha) = \mathbf{w}^T\boldsymbol{\alpha}$ verwendet. Der simulierte Response hat $k = 7$ Kategorien. Die Vektoren $\boldsymbol{\alpha}$ und $\boldsymbol{\gamma}$ werden wie folgt gewählt: $\boldsymbol{\alpha} = (1.4, 1.51)$, $\boldsymbol{\gamma} = (-4.5, -3.25, -2, 1.25, 2, 3.1, -1.9)$. Die Unsicherheitswahrscheinlichkeit $1 - \pi$ wird variiert. Es werden die Unsicherheitswahrscheinlichkeiten $0.1, \ldots, 0.6$ betrachtet und jeweils 30 Datensätze mit diesen Parametern simuliert.
Für jeden der simulierten Datensätze wird zum einen ein BETAMIX(c)-Modell mit der standardnormalverteilten Kovariablen W_1 für den Responsestyle, also zur Parametrisierung des α's, und der Variablen X_1 für die Präferenz angepasst. Dies entspricht der Anpassung der Daten mit dem wahren Modell. Außerdem wird auch das CUP(c)-Modell angepasst,

welches den mit $\log(\alpha) = \mathbf{w}^T\boldsymbol{\alpha}$ modellierten Responsestyle ignoriert.

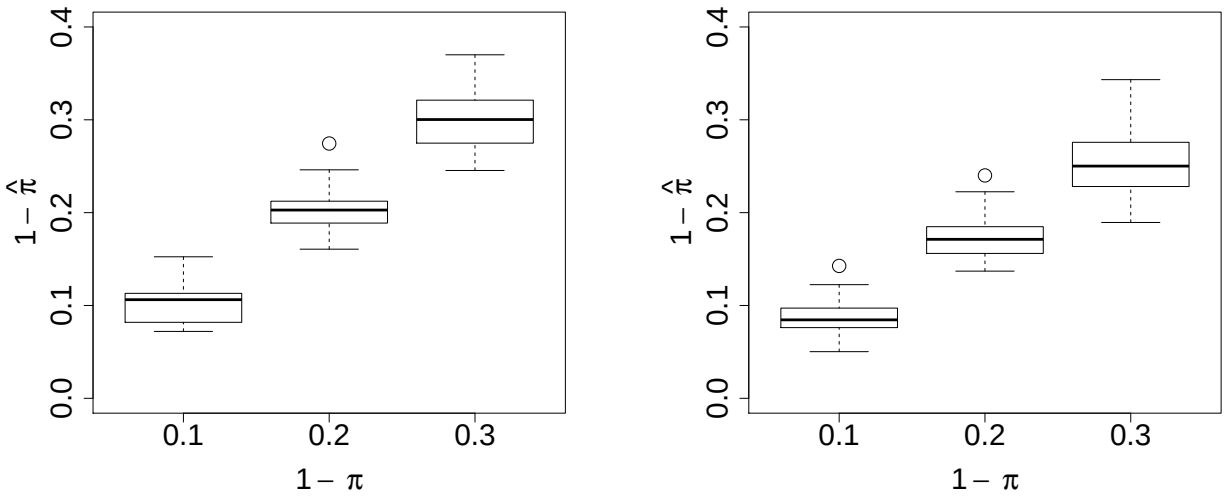

Abbildung 4.53: Simulation mit einem BETAMIX(c)-Modell mit Responsestyle als datengenerierenden Prozess mit den Werten $0.1, 0.2$ und 0.3 für die Unsicherheit $1 - \pi$: $1 - \hat{\pi}$ des BETAMIX(c) mit Responsestyle (links), $1 - \hat{\pi}$ des CUP(c)-Modells (rechts)

In Abbildung 4.53 und Abbildung 4.54 werden die Schätzungen von $1 - \hat{\pi}$, welche durch das BETAMIX(c)-Modell (links) und das CUP(c)-Modell (rechts) erhalten werden, dargestellt. In Abbildung 4.53 werden die Schätzungen, die für die Datensätze mit den wahren Unsicherheitswahrscheinlichkeiten $0.1, 0.2$ und 0.3 erhalten werden, betrachtet. Für Abbildung 4.54 erfolgt die Verwendung der Datensätze mit den wahren Unsicherheitswahrscheinlichkeiten $0.4, 0.5$ und 0.6.
Die Schätzungen $1 - \hat{\pi}$ des BETAMIX(c)-Modells (siehe linke Abbildungen) scheinen unverzerrt. Wird der Responsestyle ignoriert, so scheinen die Schätzungen der Unsicherheitswahrscheinlichkeit verzerrt zu sein. Alle sechs betrachteten Werte für $1 - \pi$ werden durch das CUP(c)-Modell unterschätzt. Außerdem scheint die Unterschätzung mit zunehmendem Wert für $1 - \pi$ zuzunehmen.

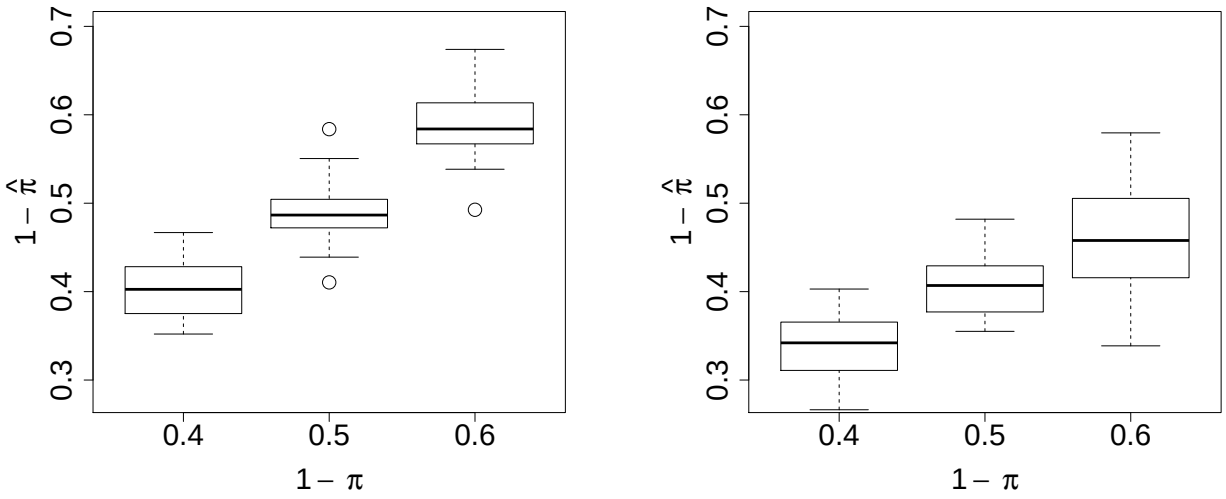

Abbildung 4.54: Simulation mit einem BETAMIX(c)-Modell mit Responsestyle als Daten generierenden Prozess mit den Werten $0.4, 0.5$ und 0.6 für die Unsicherheit $1 - \pi$: $1 - \hat{\pi}$ des BETAMIX(c) mit Responsestyle (links), $1 - \hat{\pi}$ des CUP(c)-Modells (rechts)

Die Schätzungen mittels BETAMIX(c)- und CUP(c)-Modell des Koeffizienten γ_1 sind in Abbildung 4.55 zu sehen. Die Schätzungen des BETAMIX(c)-Modells weisen nur geringe Abweichungen vom wahren Wert $\gamma_1 = -1.9$ auf. Die Schätzungen des CUP(c)-Modells scheinen jedoch mit zunehmender Unsicherheit zur 0 hin geschrumpft zu werden.

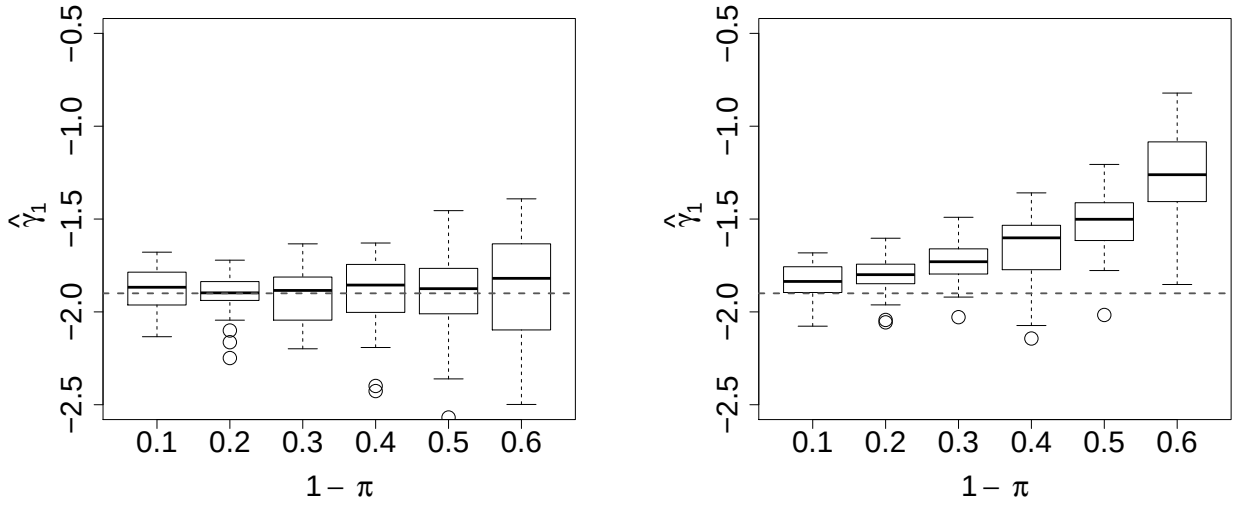

Abbildung 4.55: Simulation mit einem BETAMIX(c)-Modell mit Responsestyle als Daten generierenden Prozess mit variierenden Unsicherheitswahrscheinlichkeit $1 - \pi$ und $\gamma_1 = -1.9$: $\hat{\gamma}_1$ des BETAMIX(c) mit Responsestyle (links),$\hat{\gamma}_1$ des CUP(c)- Modells (rechts)

Für die Ergebnisse in Abbildung 4.55 wurde ein negativer Effekt zur Datengenerierung verwendet. Dass eine Schrumpfung in Richtung 0 auch bei einem positiven Effekt auftritt, wird im Folgenden gezeigt. Das datengenerierende Modell ist wie zuvor ein BETAMIX(c)-Modell mit Responsestyle. Bis auf den Vektor γ entsprechen die wahren Parameter den zu Beginn des Abschnitts 4.7 Beschriebenen. Der Koeffizientenvektor wird wie folgt gewählt: $\gamma = (-4.5, -3.25, -2, 1.25, 2, 3.1, 2)$. Der Effekt γ_1 ist hier somit positiv. Für die jeweils 30 simulierten Datensätze mit gleicher Unsicherheitswahrscheinlichkeit wird ein BETAMIX(c)-Modell mit den Variablen X_1 und W_1 und ein CUP(c)-Modell mit der Variablen X_1 angepasst. Die Schätzungen des Effekts γ_1 sind in Abbildung 4.56 dargestellt. Links sind die Schätzungen des BETAMIX(c)-Modells und rechts die des CUP(c)-Modells zu sehen.
Bei der Schätzung des Koeffizienten mit dem BETAMIX(c)-Modell ist eine Zunahme der Variabilität bei zunehmender Unsicherheit zu erkennen. Außerdem scheint auch eine leichte Verzerrung der Schätzungen bei den größeren Unsicherheitswahrscheinlichkeiten aufzutreten, welche jedoch nicht mit der bei Ignorieren des Responsestyles auftretenden Verzerrung zu vergleichen ist. Diese Verzerrung ist im rechten Teil der Abbildung 4.56 bei Zunahme der Unsicherheit und somit zunehmender Wichtigkeit des Responsestyles

deutlich zu erkennen.

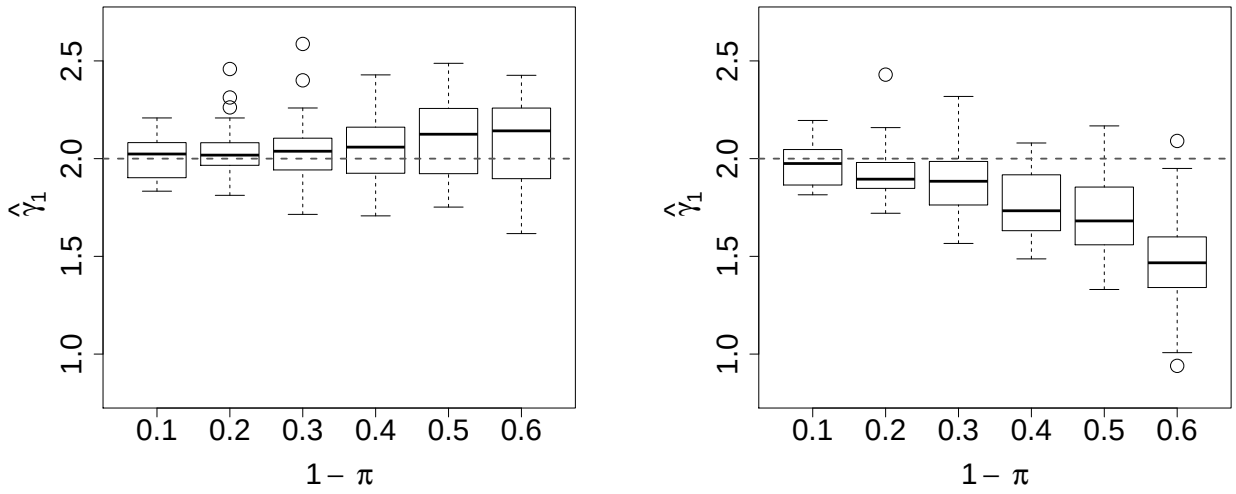

Abbildung 4.56: Simulation mit einem BETAMIX(c)-Modell mit Responsestyle als datengenerierenden Prozess mit variierenden Unsicherheitswahrscheinlichkeit $1 - \pi$ und $\gamma_1 = 2$: $\hat{\gamma}_1$ des BETAMIX(c) mit Responsestyle (links),$\hat{\gamma}_1$ des CUP(c)-Modells (rechts)

5 Anwendungsbeispiele

Im folgenden Kapitel werden die in Kapitel 2 vorgestellten Modelle auf unterschiedliche reale Datensätze angewandt. Es ist von Interesse, ob sich die Anpassung der Daten durch die komplexere Modellierung der BETAMIX-Modelle im Vergleich zu den CUP- und CUB-Modellen verbessert. Außerdem wird auch die Modellierung eines Responsestyles mit dem BETAMIX(c)-Modell betrachtet.

5.1 Datensatz *SHIW*

Der für die folgende Beispiele verwendete Datensatz basiert auf der SHIW (=Survey of Household Income and Wealth) Studie, welche im Jahre 2010 von der Bank of Italy durchgeführt wurde. Für die Analysen der ordinalen Zielgrößen „HAPPY“ wird ein Teildatensatz verwendet, welcher auch in Tutz et al. (2014) analysiert wurde.
Die für die Analyse verwendeten Variablen werden im Folgenden vorgestellt.

HAPPY:	Indikator für die Zufriedenheit (von 1: sehr unzufrieden bis 10: sehr zufrieden)
CIT:	Italienische Staatsangehörigkeit (1:ja, 2: nein)
STACIV:	Familienstand (1: verheiratet, 2: unverheiratet, 3: geschieden, 4: verwitwet)
ETA:	Alter (um 60 zentriert)
AREA3:	Wohnort (1: Nord, 2: Zentrum, 3: Süd)
FIDGEN:	Vertrauen in andere Personen (von 1 bis 10)
KLIMA:	Atmosphäre während des Interviews (von 1: niedrig bis 10: hoch)
CONDGEN:	Haushaltseinkommen reicht der Familie bis zum Ende des Monats (1: mit großen Schwierigkeiten bis 5: sehr einfach)
Y:	Familieneinkommen (um 33050 zentriert)

Das Balkendiagramm in Abbildung 5.1 zeigt, dass die extremen Kategorien (1, 2, 9, 10) im Vergleich zu den mittleren Kategorien eine geringere Häufigkeit ausweisen.

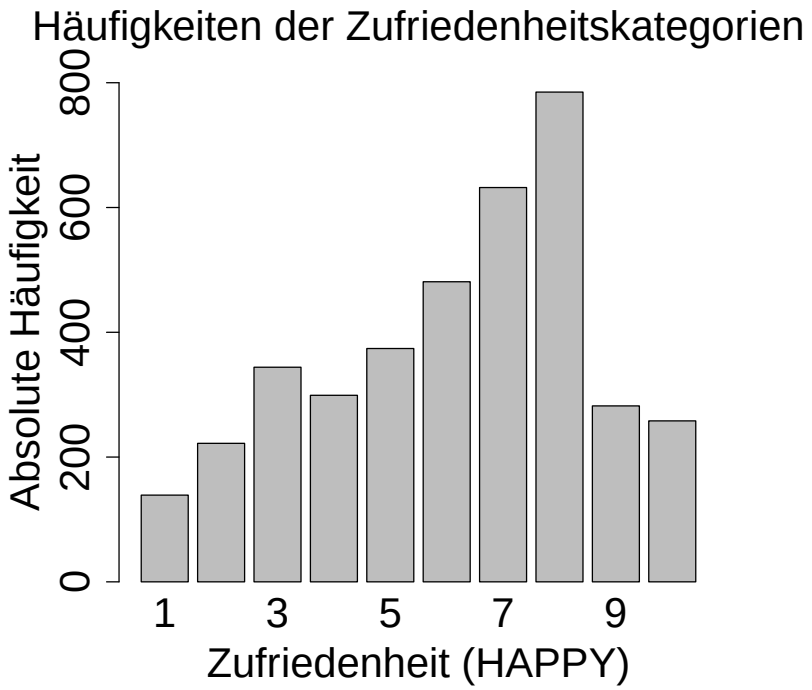

Abbildung 5.1: Balkendiagramm für die Responsevariable $HAPPY$

In den folgenden Analysen wird die Auswirkung des Parameters α der BETAMIX-Modelle auf die Datenanpassung betrachtet. Dabei wird die Anpassung der BETAMIX-Modelle mit der Datenanpassung weiterer Modelle verglichen. Außerdem werden die Unsicherheitswahrscheinlichkeiten $(1 - \pi)$ und die geschätzten Koeffizienten betrachtet.

Wahl des Parameters α

Zunächst wird betrachtet, ob laut Anpassungskriterium eine komplexere Modellierung mittels BETAMIX-Modell eine bessere Anpassung erzielt oder ob die weniger komplexen Modelle CUB, CUP oder Modelle ohne Mischungsansatz, wie z.B. das kumulative Logit-Modell, zu bevorzugen sind. Dazu wird die Anpassung der in Kapitel 2 beschriebenen Modelle mit der Responsevariablen „$HAPPY$“ und unterschiedlichen Kovariablen betrachtet.
Die Simulationsergebnisse sprechen für eine Schätzung des Parameters α statt des Abtastens eines vorgegebenen Gitters, da man mit beiden Möglichkeiten ähnliche Ergebnisse erhält, bei der Schätzung jedoch nur ein Modell anpassen muss. Die betrachten Beispiele

werfen nochmals einen Blick auf das Abtasten eines Gitters um den aus den Simulationen gewonnenen Eindruck zu verstärken.
Das für die folgenden BETAMIX(c)-Modelle verwendete Gitter des Parameters α beinhaltet folgende Werte: $0.1, 0.2, \ldots, 1.9, 2.0, 2.5, 3.0, \ldots, 9.5, 10, 15, 20$. Es erfolgt somit die Anpassung von jeweils 36 BETAMIX(b)- und BETAMIX(c)-Modellen. Die zugehörigen AIC- und BIC-Werte werden in den folgenden Abbildungen mit durchgezogen Linien dargestellt. Die gestrichelten Linien stellen hingegen den AIC- oder BIC-Wert eines weniger flexiblen Modells (CUB-, CUP-, Adjacent Categories- oder kumulatives Logit-Modell) dar. Da diese Modelle nicht vom Parameter α abhängig sind, werden die Werte als konstant für alle festen α's dargestellt.
Außerdem erfolgt neben der Anpassung der Modelle mit vorgegebenen α auch eine Schätzung des Parameters α innerhalb des BETAMIX(c)-Modells. Der AIC-Wert dieses Modells und der geschätzte Wert $\hat{\alpha} = \exp(\hat{\alpha}_0)$ wird ebenfalls in den Grafiken dargestellt.
Zunächst wird die Anpassung der Modelle mit der Kovariable Alter („*ETA*“) betrachtet. Für die Schätzung der Wahrscheinlichkeit π werden keine Kovariablen verwendet.

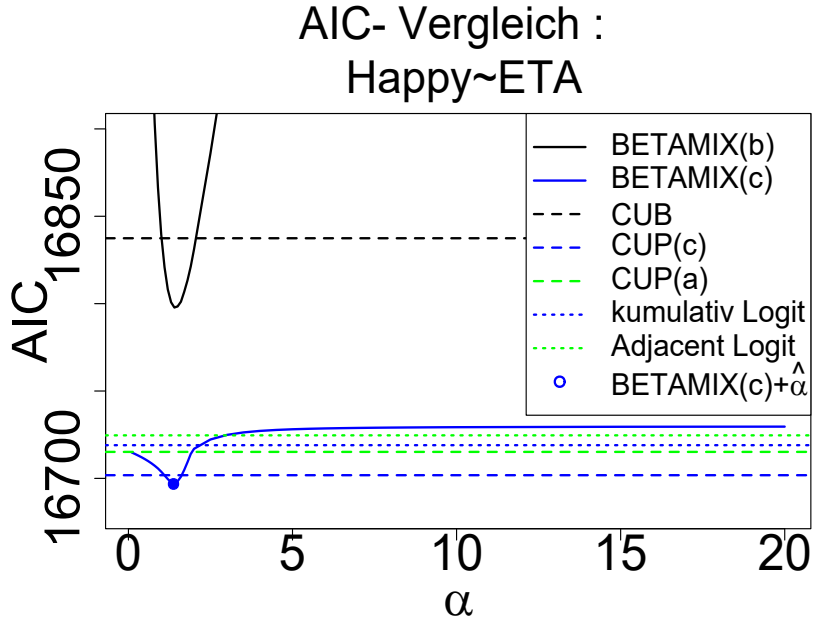

Abbildung 5.2: Vergleich der AIC- Werte für Modelle mit „ETA“ als Kovariable

Die flexibleren Modelle BETAMIX(b) und BETAMIX(c) scheinen, verglichen mit den

Mischmodellen CUB und CUP(c), für bestimmte Werte von α bessere Anpassungen zu erzielen. Der kleinste AIC-Wert des BETAMIX(b)-Modells liegt bei $\alpha = 1.4$. Auch das BETAMIX(c)-Modell weist ein Minimum für das AIC Kriterium bei $\alpha = 1.4$ auf. Laut AIC ist die Datenanpassung bei Verwendung des BETAMIX(c) mit dem Parameter $\alpha = 1.4$ verglichen mit den anderen betrachteten Modellen am besten (siehe Abbildung 5.2).
Das BETAMIX(c)-Modell, welches den Parameter α durch eine Konstante schätzt, liefert $\hat{\alpha} = \exp(\hat{\alpha}_0) = 1.373$ und einen AIC-Wert von 16696.8. Wie auch in den Simulationen zu beobachten ist, unterscheiden sich das geschätzte (1.373) und das abgetastete (1.4) α kaum. Das geschätzte α und der zugehörige AIC-Wert wird in Abbildung 5.2 mit einem blauen Punkt dargestellt. Der Abbildung ist weiterhin zu entnehmen, dass das BETAMIX(c)-Modell, welches für die Parametrisierung des Parameters α ein Intercept-Modell verwendet, die Daten am besten anzupassen scheint.

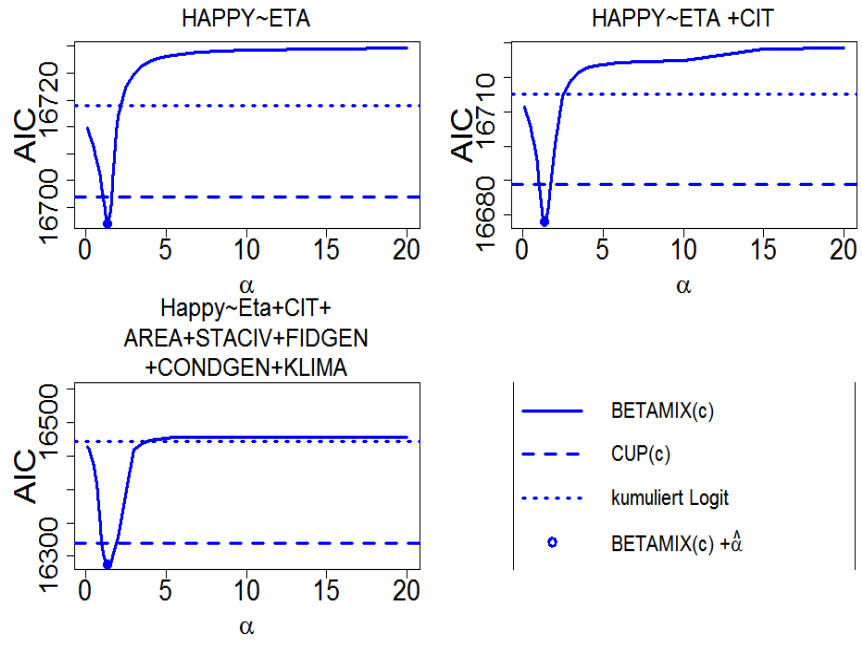

Abbildung 5.3: SHIW: AIC- Vergleich für kumulierte Modelle ohne Parametrisierung von π

In Abbildung 5.3 sind neben dem AIC-Wert des kumulativen Logit-Modells und des CUP(c)-Modells auch die AIC-Werte der BETAMIX(c)-Modelle mit festen α's (durchgezogene Linie) und geschätztem (Punkt) α zu sehen. Die jeweils verwendete Kovariablen sind dem Titel zu entnehmen.

Für die Werte oben links wird beispielsweise das Alter („ETA“) als Kovariable verwendet. Die in Abbildung 5.3 betrachteten Modelle verwenden zur Schätzung der Wahrscheinlichkeit π keine Kovariablen. Für alle drei Variablenkombinationen scheint es ein lokales Minimum der AIC-Werte der BETAMIX(c)-Modelle mit festen α zu geben, welches in allen drei Fällen bei $\alpha = 1.4$ liegt.
Das geschätzte $\hat{\alpha} = \exp(\hat{\alpha}_0)$ des BETAMIX(c)-Modells und der zugehörige AIC-Wert wird in den Abbildungen mit einem Punkt dargestellt. Der Abbildung ist zu entnehmen, dass der geschätzte Wert in allen drei Fällen ungefähr dem α-Wert entspricht, welches das AIC-Kriterium minimiert.
Die flexiblere Modellierung des BETAMIX(c)-Modells scheint für die Modelle „$HAPPY \sim ETA$“, „$HAPPY \sim ETA + CIT$“ und
„$HAPPY \sim ETA + CIT + AREA + STACIV + FIDGEN + CONDGEN + KLIMA$“ angebracht zu sein, da die Anpassung des komplexeren BETAMIX(c)-Modells gegenüber dem CUP(c)- und kumulativen Logit-Modell besser ist. Das BETAMIX(c)-Modell, welches den Parameter α schätzt, passt die Daten laut AIC-Kriterium jedoch am besten an.
Weiterhin lässt sich erkennen, dass das BETAMIX(c)-Modell mit den sieben Kovariablen und dem geschätzten α gegenüber den Modellen mit weniger Kovariablen zu bevorzugen ist.

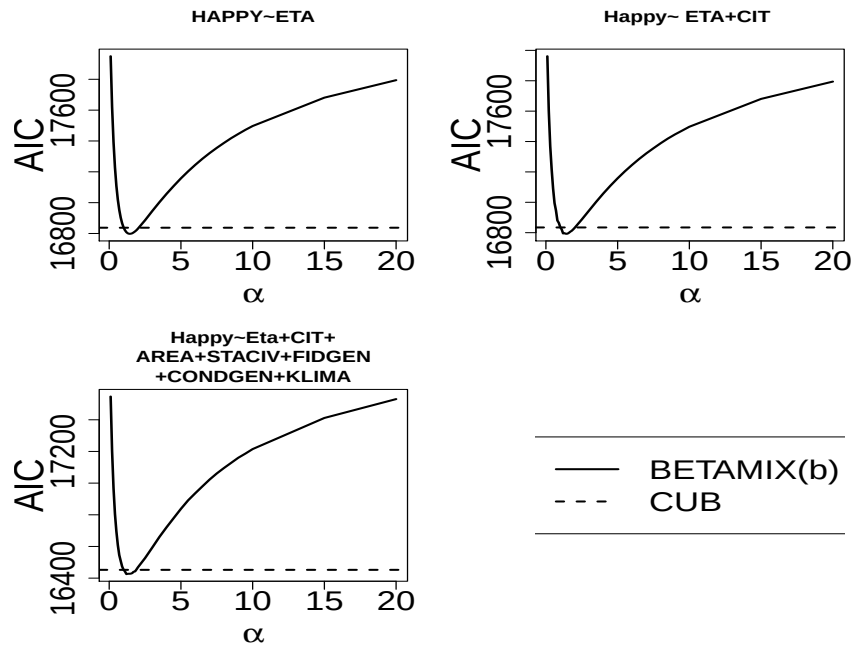

Abbildung 5.4: SHIW: Vergleich der AIC-Werte für CUB- und BETAMIX(b)-Modelle, die keine Kovariablen zur Parametrisierung des Parameters π verwenden

Die Verwendung des CUB- und des BETAMIX(b)-Modells für die SHIW Daten führt verglichen mit den kumulativen Modellen (CUP(c)-, BETAMIX(c)- und kumulatives Logit-Modell), welche in Abbildung 5.3 zu sehen sind, zu höheren AIC-Werten (vgl. Abbildung 5.4). Dies wurde auch in Tutz et al. (2014) gezeigt. Für alle Kovariablenkombinationen erkennt man, dass die komplexere Modellierung mit dem BETAMIX(b)-Modell die Datenanpassung verglichen mit dem CUB-Modell verbessern kann. Auch bei den BETAMIX(b)-Modellen minimiert der Wert $\alpha = 1.4$ bei den betrachteten Variablenkombinationen das AIC-Kriterium.
Wie auch beim AIC gibt es für das BIC des BETAMIX(c)-Modells ein eindeutiges Minimum, welches bei $\alpha = 1.4$ liegt (vgl. Abbildung 5.5). Im Gegensatz zum AIC-Kriterium scheint dem BIC-Kriterium nach zu urteilen für das Modell „$HAPPY \sim ETA$"das CUP(c)-Modell eine bessere Datenanpassung zu erzielen als eines der flexibleren BETAMIX-Modelle.

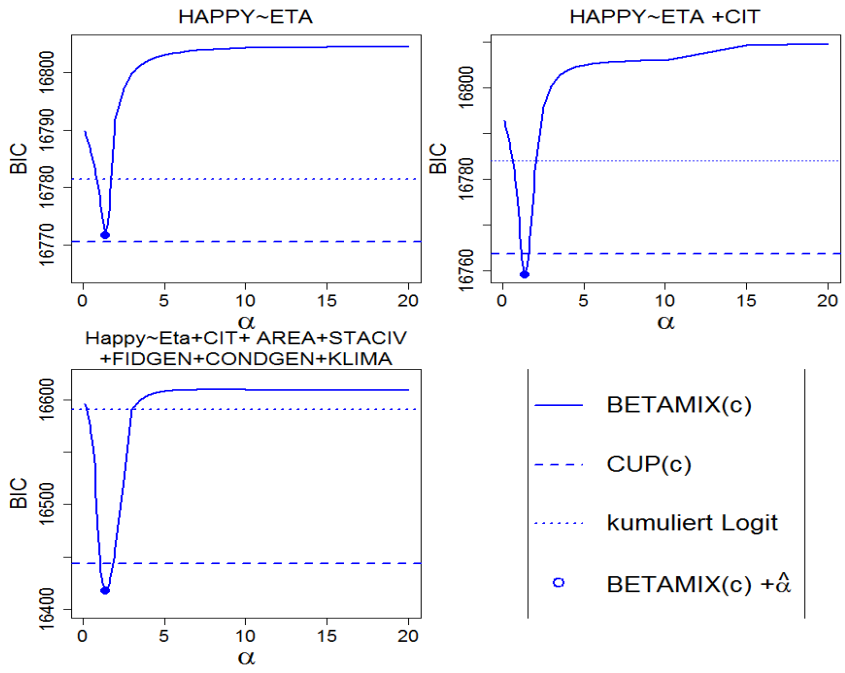

Abbildung 5.5: SHIW: Vergleich der BIC-Werte für CUP(c)- und BETAMIX(c)-Modelle, die keine Kovariablen zur Parametrisierung des Parameters π verwenden

Die im Folgenden betrachteten Modelle verwenden einen Kovariablenvektor **z** zur Parametrisierung der Wahrscheinlichkeit π_i. In Abbildung 5.6 werden die AIC-Werte für BETAMIX(c)-, CUP(c)- und kumulative Logit-Modelle dargestellt. Die BETAMIX(c)- und CUP(c)-Modelle verwenden Kovariablen für die Spezifizierung des kumulativen Logit-Modells in der Präferenzkomponente und der Wahrscheinlichkeit π. Das einfache kumulative Modell, dessen AIC-Wert ebenfalls aus der Abbildung 5.6 zu entnehmen ist, enthält keine Mischkomponente. Daher werden auch nur die Variablen für das kumulative Modell verwendet, die im BETAMIX(c)- und CUP(c)-Modell für die Präferenzkomponente gewählt werden.
Für die Abbildung oben links in Abbildung 5.6 besteht der Kovariablenvektor **x** aus der Variablen Alter („ETA“). Für die Parametrisierung der Wahrscheinlichkeit π wird ebenfalls das Alter (“ETA„) verwendet.
Die Modelle oben rechts und unten links beinhalten sieben **x**-Variablen. Die **x**-Variablen sind „ETA“, „CIT“, „AREA“, „STACIV“, „FIDGEN“, „CONDGEN“und „KLIMA“. Bei Berücksichtigung der Dummycodierung der Variablen „AREA“und „STACIV“erhält man zehn Variablen. Dies ist dem Titel der Grafik zu entnehmen. Die BETAMIX(c)- und CUP(c)-

Modelle oben rechts verwenden neben den **x**- Variablen vier Variablen für den Vektor **z**. Zu diesen Variablen gehören „STACIV“, „AREA“, „FIDGEN“und„KLIMA“. Für die AIC-Werte links unten wurde $\mathbf{x} = \mathbf{z}$ gewählt.
Für die in Abbildung 5.6 betrachteten Variablenkombinationen lässt sich erkennen, dass die AIC-Werte bis zu $\alpha = 1.2$ oder $\alpha = 1.3$ abnehmen. Bis zu einem Wert von $\alpha = 2$ nehmen die AIC-Werte wieder zu. Für Werte $\alpha \leq 2$ wiederum ist eine Abnahme zu erkennen.
Im Gegensatz zu den BETAMIX(c)-Modellen, die keine Kovariablen für π verwenden, scheinen die lokalen Minima der über das Gitter erhaltenen AIC-Werte nicht auch globale Minima zu sein. Bei den drei in Abbildung 5.6 betrachteten Fällen scheint ein größeres α eine bessere Anpassung an die Daten zu erzielen. Die Schätzung des Parameters α hingegen beträgt für das Modell oben links 1.19. In den zwei weiteren Modelle wird für α der Wert $\exp(\hat{\alpha}_0) = 1.23$ geschätzt. Man sieht, dass diese Schätzungen in etwa den α-Werten der lokalen Minima entsprechen.
Die Simulationsszenarien 3 und 4 verwenden Variablen für die Präferenzkomponente und für die Wahrscheinlichkeit π. Die Werte der gewählten α's unterscheiden sich in diesen Simulationsszenarien kaum von den geschätzten Werten für $\exp(\hat{\alpha}_0)$. Für die in Abbildung 5.6 betrachteten BETAMIX-Modelle liegt jedoch ein Unterschied zwischen gewählten und geschätzten Werten für α vor. Da bei der Schätzung ein lokales Maximum der Likelihood gefunden wird, kann eine Änderung der Startwerte eventuell Abhilfe leisten. Die Änderung der Werte für die bisherigen Startwert-Argumente in der Funktion zur Anpassung eines BETAMIX(c)-Modells scheint für die in Abbildung 5.6 betrachteten Beispiele jedoch keinen anderen Schätzwert für $\exp(\hat{\alpha}_0)$ zu erzielen. Aus diesem Grund kann es bei BETAMIX(c)-Modellen mit Kovariablen für die Wahrscheinlichkeit π sinnvoll sein einen Blick auf die Wahl des α's über ein Gitter zu werfen.

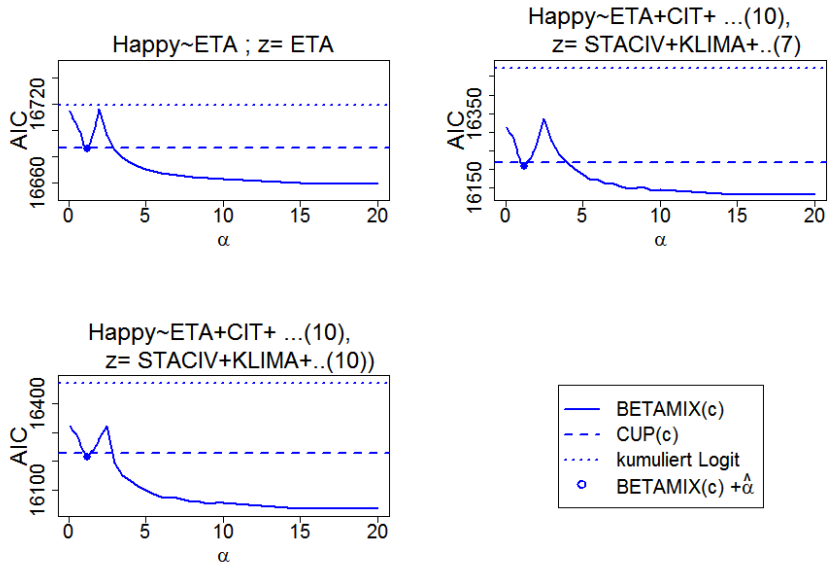

Abbildung 5.6: SHIW: Vergleich der AIC-Werte für Modelle mit einer Parametrisierung der Wahrscheinlichkeit π durch Kovariablen

Verwendet man keine Kovariablen für die Schätzung der Wahrscheinlichkeit π, so erkennt man in Abbildung 5.3, dass das Modell unten links die Daten am besten anpasst. Der Variablenvektor $\mathbf{x}$ aus diesem Modell wird auch für das obere rechte und untere Modell in Abbildung 5.6 verwendet. Wird zu diesem Variablenvektor noch ein weiterer Vektor $\mathbf{z}$ für die Wahrscheinlichkeiten π_i hinzugenommen, so scheint eine Verbesserung der Datenanpassung zu erfolgen, was bei Betrachtung der AIC-Werte aus Abbildung 5.6 deutlich wird. Den kleinsten AIC-Wert kann man beim BETAMIX(c)-Modell mit festem $\alpha = 20$ und den Kovariablenvektoren $\mathbf{x}_i = \mathbf{z}_i = (ETA_i, CIT_i, STACIV_i, AREA_i$ $FIDGEN_i, CONDGEN_i, KLIMA_i)$ beobachten.
Im Folgenden werden jedoch die Schätzungen der Koeffizienten γ und $\boldsymbol{\beta}$ für das das BETAMIX(c)- Modell mit den Kovariablen
$\mathbf{x}_i = (ETA_i, CIT_i, STACIV_i, AREA_i, FIDGEN_i, CONDGEN_i, KLIMA_i)$ und
$\mathbf{z}_i = (STACIV_i, AREA_i, FIDGEN_i, KLIMA_i)$, sowie der Schätzung des Parameters α betrachtet. Der Grund hierfür ist, dass dieses Modell den kleinsten AIC-Wert unter den betrachteten BETAMIX(c)-Modellen, welche das α schätzen, hat. Der AIC-Wert für dieses Modell beträgt 16214.89. Die Parameterschätzungen sind in Tabelle 5.1 dargestellt.

	Kovariablen	Koeffizienten
$\boldsymbol{\beta}$	Konstante (β_0)	0.04
	Familienstand: Unverheiratet	0.32
	Familienstand: Geschieden	0.69
	Familienstand: Verwitwet	0.53
	Wohnort: Zentrum	0.93
	Wohnort: Süd	0.31
	Vertrauen	0.08
	Atmosphäre	-0.12
γ	Familienstand: Unverheiratet	1.31
	Familienstand: Geschieden	1.52
	Familienstand: Verwitwet	1.64
	Wohnort: Zentrum	-0.56
	Wohnort: Süd	0.39
	Vertrauen	-0.09
	Ausreichendes Einkommen	-0.33
	Atmosphäre	-0.34
	Staatsangehörigkeit	0.88
	Alter	0.02
	α_0	0.24
	$\frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}_i)$	0.51

Tabelle 5.1: Schätzungen der Parameter $\boldsymbol{\beta}$ und γ des BETAMIX(c)-Modells für die SHIW Studie

Für die CUB-Modelle, sowie für die CUP-Modelle gibt es die Möglichkeit die Effekte der Kovariablen, die sowohl für die Präferenzkomponente als auch für die Unsicherheitskomponente verwendet werden, grafisch darzustellen (vgl. Tutz et al., 2014, S.13f).
Bei der Verwendung des kumulativen Logit-Modells in der Präferenzkomponente ist eine einfache Interpretation der Koeffizienten γ über die Chancen möglich. Diese Interpretationsmöglichkeit lässt sich folgender Darstellung des kumulativen Logit-Modells entnehmen:

$$\frac{P(Y \leq r|\mathbf{x})}{P(Y > r|\mathbf{x})} = \exp(\gamma_{0r}) \exp(\mathbf{x}^T\boldsymbol{\gamma}) = e^{\gamma_{0r}}(e^{\gamma_1})^{x_1} \cdots (e^{\gamma_p})^{x_p}$$

Die Chance auf die Kategorie r oder niedriger im Verhältnis zu einer höheren Kategorie ändert sich um den Faktor $\exp(\gamma_j)$, wenn die j-te Variable um eine Einheit zunimmt. Dies gilt bei Konstanthaltung aller anderen Kovariablen und gilt unabhängig von der Responsekategorie.
In Abbildung 5.7 wird der Faktor $\exp(\gamma_j)$ gegen die Unsicherheitswahrscheinlichkeit $1 - \pi_j$ für die Kovariablen „Familienstand“und „Wohnort“abgetragen. Die Koeffizienten für die Unsicherheit sind im Modell $logit(\pi_i) = \mathbf{z}_i^T\boldsymbol{\beta}$ beinhaltet. Um eine Skala für die Unsicherheit $1 - \pi$ zu erhalten, werden die anderen Kovariablen auf einen festen Wert gesetzt. Auch für $\exp(\mathbf{x}^T\boldsymbol{\gamma})$ trifft dies auf den Variablenvektor zu. Die gewählten festen Werte sind Kategorie 1 für Vertrauen in andere Personen und Interviewatmosphäre, Kategorie 3 für das ausreichende Einkommen während für die anderen Variablen 0 gewählt wird.
Ein großer Wert für $\exp(\mathbf{x}^T\boldsymbol{\gamma})$ zeigt eine Präferenz der kleineren Responsekategorien an. In der SHIW Studie weisen große Werte für $\exp(\mathbf{x}^T\boldsymbol{\gamma})$ auf Unzufriedenheit hin.

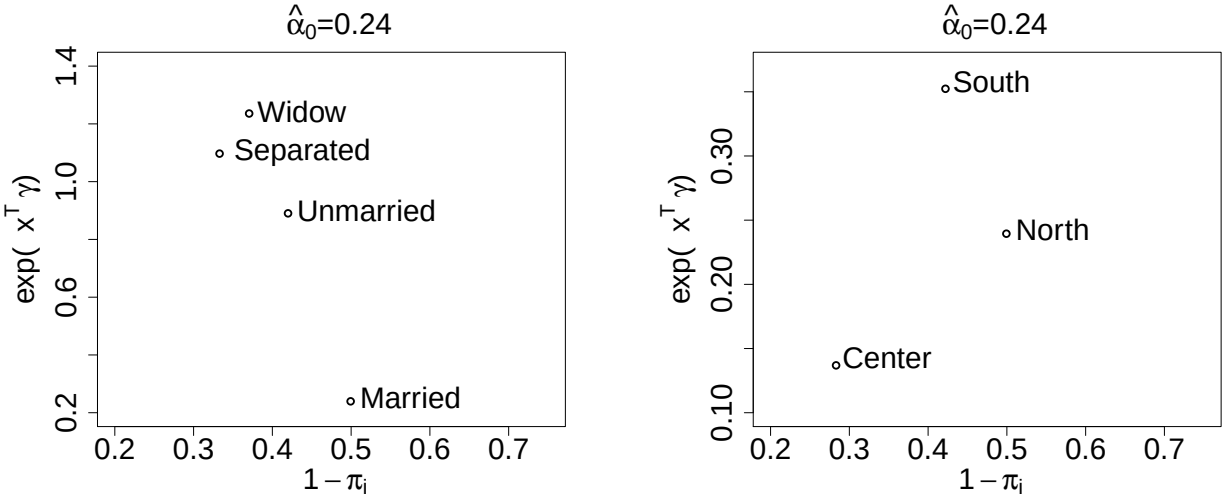

Abbildung 5.7: Effekte der kategorialen Kovariablen Familienstand (links) und Wohnort (rechts) in der Präferenz und der Unsicherheitskomponente

Die Form der Betabinomialverteilung in der Unsicherheitskomponente wird durch den Parameter α bestimmt. Dieser Parameter ist im Modell $\log(\alpha) = \mathbf{w}^T\boldsymbol{\alpha}$ beinhaltet. In dem BETAMIX(c)-Modell, welches in Tabelle 5.1 und Abbildung 5.7 betrachtet wird, ist das Modell für α nur durch $\alpha = \exp(\alpha_0)$ bestimmt, da keine Kovariablen zur Parametrisierung verwendet werden. Es scheint eine leichte Tendenz zu den mittleren Kategorien zu bestehen, da $\exp(\alpha_0) = 1.27$. In Abbildung 5.7 ist zu erkennen, dass der Familienstand „verwitwet“ („widow“) einen vergleichsweise höheren Wert für $\exp(\mathbf{x}^T\boldsymbol{\gamma})$, aber einen geringeren Wert für die Unsicherheit aufweist. Befragte mit Familienstand „verheiratet“ („married“) hingegen scheinen glücklicher, aber unsicherer zu sein. Außerdem scheinen Befragte mit dem Wohnort „Center“ im Vergleich zu den in Süden oder Norden lebenden Personen zu Kategorien der Zufriedenheit zu tendieren und nur wenig Unsicherheit aufzuweisen.

Kovariablen zur Parametrisierung von α

Die Modellierung eines Responsestyles ist durch die Verknüpfung des Parameters α mit Kovariablen möglich. Für die SHIW Studie werden unterschiedliche BETAMIX(c)-Modelle angepasst. Das BETAMIX(c)-Modell mit den Kovariablenvektoren $\mathbf{x}_i = \mathbf{w}_i = (ETA_i, CIT_i, STACIV_i, AREA_i,$

$FIDGEN_i, CONDGEN_i, KLIMA_i, ETA_i^2, ETA_i^3, FIDGEN_i^2, FIDGEN_i^3$) ist unter den betrachteten BETAMIX-Modellen, welche einen Responsestyle modellieren, zu bevorzugen. Der AIC-Wert des Modells beträgt hier $AIC = 15997.9$. Vergleicht man diesen AIC-Wert mit dem Wert des BETAMIX(c)-Modells aus dem vorherigen Abschnitt ($AIC = 16214.89$) sieht man, dass die Berücksichtigung eines Responsestyles eine Verbesserung der Datenanpassung erzielt.
Die Tabelle 5.2 zeigt die geschätzten Koeffizienten $\boldsymbol{\alpha}$ und γ für dieses angepasste BETAMIX(c)-Modell. Auch in Tutz und Berger (2015) wird ein Modell, welches einen Responsestyle ohne Mischverteilungsansatz berücksichtigt, für die SHIW Studie angepasst. Dabei wird sowohl für die Präferenz als auch für den Responsestyle ein Kovariablenvektor verwendet, der jedoch neben den zuvor genannten Kovariablen auch die Variable Geschlecht verwendet ((vgl. Tutz und Berger, 2015, S.7ff)).

Bei der Nutzung des „Rating Scale model acounting for Response Style“ (RSRS), welches in Tutz und Berger (2015) beschrieben wird, weisen die Kategorien „verheiratet“, „verwitwet“,„geschieden“ verglichen mit der Kategorie „unverheiratet“ eine Präferenz der höheren Kategorien der Responsevariablen auf. Bezüglich des Responsestyles lässt sich mittels des RSRS erkennen, dass unverheiratete und verwitwete Personen eher zu den mittleren Kategorien tendieren, wohingegen bei Verheirateten eine Tendenz zu den extremere Kategorien zu beobachten ist. Für den Wohnort ergibt die Analyse mit dem RSRS Modell, dass Personen, die im Zentrum leben, vergleichsweise glücklicher sind. Außerdem scheinen im Süden lebende Personen verglichen zu den im Norden weniger extrem zu antworten. Die Kategorie „Zentrum“ scheint im Vergleich zur Kategorie „Norden“ eine Tendenz zu den extremeren Kategorien aufzuweisen. Dieser Unterschied im Responsestyle kann jedoch vernachlässigt werden. Dies lässt sich dem berechneten Konfidenzintervall entnehmen (vgl. Tutz und Berger, 2015, S.8f).

In Abbildung 5.8 sind die Effekte der Kovariablen „Familienstand“ und „Wohnort“ für die Präferenz und den Responsestyle abgebildet. Werte $\exp(\gamma_j) > 1$ zeigen eine Bevorzugung kleiner Responsekategorien an. Große Werte für $\exp(\alpha_j)$ weisen darauf hin, dass eine Tendenz zu den mittleren Responsekategorien vorliegt.
Verglichen mit der Referenzkategorie „verheiratet“ scheinen Personen aus den anderen Kategorien eher Responsekategorien, die auf Unzufriedenheit hinweisen, zu bevorzugen. Auch im RSRS scheinen Personen aus der Kategorie „verheiratet“ im Vergleich zu „un-

	Kovariablen	Koeffizienten
$\boldsymbol{\alpha}$	Konstante (α_0)	1.53
	Familienstand: Unverheiratet	1.06
	Familienstand: Geschieden	0.95
	Familienstand: Verwitwet	1.27
	Wohnort: Zentrum	-0.42
	Wohnort: Süd	0.06
	Vertrauen	0.69
	Ausreichendes Einkommen	-0.12
	Atmosphäre	-0.29
	Staatsangehörigkeit	0.51
	Alter	-0.007
	Alter^2	-0.0007
	Alter^3	0.00002
	Vertrauen^2	-0.08
	Vertrauen^3	0.003
γ	Familienstand: Unverheiratet	1.1
	Familienstand: Geschieden	1.28
	Familienstand: Verwitwet	1.20
	Wohnort: Zentrum	-0.38
	Wohnort: Süd	0.50
	Vertrauen	-1.76
	Ausreichendes Einkommen	-0.35
	Atmosphäre	-0.38
	Staatsangehörigkeit	0.83
	Alter	0.01
	Alter^2	-0.0003
	Alter^3	0.00003
	Vertrauen^2	0.35
	Vertrauen^3	-0.02
	$(1 - \hat{\pi})$	0.63

Tabelle 5.2: Schätzungen der Parameter γ und $\boldsymbol{\alpha}$ des BETAMIX(c)-Modells mit Responsestyle für die SHIW Studie

verheiratet“ zufriedener zu sein. Außerdem scheint bei der Anpassung des BETAMIX(c)-Modells die Tendenz der Kategorien „unverheiratet“,„verwitwet“, „geschieden“ im Vergleich zu „verheiratet“ eher in Richtung der mittleren Kategorien zu gehen. Im RSRS scheinen Verwitwete im Vergleich zu Unverheirateten eher zu den mittleren Kategorien zu tendieren, wohingegen Verheiratete im Vergleich zu Unverheirateten zu den äußeren Kategorien tendieren. Diese Tendenz lässt sich auch im BETAMIX(c)-Modell erkennen (vgl. Abbildung 5.8).
Bei der Variablen „Wohnort“scheint das Wohnen im Zentrum zufriedener zu machen. Dies ergibt sich auch unter Nutzung des RSRS (vgl. Tutz und Berger, 2015, S.8)). Außerdem ist dem Wert des Koeffizienten $\alpha_{Zentrum}$ zu entnehmen, dass Befragte mit der Ausprägung „Zentrum“im Vergleich zu denen mit der Ausprägung „Norden“zu den extremeren Kategorien tendieren, was sich auch bei Anwendung des RSRS ergibt.

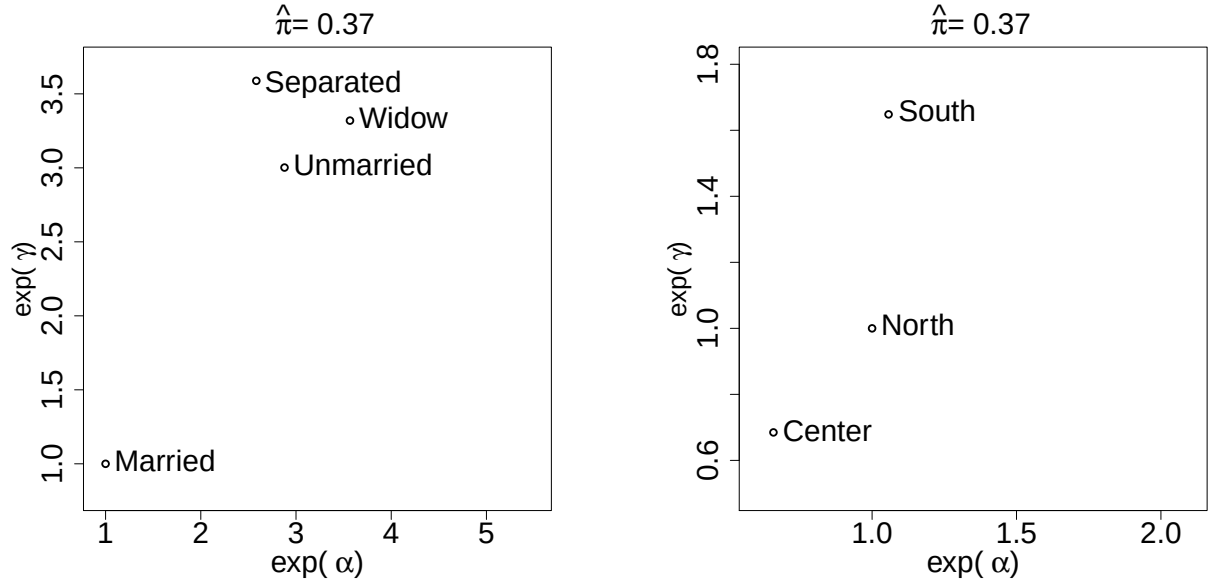

Abbildung 5.8: Effekte der kategorialen Kovariablen Familienstand (links) und Wohnort (rechts) für die Präferenz und den Responsestyle

5.2 Datensatz *Allbus*

Für das zweite Datenbeispiel werden Daten der allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 2012 verwendet. Der genutzte Datensatz besteht aus

Informationen zu 2946 Personen. Für die Analysen wird das Vertrauen in das Gesundheitssystem („healthsys“), welches auf einer Skala von 1 (überhaupt kein Vertrauen) bis 7 (sehr großes Vertrauen) gemessen wird, als Responsevariable verwendet. In Abbildung 5.9 sind die absoluten Häufigkeiten der einzelnen Kategorien der Variablen „healthsys“zu sehen. Die Häufigkeitsverteilung ist rechtssteil und der Modus liegt bei Kategorie 5.

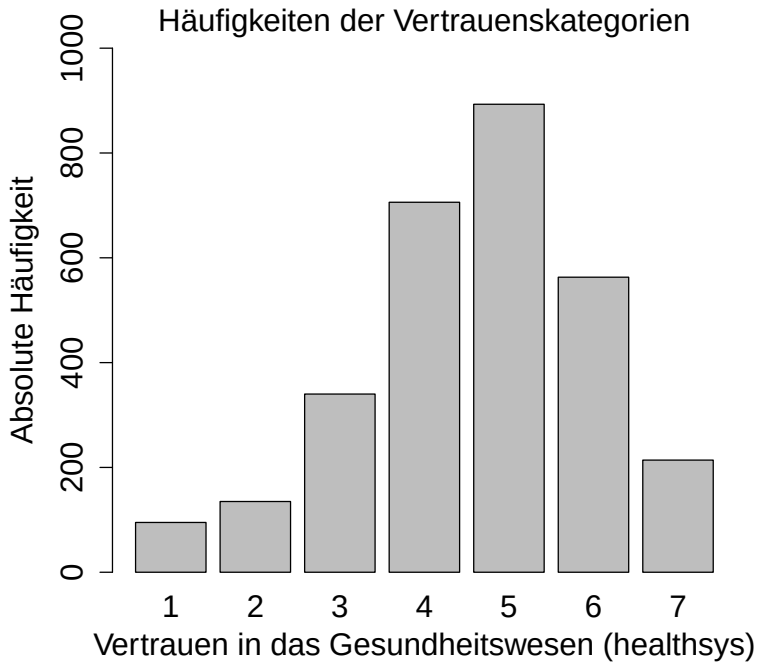

Abbildung 5.9: Balkendiagramm für den Response *healthsys*

Als erklärende Variablen werden folgende verwendet:

female:	Geschlecht (0: männlich, 1: weiblich)
ee2:	Einkommen in Tausend Euro
age2:	Alter (zentriert um 50 Jahre)
health:	Gesundheitszustand des Befragten (von 1: sehr gut bis 5: schlecht)
german:	Staatsangehörigkeit (0: andere, 1: deutsch)
east:	Durchführung des Interviews in Ostdeutschland
happy:	allgemeine Zufriedenheit (von 0: sehr unzufrieden bis 10: sehr zufrieden)

Wahl des Parameters α

Es wird ein BETAMIX(c)-Modell mit dem Kovariablenvektor
$\mathbf{x}_i = (german_i, ee2_i, age2_i, east_i, happy_i)$ angepasst. Für den Vektor $\mathbf{z}_i$, welcher die Wahrscheinlichkeit π parametrisiert, wird eine binären Version der Variablen „health "mit den Referenzkategorien 1,2 und 3 genutzt. Als Modell für den Parameter α aus der Unsicherheitskomponente des BETAMIX-Modells wird das Interceptmodell verwendet.
In Tabelle 5.3 sind die Schätzungen der Koeffizienten γ, $\boldsymbol{\beta}$ und α_0 sowohl für das BETAMIX(c)-Modell als auch für das CUP(c)-Modell dargestellt.

	Kovariablen	BETAMIX(c)	CUP(c)
$\boldsymbol{\beta}$	Konstante (β_0)	2.00	2.06
	Gesundheitszustand: weniger gut/schlecht	-1.68	-1.68
γ	Staatsangehörigkeit: deutsch	1.19	1.18
	Einkommen	0.07	0.06
	Alter	-0.007	-0.007
	allgemeine Zufriedenheit	-0.22	-0.21
	Ost	-0.39	-0.39
	α_0	0.07	
	$\frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}_i)$	0.16	0.16
	AIC	9962.37	9960.25

Tabelle 5.3: Schätzungen der Parameter γ und $\boldsymbol{\beta}$ des BETAMIX(c)- und CUP(c)-Modells für den ALLBUS Datensatz

Man erkennt, dass sich die Schätzungen der Parameter γ und $\boldsymbol{\beta}$ des CUP(c)- und BETAMIX(c)-Modells kaum voneinander unterscheiden. Dies liegt an der Schätzung des Parameters α des BETAMIX(c)-Modells. Für diesen Parameter erhält man den Wert $\exp(\hat{\alpha}_0) = 1.07$. Somit weicht die Verteilung für die Unsicherheitskomponente des BETAMIX(c)-Modells kaum von der diskreten Gleichverteilung ab. Auch der Unterschied der betrachteten AIC-Werte ist sehr gering. Der Parameter α, der im BETAMIX(c)-Modell zusätzlich auftritt, trägt hauptsächlich zu dieser Differenz bei, da eine Bestrafung des zusätzlichen Parameters im AIC-Kriterium des BETAMIX-Modells erfolgt. Hinsichtlich der geschätzten Koeffizienten macht es in diesem Fall jedoch keinen Unterschied, ob man das BETAMIX(c)- oder das CUP(c)-Modell verwendet. An diesem Beispiel erkennt man, dass das BETAMIX(c)-Modell

den Spezialfall CUP(c) gut findet.
Die Verwendung von Kovariablen für die Parametrisierung der Wahrscheinlichkeit π des BETAMIX(c)-Modells hat in der SHIW-Studie zu unterschiedlichen Ergebnissen für den Parameter α bei einer Gitter basierten Wahl und einer Schätzung geführt. Da das BETAMIX(c)-Modell aus Tabelle 5.3 ebenfalls einen **z**-Vektor verwendet, werden in Abbildung 5.10 die AIC-Werte, die sich über ein Gitter mit festen Werten für α ergeben, betrachtet.

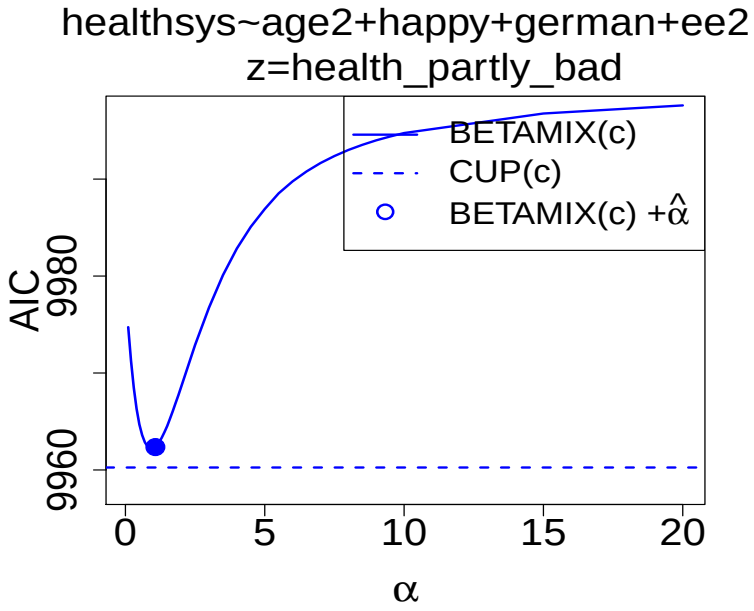

Abbildung 5.10: ALLBUS: AIC-Werte für BETAMIX-Modelle mit festem α

Im Gegensatz zu den Beispielen in Abbildung 5.6 hat die AIC-Kurve aus Abbildung 5.10 ein lokales Minimum, was auch gleichzeitig das Globale zu sein scheint. Somit ist hier, wie auch in den Simulationen, kein größerer Unterschied zwischen einer Wahl und einer Schätzung von α zu beobachten.

Kovariablen zur Parametrisierung von α

Das BETAMIX(c)-Modell mit den Vektoren $\mathbf{x}_i = (german_i, ee2_i, age2_i, east_i, happy_i)$ und $\mathbf{w}_i = (health\_partly\_bad)$ ist ähnlich zu dem in Tabelle 5.3 betrachteten Modell. Die Variable „health_partly_bad“ bezeichnet dabei die binäre Version der Variablen „health“

mit den Kategorien 1,2 und 3 als Referenz. Statt der Verknüpfung der Variablen „health_partly_bad“ mit der Wahrscheinlichkeit π (wie in Tabelle 5.3), wird die Variable hier für den Responsestyle genutzt.

	Kovariablen	BETAMIX(c)
$\boldsymbol{\alpha}$	Konstante (α_0)	2.62
	Gesundheitszustand: weniger gut/schlecht	-2.62
γ	Staatsangehörigkeit: deutsch	1.42
	Einkommen	0.07
	Alter	-0.008
	allgemeine Zufriedenheit	-0.29
	Ost	-0.44
	$1 - \hat{\pi}$	0.30
	AIC	9964.97

Tabelle 5.4: BETAMIX(c)-Modell mit Responsestyle-Effekten: Verwendung des **z**-Vektors aus Tabelle 5.3 als **w**-Vektor

Die geschätzten Koeffizienten γ unterscheiden sich zwischen dem Modell mit **z**-Vektor und dem mit **w**-Vektor nicht stark (vgl. Tabelle 5.3 und Tabelle 5.4). Aus Tabelle Tabelle 5.4 lässt sich eine Tendenz zu den mittleren Responsekategorien für Personen mit einem guten Gesundheitszustand (Kategorien 1,2 und 3 der Variablen „health“) im Vergleich zu den Personen mit schlechten Gesundheitszustand (Kategorie 4 und 5 der Variablen„health“) entnehmen ($\exp(\hat{\alpha}_0) = 13.7$). Für Personen, die ihren Gesundheitszustand mit weniger gut oder schlecht einschätzen, scheint jedoch weder eine Tendenz zu den mittleren noch zu den äußeren Responsekategorien vorzuliegen. Grund hierfür ist $\exp(\hat{\alpha}_0 + \hat{\alpha}_{helath\_partly\_bad}) = 1$. Dieses BETAMIX(c)-Modell ist jedoch laut AIC-Kriterium dem CUP(c)-Modell aus Tabelle 5.3 nicht zu bevorzugen. Der Unterschied der AIC-Werte ist jedoch gering.

Ein BETAMIX(c)-Modell mit Responsestyle, das einen kleineren AIC-Wert als das CUP(c)-Modell aus Tabelle 5.3 aufweist, ist Folgendes: Die Variablen für die Präferenzkomponente (Vektor **x**) sind „german“, „ee2“, „age2“, „health_partly_bad“, „female“, „happy“, „east“. Die Variablen „health_partly_bad“, „female“, „happy“ werden für den Responsestyle (Vektor **w**) genutzt. Die Koeffizientenschätzungen zu diesem Modell sind in Tabelle 5.5 dargestellt.

	Kovariablen	BETAMIX(c)
$\boldsymbol{\alpha}$	Konstante (α_0)	2.23
	Gesundheitszustand: weniger gut/schlecht	-4.2
	Geschlecht: weiblich	-1.75
	allgemeine Zufriedenheit	0.44
γ	Staatsangehörigkeit: deutsch	1.41
	Einkommen	0.08
	Alter	-0.009
	allgemeine Zufriedenheit	-0.27
	Gesundheitszustand: weniger gut/schlecht	0.25
	Ost	-0.43
	Geschlecht: weiblich	0.31
	$1 - \hat{\pi}$	0.30
	AIC	9949.68

Tabelle 5.5: Schätzungen der Parameter γ und $\boldsymbol{\alpha}$ des BETAMIX(c)-Modells mit Responsestyle-Effekten für den ALLBUS Datensatz

Mit zunehmender allgemeinen Zufriedenheit scheint die Tendenz zu den mittleren Kategorien zuzunehmen. Bei der Variablen „health_partly_bad“, sowie bei „female“ liegt im Vergleich zu der jeweiligen Referenzkategorie eine Tendenz zu den extremen Kategorien der Responsevariablen „healthsys“ vor. Die Wahrscheinlichkeit für diesen Responsestyle $(1 - \hat{\pi})$ beträgt 0.30.
Aus dem kumulativen Logit-Modell des BETAMIX(c)-Modells lassen sich folgende Aussagen treffen: Die Chance auf das Vertrauen in das Gesundheitssystem mit Stärke r (bzw. Kategorie r) oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist für deutsche Staatsbürger um den Faktor $\exp(1.41) = 4.09$ höher als für nicht deutsche Staatsbürger.
Für im Osten lebende Personen verringert sich diese Chance im Vergleich zu im Westen lebenden Personen um den Faktor 0.65. Die Chance auf die Präferenz der Kategorien $\{1, \ldots, r\}$ im Verhältnis zu den Kategorien $\{r + 1, \ldots, k\}$ nimmt auch für ein zunehmendes Alter und zunehmender allgemeinen Zufriedenheit ab. Bei den kategorialen Variablen „health_partly_bad“ ,„female“ und „german“ und der stetigen Variablen Einkommen ist jedoch eine Zunahme dieser Chance zu erkennen. Diese Zunahme der Chance kann als Präferenz der niedrigen Responsekategorien und somit als geringes Vertrauen in das

Gesundheitssystem interpretiert werden.

6 Zusammenfassung und Ausblick

Das BETAMIX-Modell ist ein aus zwei Komponenten bestehendes Mischmodell, welches für die Analyse einer ordinalen Responsevariablen geeignet ist. Das Modell besteht aus einem beliebigen ordinalen Regressionsmodell und einer beschränkten Version der Betabinomialverteilung. Erfolgt im BETAMIX(c)-Modell die Annahme, dass der Parameter α für alle Individuen gleich ist, so besteht die Möglichkeit den Parameter zu schätzen oder über das Abtasten eines Gitters zu wählen.
Anhand der Simulationsszenarien lässt sich erkennen, dass beide Möglichkeiten zu Werten führen, die nur geringe Abweichungen zum wahren Wert aufweisen. Die Ergebnisse für das Abtasten und das Schätzen sind nicht nur gut, sie unterscheiden sich auch kaum voneinander. Die Schätzung ist weniger zeitaufwendig als das Abtasten. Aus diesem Grund ist die Schätzung des Parameters zu empfehlen.
Nicht nur der Parameter α wird gut geschätzt, sondern auch die anderen Parameter des BETAMIX(c)-Modells. Wird ein Responsestyle simuliert, was durch Verknüpfung von Kovariablen mit dem Parameter α erzielt wird, werden die Koeffizienten den entsprechenden Simulationsszenarien zufolge gut geschätzt. Den Simulationsergebnissen kann man auch entnehmen, dass die Effekte erklärender Variablen in Richtung null geschrumpft werden, wenn der Responsestyle ignoriert wird. Dies erfolgt beispielsweise durch die Anpassung eines CUP- Modells, statt der Anpassung eines BETAMIX-Modells mit der Parametrisierung des α's durch Kovariablen.
Ein Vorteil des BETAMIX-Modells stellt wie bei CUP- und CUB-Modell die grafische Darstellung der Effekte dar. Dabei können die Effekte der Präferenzkomponente sowohl mit der Unsicherheit als auch mit den Effekten des Responsestyles in einer Grafik dargestellt werden.
In dieser Arbeit wurde für die Modellwahl hauptsächlich das AIC- und BIC-Kriterium betrachtet. Dabei wäre es weiterhin interessant die Prognosegüte für die Modellwahl heranzuziehen. Man könnte beispielsweise die hier über das AIC-Kriterium gewählten α auch mittels Kreuzvalidierung wählen. Eine Simulationsstudie hierzu könnte Aufschluss geben, ob die so gewählten Werte für α besser als die über das AIC gewählte oder die geschätzten Werte sind.
Aufgrund der Verwendung des EM-Algorithmus für die Schätzung der Koeffizienten des BETAMIX-Modells werden die Standardfehler der Koeffizienten nicht automatisch erhalten. Für die Interpretation und für die Beurteilung der geschätzten Effekte sind

die Standardfehler jedoch von Interesse. In Tutz et al. (2014) werden die Standardfehler der Koeffizienten mittels nichtparametrischen Bootstrap geschätzt. Dies sollte auch für die BETAMIX-Modelle möglich sein. Neben der Bootstrap-Methode (Efron und Tibshirani (1998)) bietet auch eine Erweiterung des EM-Algorithmus, der Supplemented-EM-Algorithmus (Little und Rubin (1987)), eine Möglichkeit zum Erhalt der Standardfehler. Im BETAMIX-Modell können Kovariablen mit der Präferenzkomponente und der Mischwahrscheinlichkeit oder der Präferenzkomponente und dem Parameter α verknüpft werden. Da alle in einem Datensatz zur Verfügung stehenden Kovariablen für die zwei Verknüpfungen verwendet werden können, können doppelt so viele Koeffizienten wie Kovariablen zur Schätzung anfallen. Dies macht eine Variablenselektion im BETAMIX-Modell notwendig, was eventuell mit einer Penalisierung der log-Likelihood erreicht werden kann.

Literaturverzeichnis

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society 46*(1), 1–30.

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *International Marketing Review 18*, 301–324.

Efron, B. und R. J. Tibshirani (1998). *An Introduction to the Bootstrap*. Monographs on statistics and applied probability. Chapman & Hall.

Gneiting, T. und A. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Society* (102), 359–376.

Iannario, M. und D. Piccolo (2012). Cub models: Statistical methods and empirical evidence. In *Modern Analysis of Customer Surveys: with applicarions using R*, S. 231–258. New York: Wiley.

Little, R. J. und D. B. Rubin (1987). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Wiley.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society 42*(2), 109–142.

Meisenberg, G. und A. Williams (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences 44*(7), 1539–1550.

Poessnecker, W. (2015). Mrsp: Multinomial response models with structured penalties. *R package version 0.6.0*.

Tutz, G. (2000). *Die Analyse kategorialer Daten*. Lehr- und Handbücher der Statistik. Oldenbourg.

Tutz, G. (2012). *Regression for categorical data*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press.

Tutz, G. und M. Berger (2015, August). Response styles in rating scales - simultaneous modelling of content-related effects and the tendency to middle or extreme categories. *Technical Report, University of Munich*.

Tutz, G., M. Schneider, M. Iannario, und D. Piccolo (2014). Mixture models for ordinal responses to account for uncertainty of choice. *Technical Report Number 175,2014, University of Munich*.

Van Herk, H., Y. Poortinga, und T. Verhallen (2004). Response styles in rating scales evidence of method bias in data from six eu countries. *Journal of Cross-Cultural Psychology 35*(3), 346–360.

Yee, T. W. (2015). Vgam: Vector generalized linear and additive models. *R package version 1.0-0.*.

A Weitere graphische Auswertungen

A.1 Szenario 1

Die folgenden Abbildungen sind weitere Ergebnisse zu dem Simulationsszenario 1 (siehe Abschnitt 4.1).

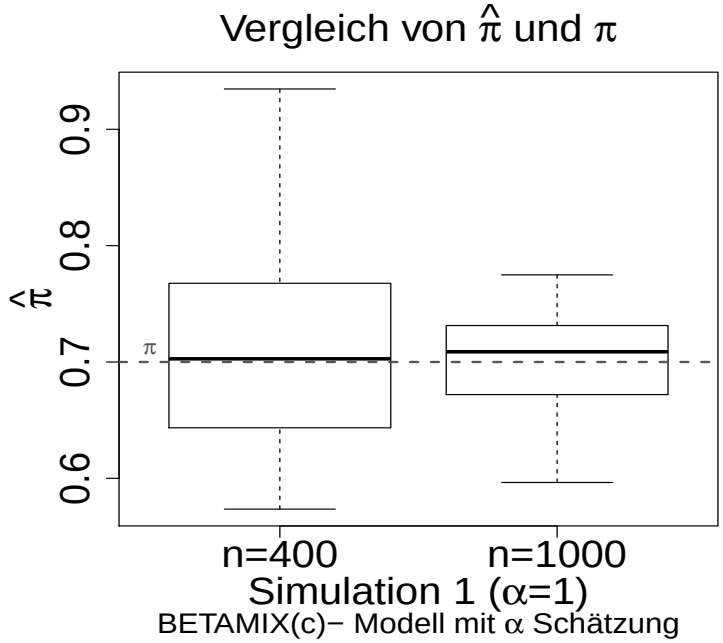

Abbildung A.1: Szenario 1($\alpha = 1$): Boxplot des Parameters $\hat{\pi}$

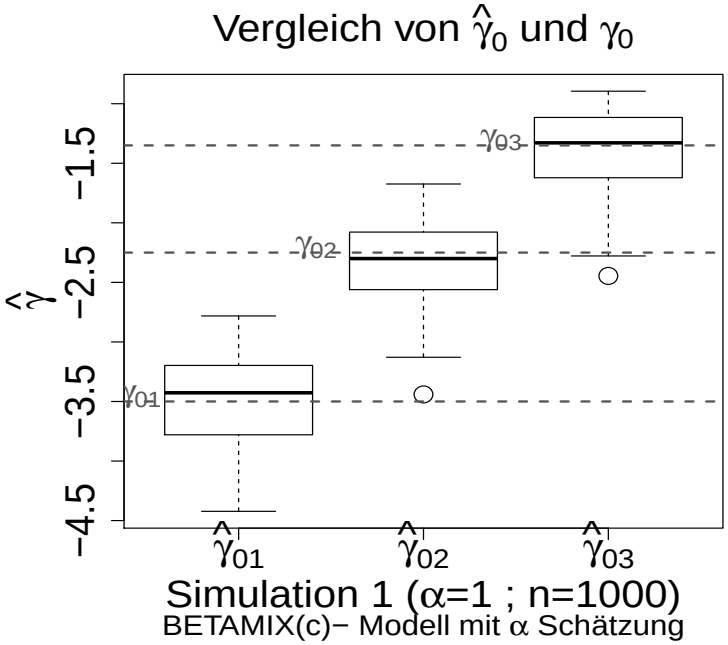

Abbildung A.2: Szenario 1 ($\alpha = 1; n = 1000$): Boxplot des Koeffizientenvektors $\hat{\gamma}_0$

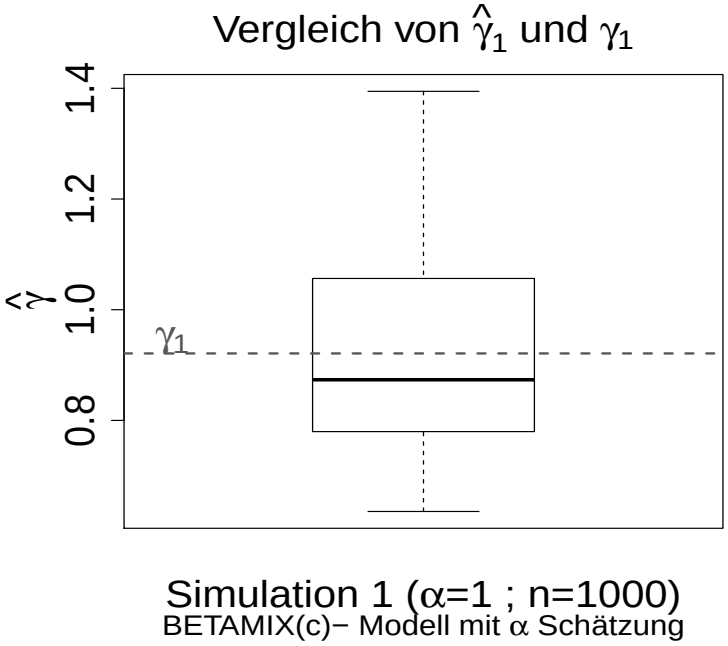

Abbildung A.3: Szenario 1 ($\alpha = 1; n = 1000$): Boxplot des Koeffizienten $\hat{\gamma}_1$

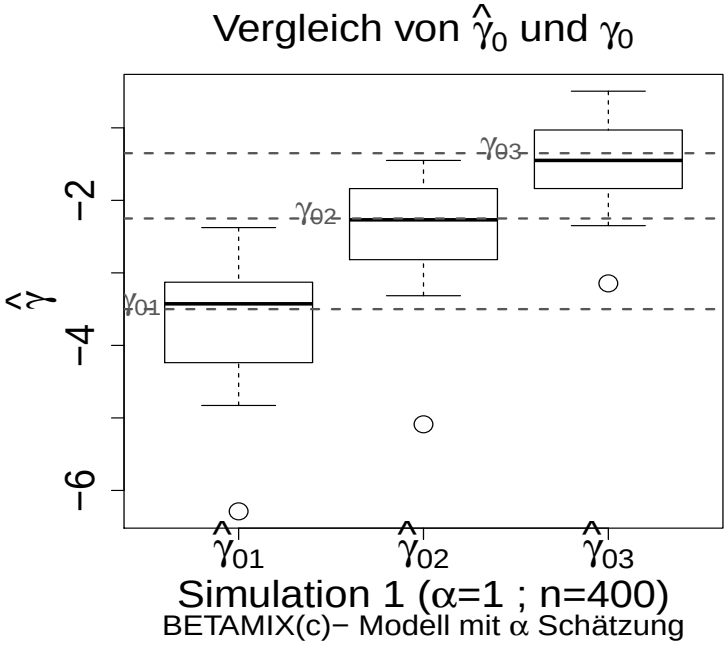

Abbildung A.4: Szenario 1 ($\alpha = 1; n = 400$): Boxplot des Koeffizientenvektors $\hat{\gamma}_0$

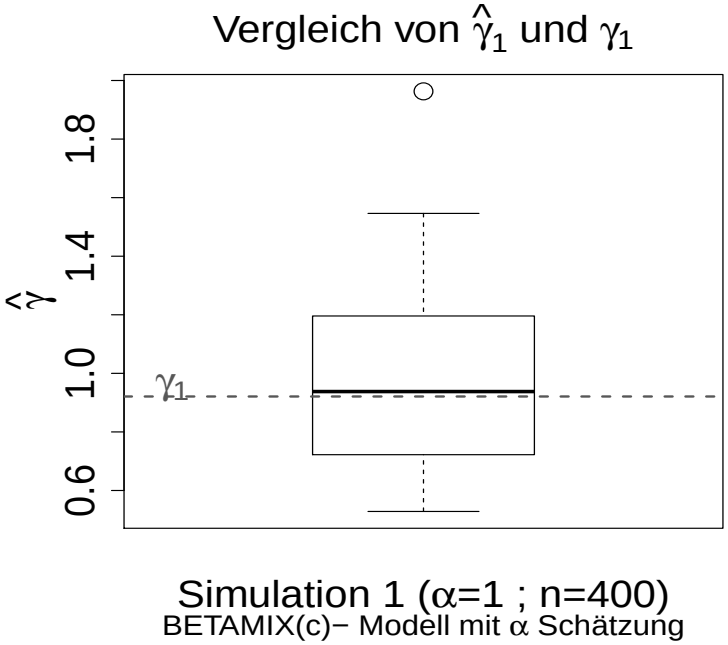

Abbildung A.5: Szenario 1 ($\alpha = 1; n = 400$): Boxplot des Koeffizienten $\hat{\gamma}_1$

Die Boxplots aus Abbildung A.6 basieren auf jeweils 18 MSE- Werten, da nur Werte kleiner

als 10 zur Erstellung des Plots verwendet wurden.

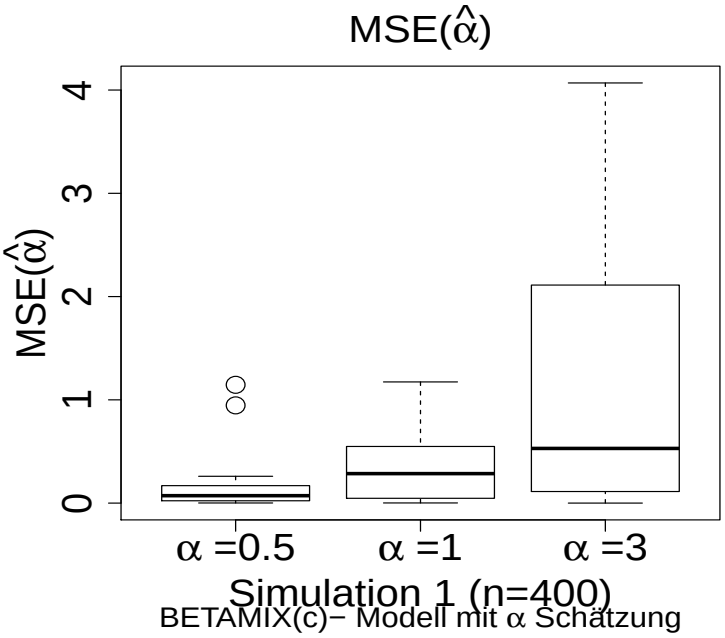

Abbildung A.6: $MSE(\hat{\alpha})$ des Szenarios 1 ($n = 400$) mit $\alpha = 0.5$, $\alpha = 1$ und $\alpha = 3$

A.2 Szenario 2

Für das Szenario 2 ($\alpha = 0.5; n = 400$) sind die mittels BETAMIX(c)-Modell geschätzten Koeffizienten in den folgenden Abbildungen dargestellt.

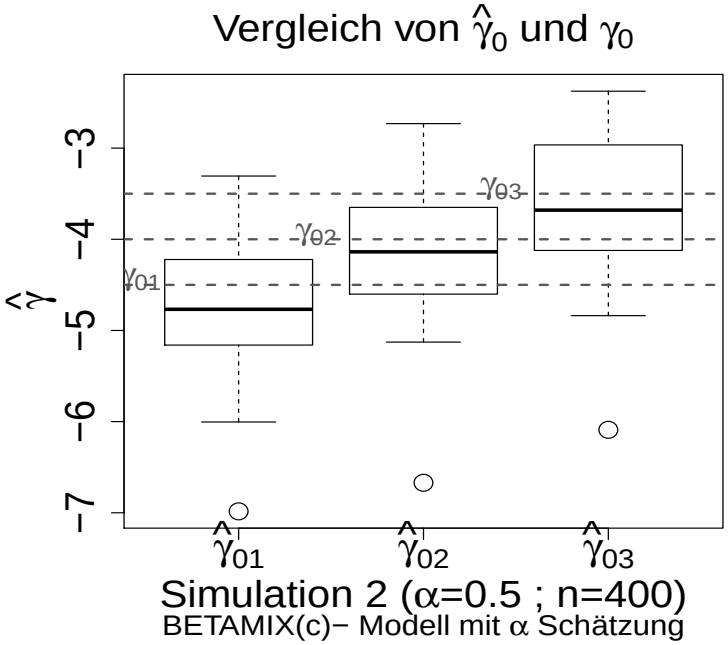

Abbildung A.7: Szenario 2 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

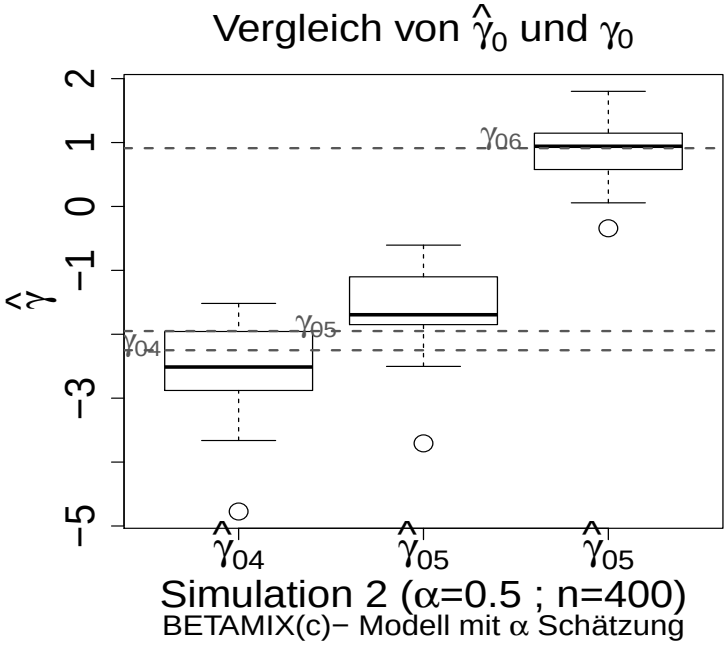

Abbildung A.8: Szenario 2 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

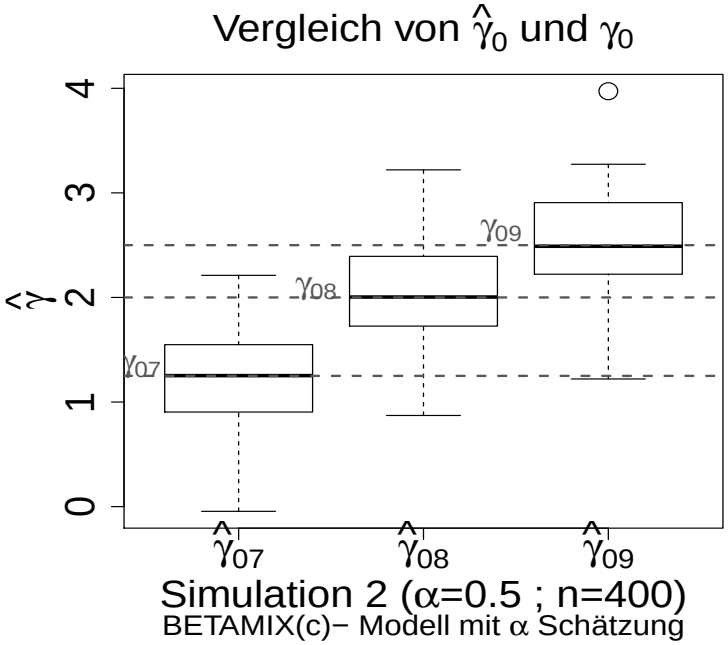

Abbildung A.9: Szenario 2 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{07}, \ldots, \gamma_{09}$

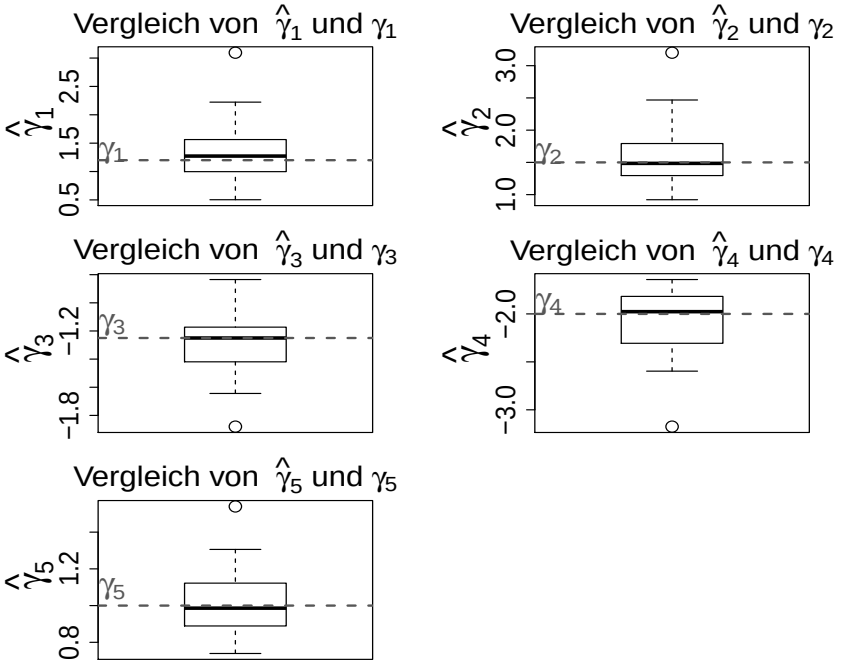

Abbildung A.10: Szenario 2 ($\alpha = 0.5; n = 400$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

Die Wahl und die Schätzung des Parameters α scheinen den folgenden Abbildungen nach

zu urteilen ähnlichen Werte für α zu liefern (vgl.Abbildung A.11 bis Abbildung A.14).

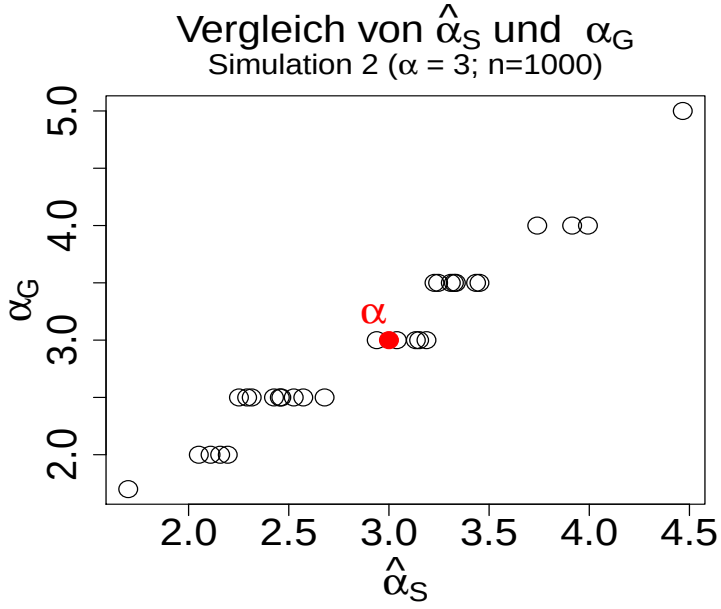

Abbildung A.11: Szenario 2 ($\alpha = 3; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

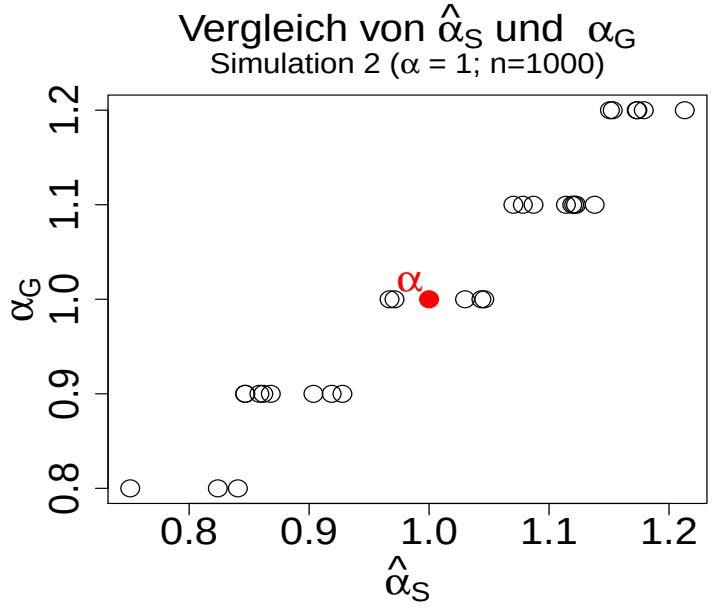

Abbildung A.12: Szenario 2 ($\alpha = 1; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

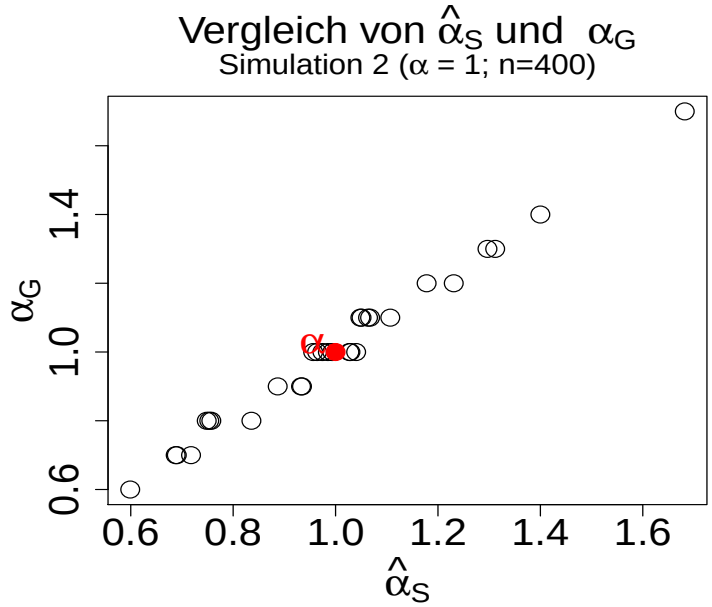

Abbildung A.13: Szenario 2 ($\alpha = 1; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

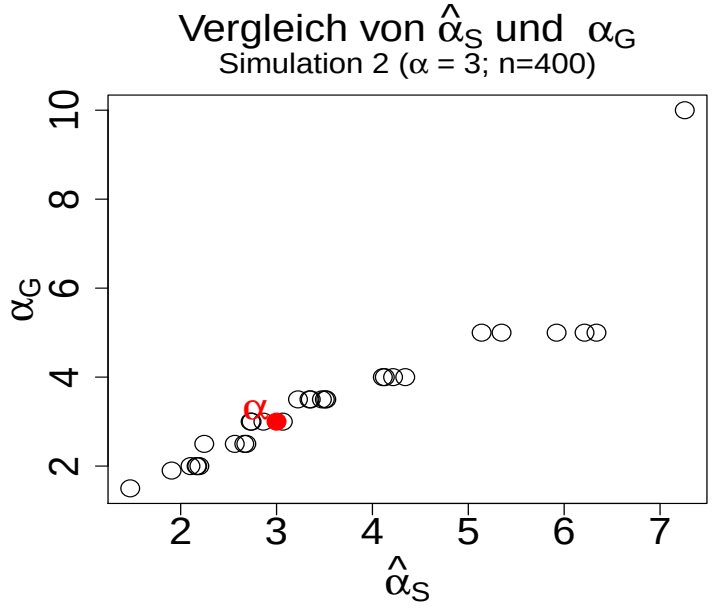

Abbildung A.14: Szenario 2 ($\alpha = 3; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Für die Abbildung A.15 wird im Gegensatz zu Abbildung 4.18 ein feineres Gitter verwendet: $(0.1, 0.12, \ldots, 0.98, 1, 1.1, 1.2, 1.3)$.

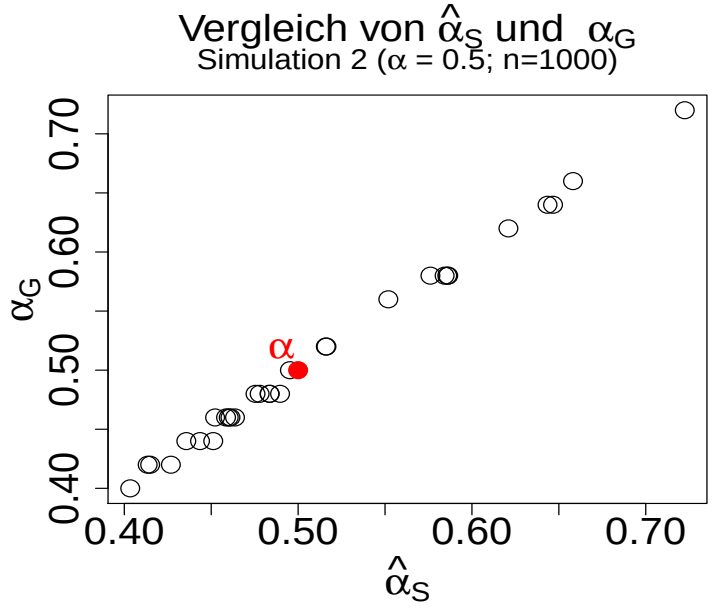

Abbildung A.15: Szenario 2 ($\alpha = 0.5; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und auf einem feinen Gitter gewählten α_G

A.3 Szenario 3

Die Schätzungen der Koeffizienten des Szeanrios 3 ($\alpha = 0.5$) sind in den folgenden Grafiken (Abbildung A.16-Abbildung A.31) zu sehen.

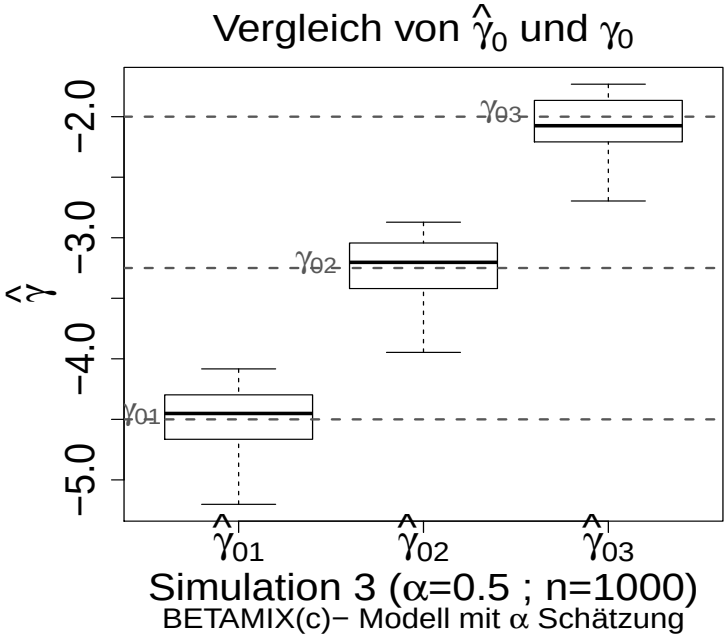

Abbildung A.16: Szenario 3 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

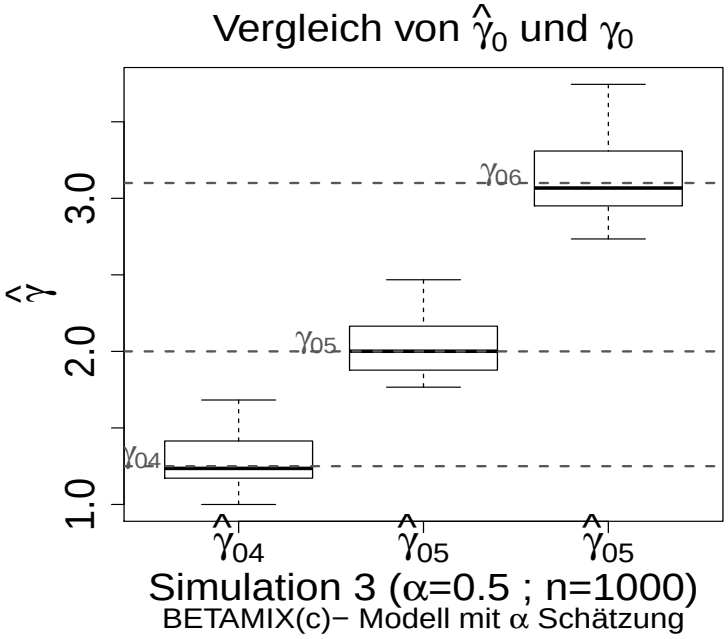

Abbildung A.17: Szenario 3 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

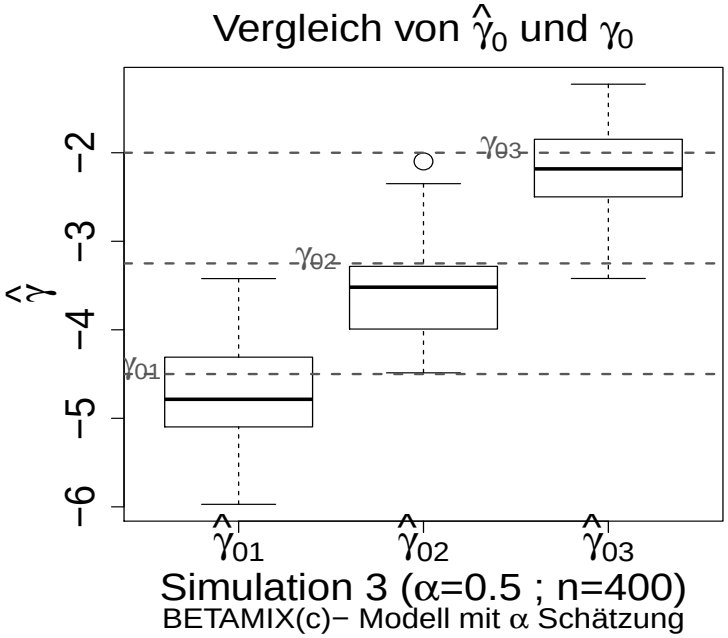

Abbildung A.18: Szenario 3 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

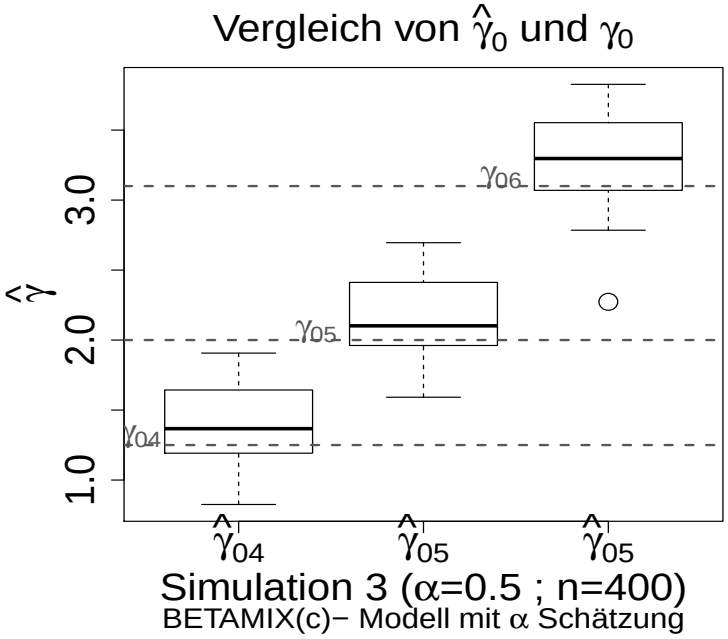

Abbildung A.19: Szenario 3 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

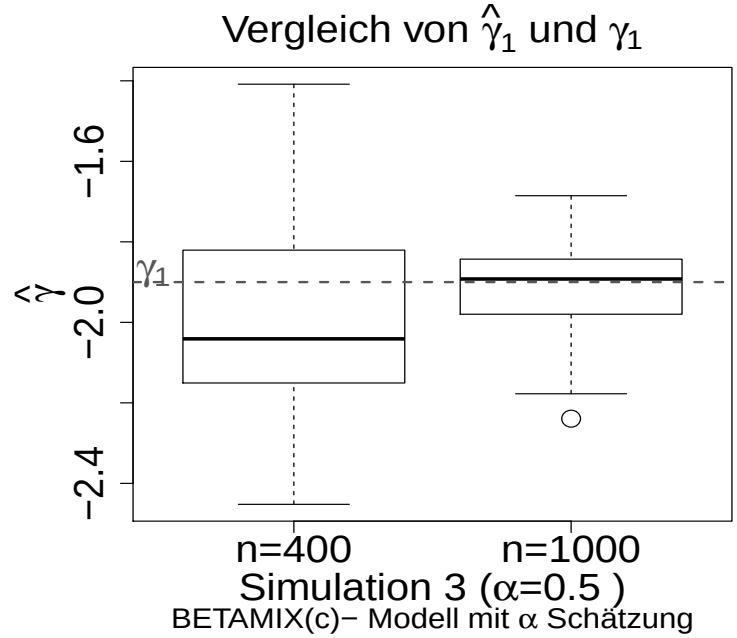

Abbildung A.20: Szenario 3 ($\alpha = 0.5$): Boxplot des Parameters γ_1

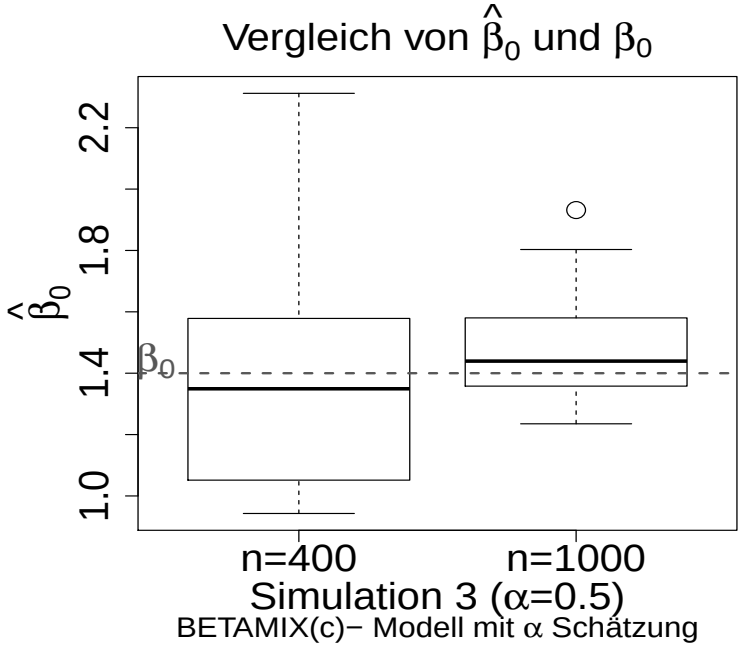

Abbildung A.21: Szenario 3 ($\alpha = 0.5$): Boxplot des Parameters β_0

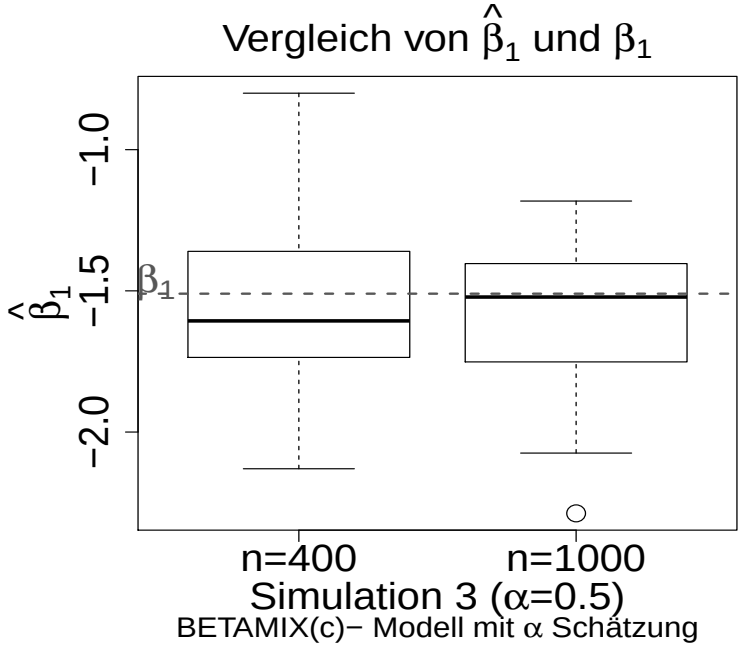

Abbildung A.22: Szenario 3 ($\alpha = 0.5$): Boxplot des Parameters β_1

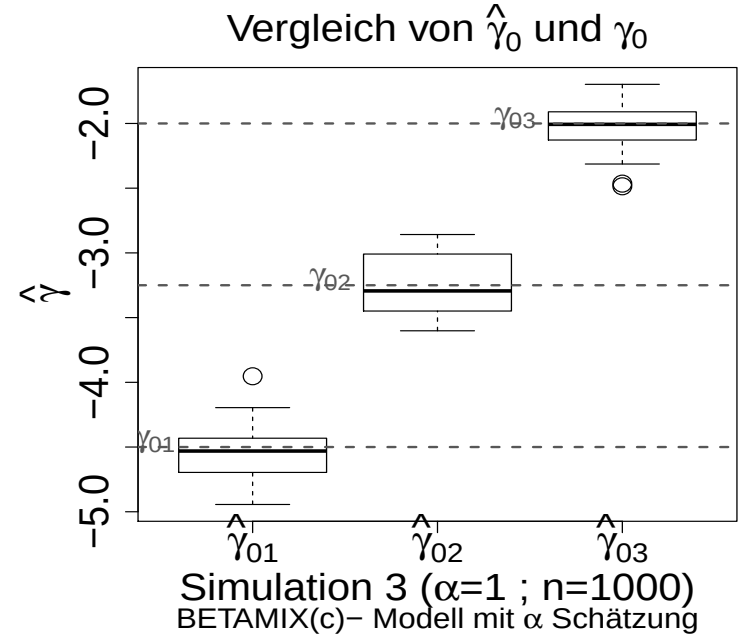

Abbildung A.23: Szenario 3 ($\alpha = 1; n = 1000$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

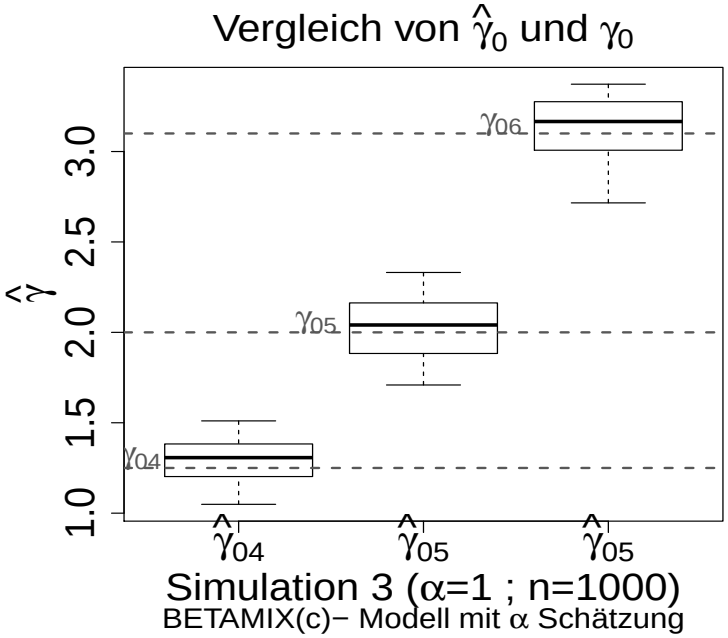

Abbildung A.24: Szenario 3 ($\alpha = 1; n = 1000$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{04}$

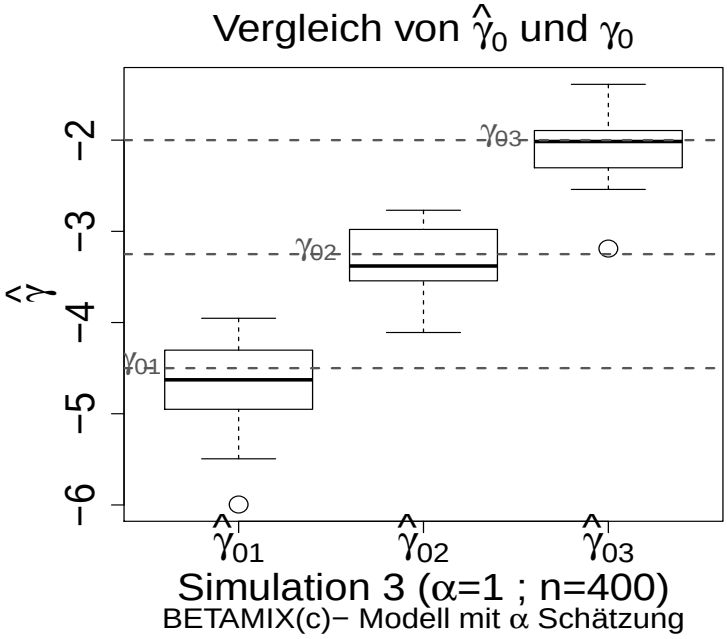

Abbildung A.25: Szenario 3 ($\alpha = 1; n = 400$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

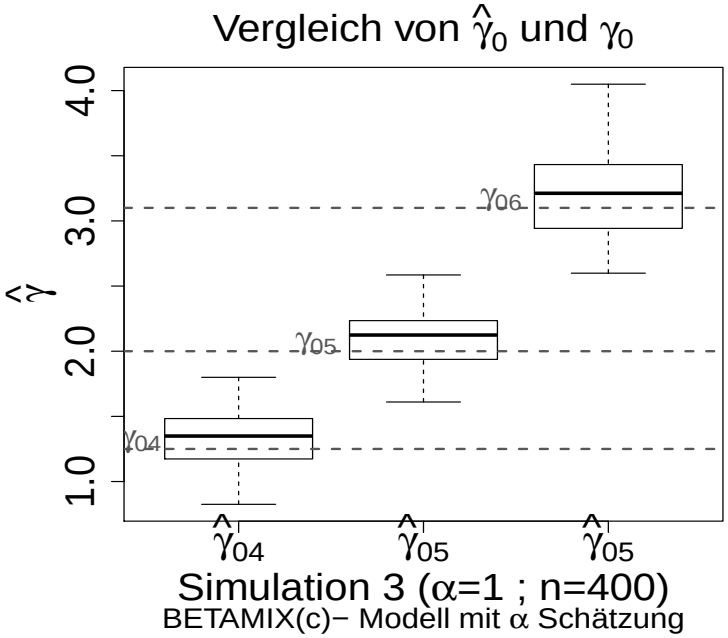

Abbildung A.26: Szenario 3 ($\alpha = 1; n = 400$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

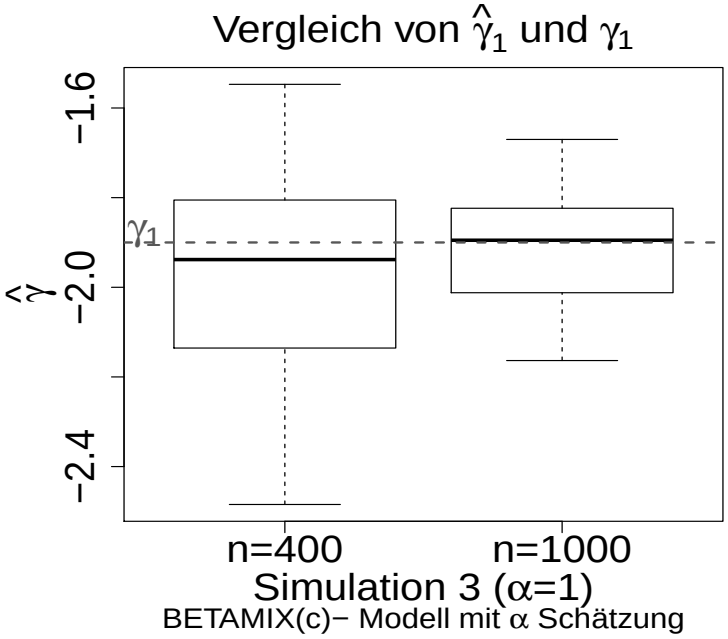

Abbildung A.27: Szenario 3 ($\alpha = 1$): Boxplot des Parameters γ_1

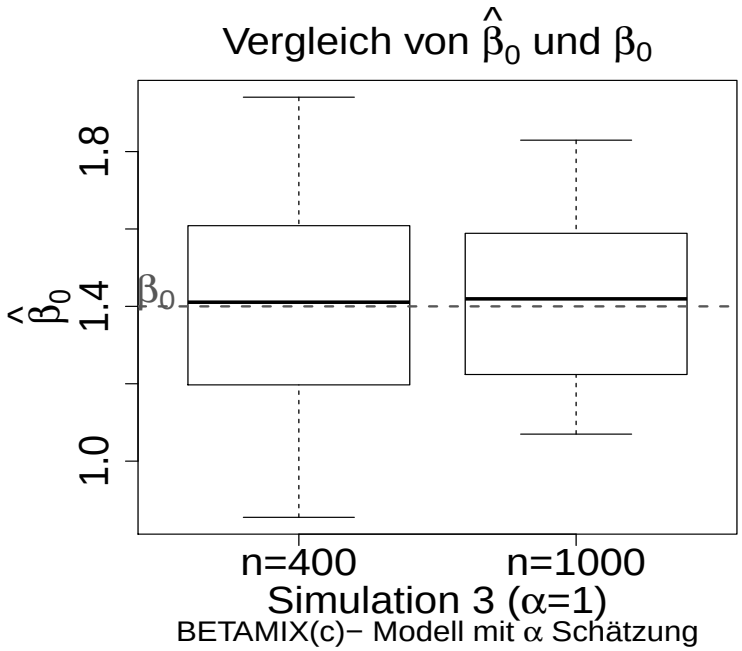

Abbildung A.28: Szenario 3 ($\alpha = 1$): Boxplot des Parameters β_0

In Abbildung A.29 lässt sich erkennen, dass der Median der geschätzten $\hat{\beta}_1$ bei beiden Beobachtungsanzahlen um ungefähr 0.1 vom wahren Wert abweicht.

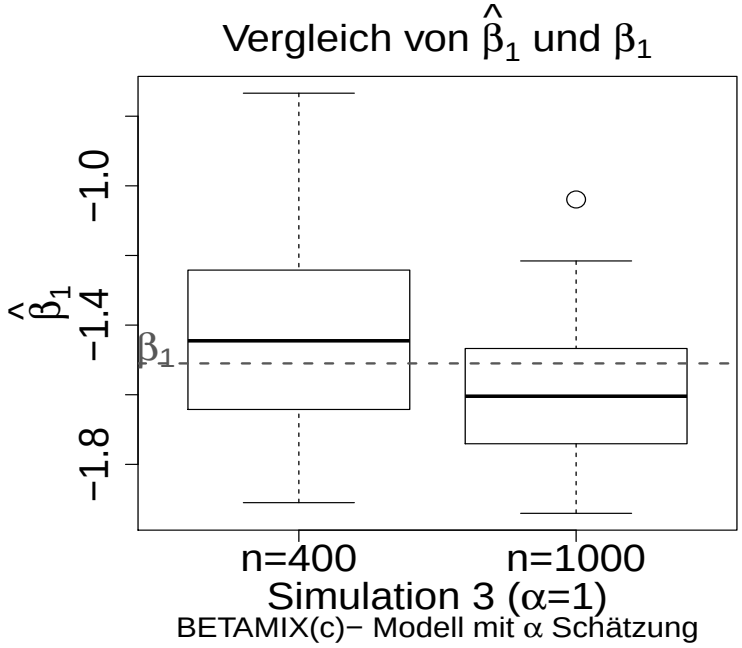

Abbildung A.29: Szenario 3 ($\alpha = 1$): Boxplot des Parameters β_1

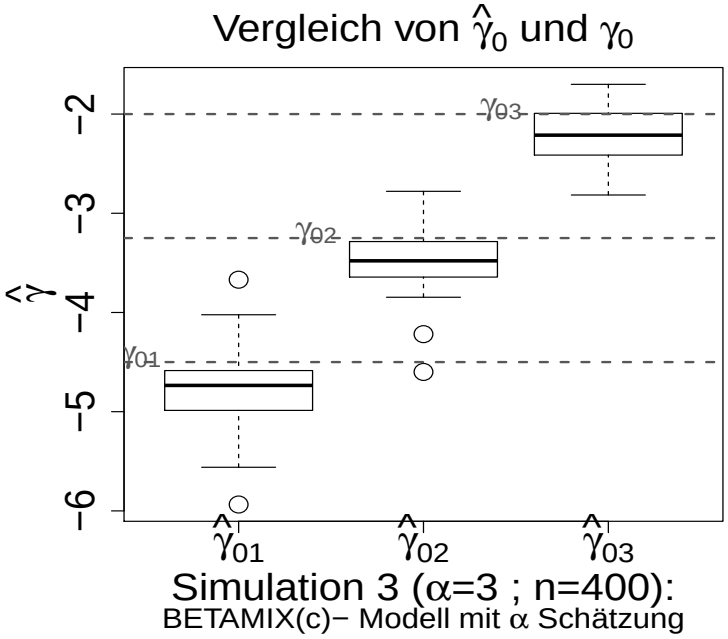

Abbildung A.30: Szenario 3 ($\alpha = 3; n = 400$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

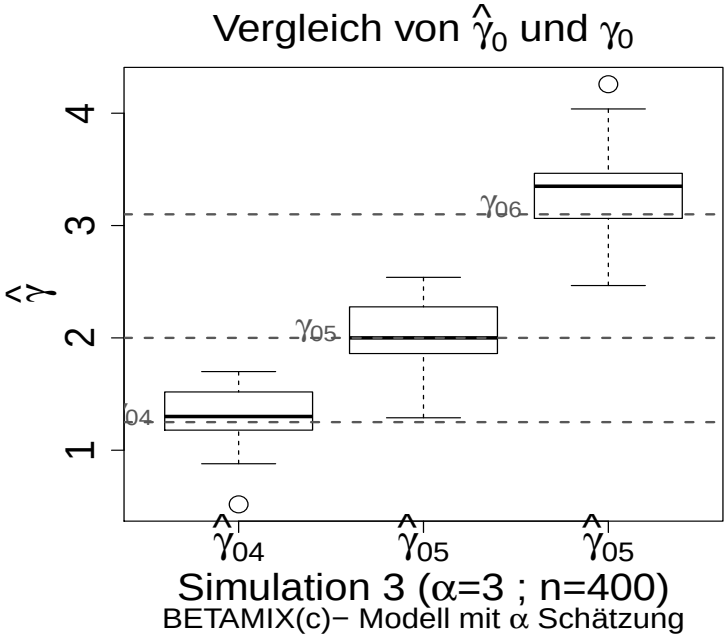

Abbildung A.31: Szenario 3 ($\alpha = 3; n = 400$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

Wie bereits in den Abbildungen 4.30 und 4.31 zu sehen ist, unterscheiden sich gewähltes und geschätztes α auch bei der Hinzunahme einer Kovariablen zur Parametrisierung der Wahrscheinlichkeit π, was in Szenario 3 gemacht wird, kaum. Abbildung A.32 zeigt ebenfalls, dass die Werte auf der 1. Winkelhalbierenden liegen.

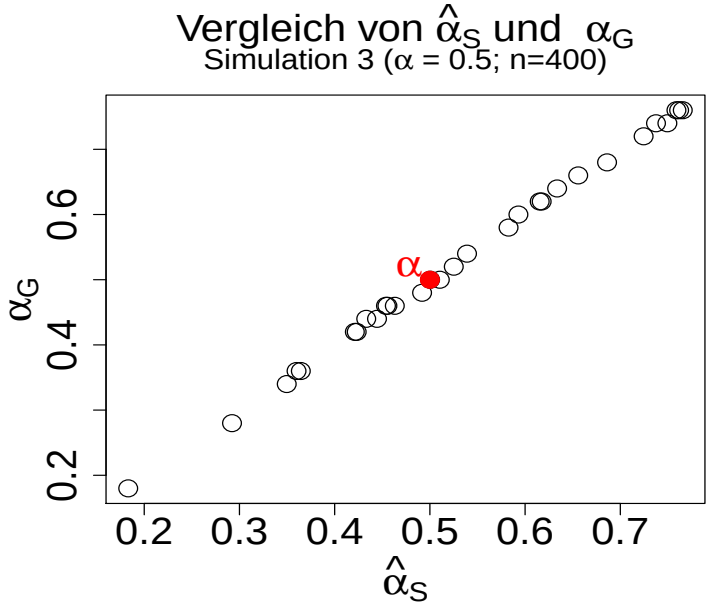

Abbildung A.32: Szenario 3 ($\alpha = 0.5; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

A.4 Szenario 4

Weitere Simulationsergebnisse zu Szenario 4 sind im Folgenden zu finden. Es sind Abbildungen zu den Schätzungen der Koeffizienten γ und β.

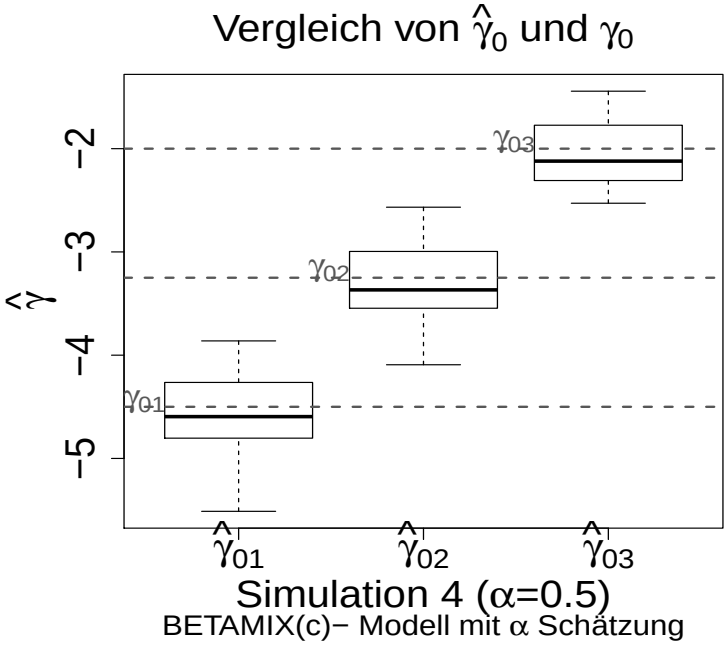

Abbildung A.33: Szenario 4 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

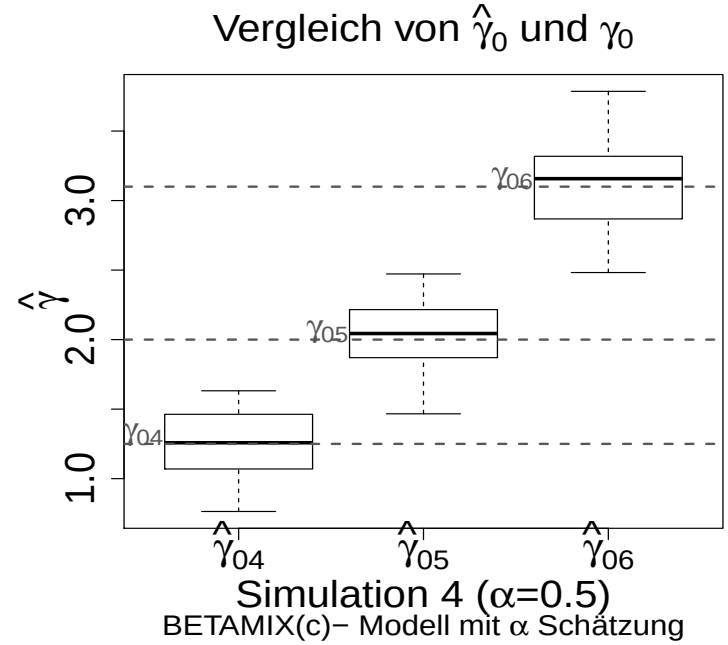

Abbildung A.34: Szenario 4 ($\alpha = 0.5; n = 1000$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

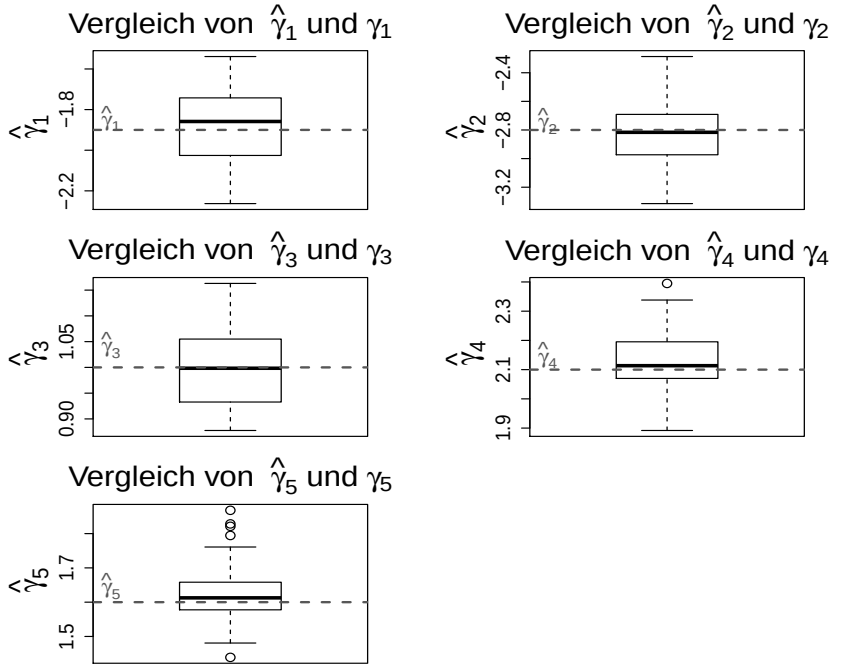

Abbildung A.35: Szenario 4 ($\alpha = 0.5; n = 1000$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

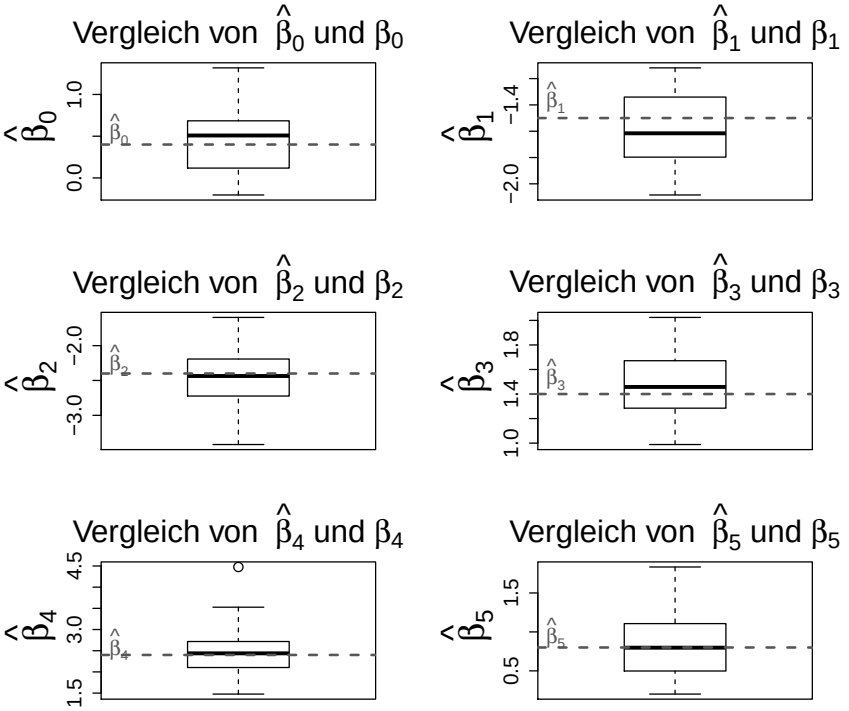

Abbildung A.36: Szenario 4 ($\alpha = 0.5; n = 1000$): Boxplot der Koeffizienten $\beta_0, \ldots, \beta_5$

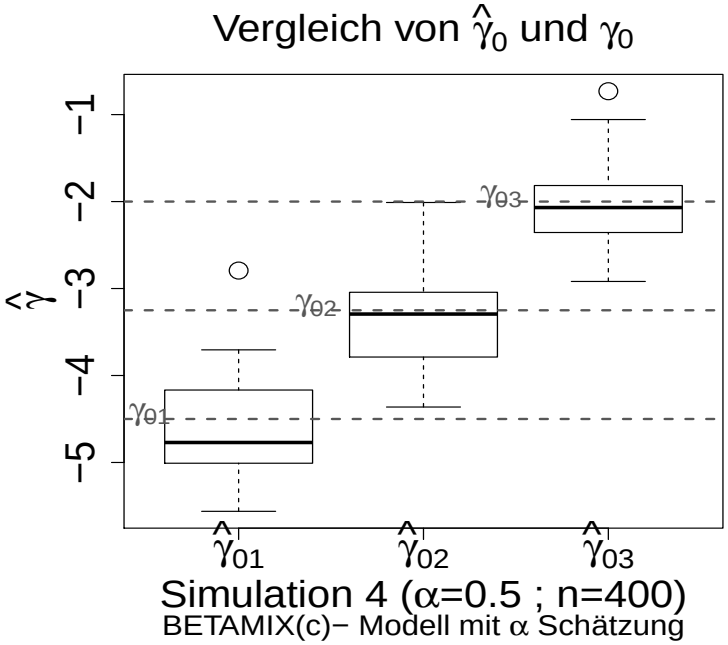

Abbildung A.37: Szenario 4 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{01}, \ldots, \gamma_{03}$

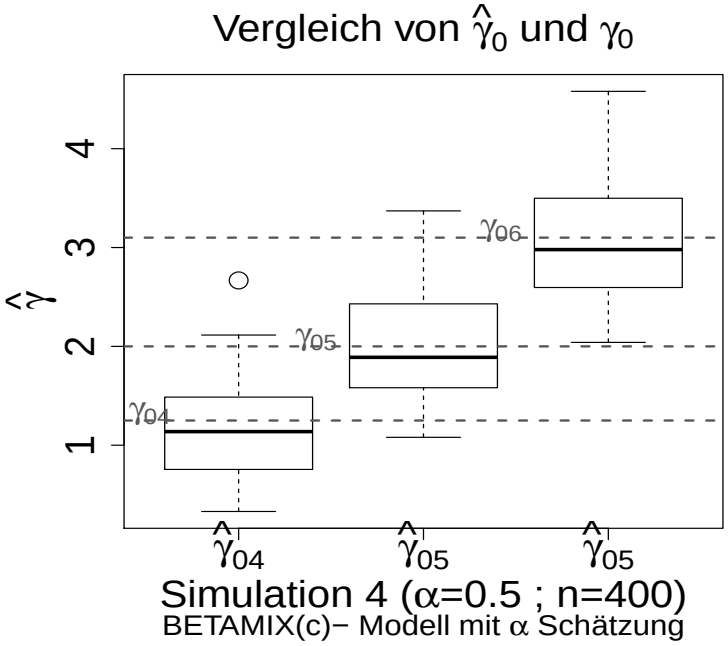

Abbildung A.38: Szenario 4 ($\alpha = 0.5; n = 400$): Boxplot der Intercepts $\gamma_{04}, \ldots, \gamma_{06}$

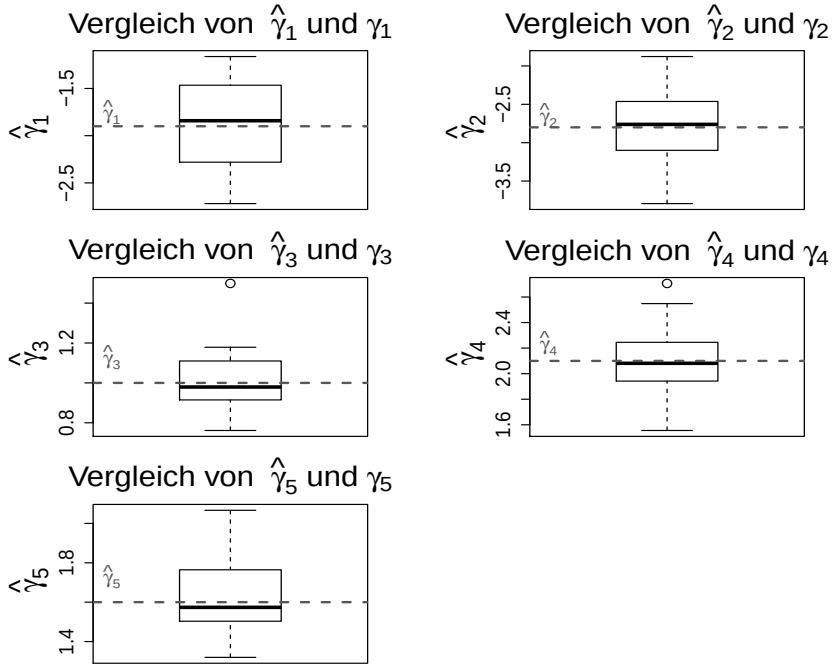

Abbildung A.39: Szenario 4 ($\alpha = 0.5; n = 400$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

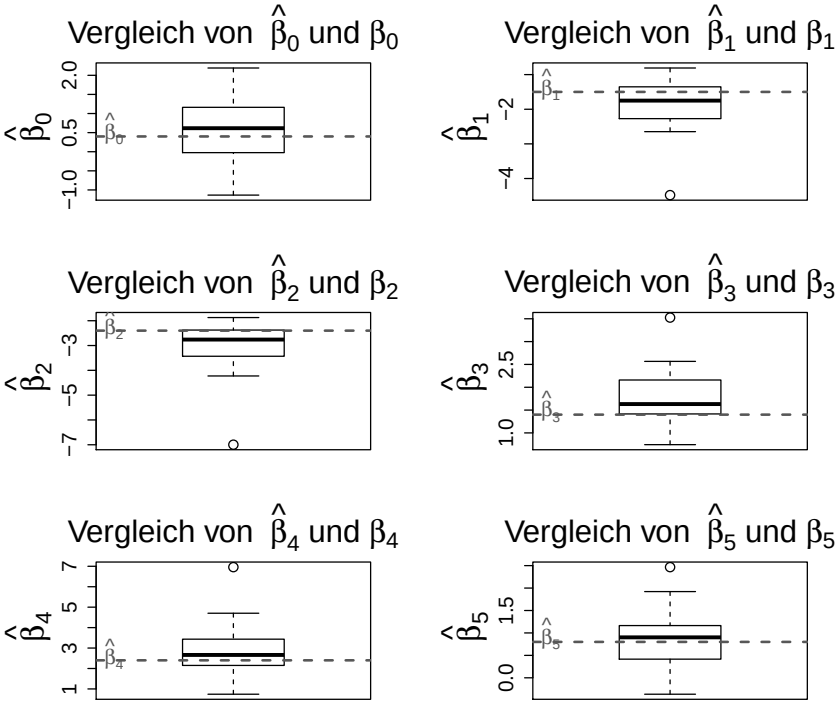

Abbildung A.40: Szenario 4 ($\alpha = 0.5; n = 400$): Boxplot der Koeffizienten $\beta_0, \ldots, \beta_5$

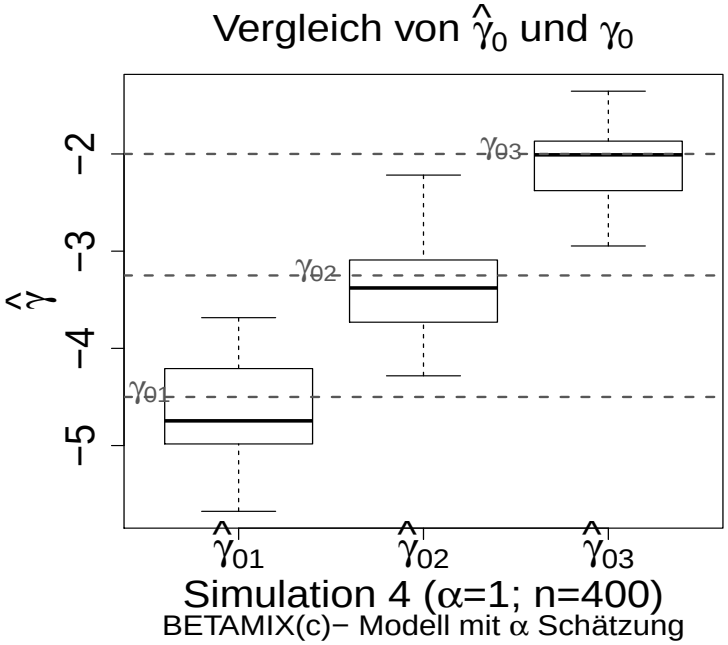

Abbildung A.41: Szenario 4 ($\alpha = 1$; $n = 400$): Boxplot der Intercepts γ_{01},γ_{02}, γ_{03}

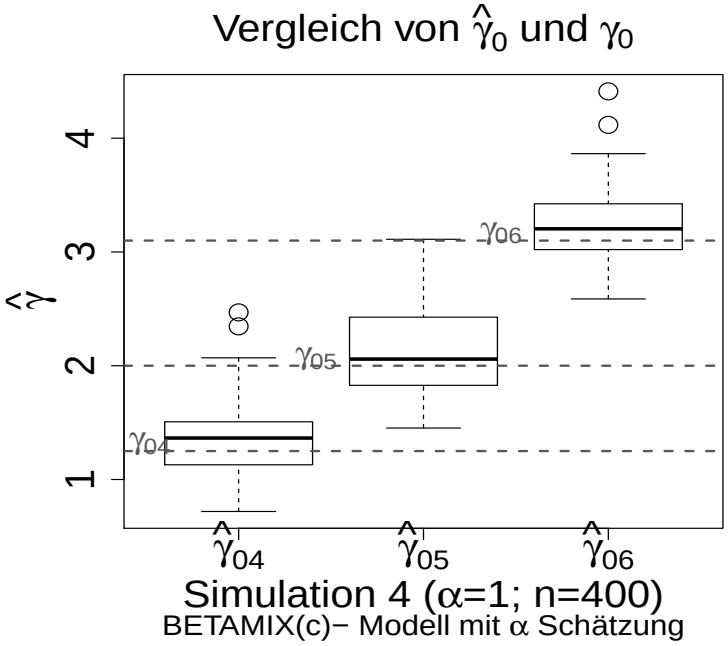

Abbildung A.42: Szenario 4 ($\alpha = 1$; $n = 400$): Boxplot der Intercepts γ_{04},γ_{05}, γ_{06}

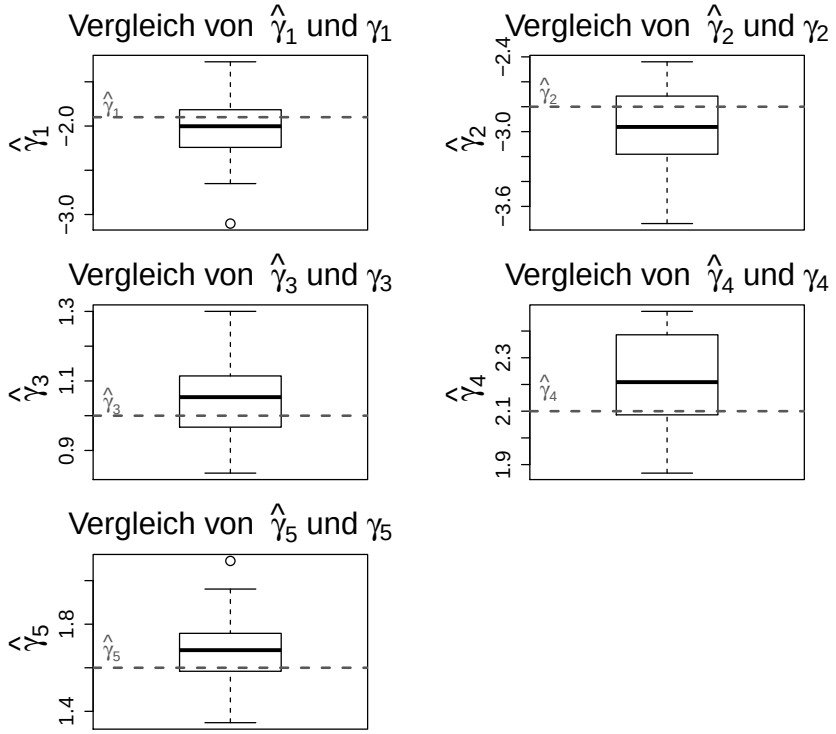

Abbildung A.43: Szenario 4 ($\alpha = 1$; $n = 400$): Boxplot der Koeffizienten $\gamma_1, \ldots, \gamma_5$

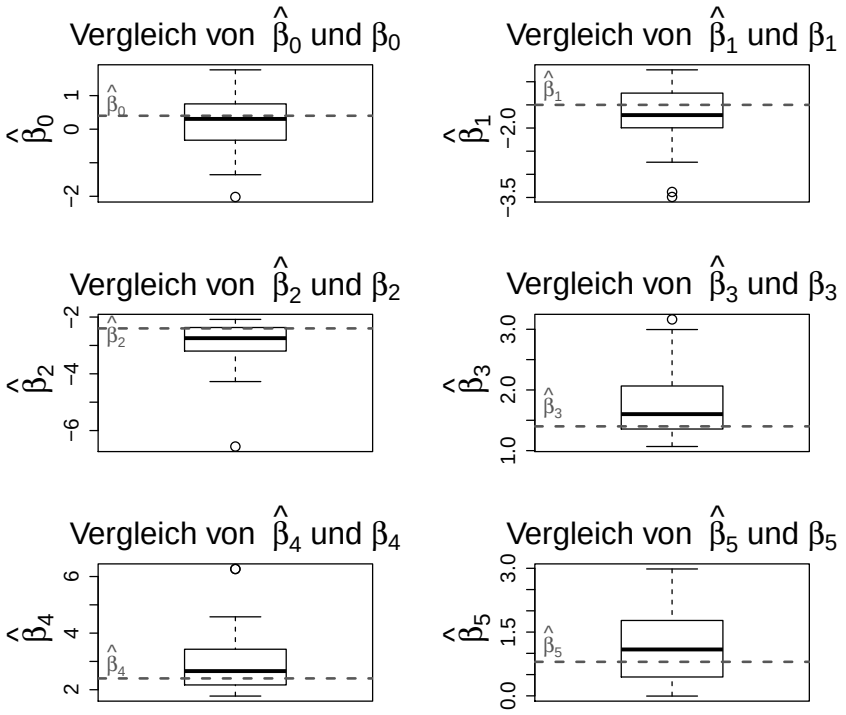

Abbildung A.44: Szenario 4 ($\alpha = 1$; $n = 400$): Boxplot der Koeffizienten $\beta_0, \ldots, \beta_5$

Ein Vergleich der gewählten und geschätzten α's ist in den folgenden Abbildungen zu sehen. Wie auch in den Szenarien zuvor liegen die betrachteten Werte $\hat{\alpha}_S, \alpha_G$ auf der 1. Winkelhalbierenden. Somit lässt sich sagen, dass die Wahl über ein Gitter annähernd den selben Wert wie die Schätzung für den Parameter α erzielt.

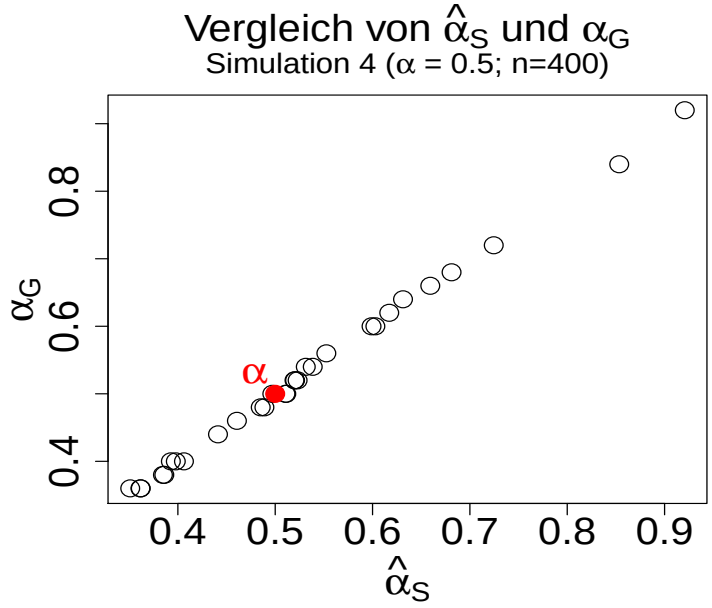

Abbildung A.45: Szenario 4 ($\alpha = 0.5; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

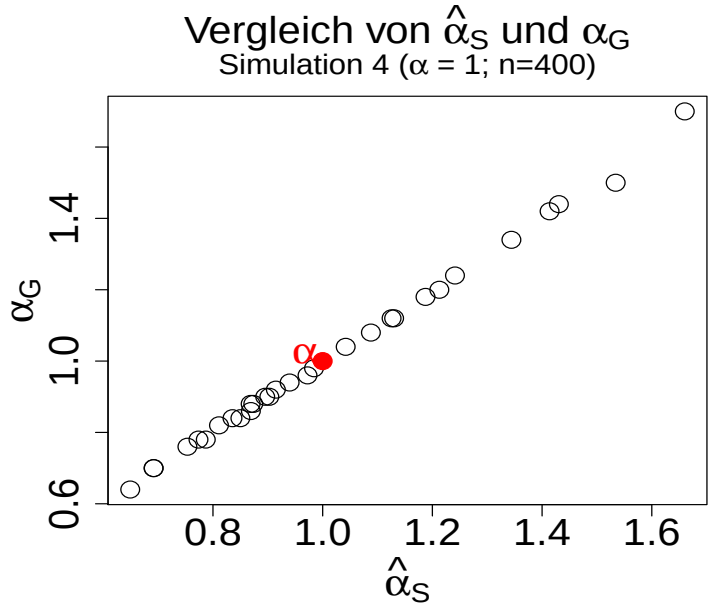

Abbildung A.46: Szenario 4 ($\alpha = 1; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

Für die Abbildungen A.47 und A.48 wird für die Wahl des α's das Gitter $(0.5, 0.7, \ldots, 2.3, 2.5, 2.6, 2.7, \ldots 3.9, 4, 4.5, 5, 6)$ verwendet.

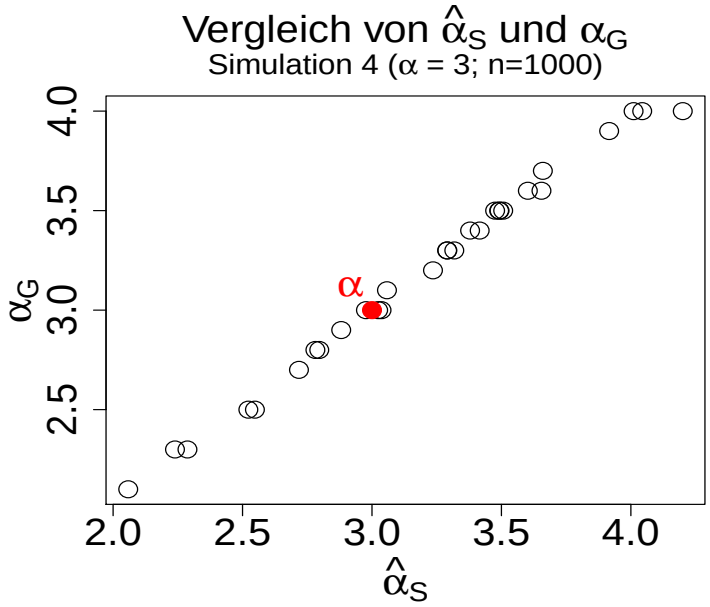

Abbildung A.47: Szenario 4 ($\alpha = 3; n = 1000$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

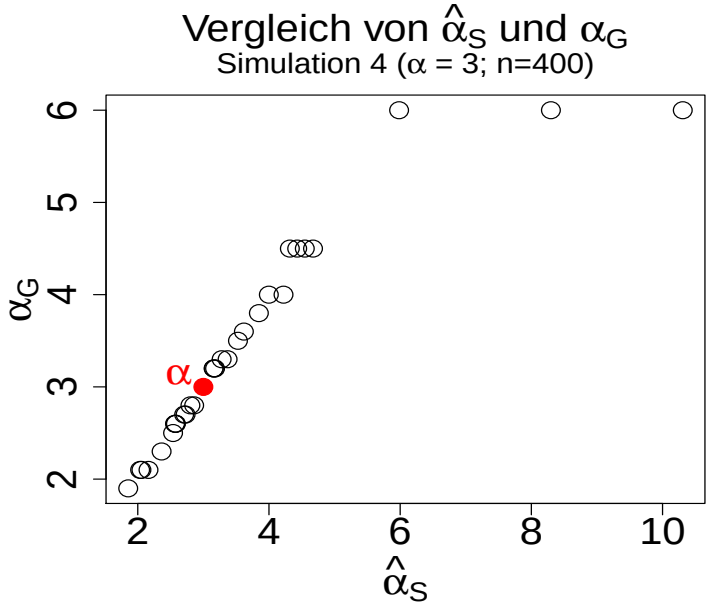

Abbildung A.48: Szenario 4 ($\alpha = 3; n = 400$): Vergleich der geschätzten $\hat{\alpha}_S$ und der gewählten α_G

B Elektronischer Anhang

Die beigefügte CD beinhaltet den vorliegenden Text im .pdf Format und den Dateiordner:

- **R-Code**: Enthält den erzeugten R-Code (Dateien mit der Endung .R) und die in diesem R-Code gespeicherten Ergebnisse (Dateien mit der Endung .RData)

Der Ordner R-Code ist in weitere Unterordner unterteilt. Im Folgenden werden diese Unterordner benannt und kurz beschrieben.

- **Plots** : Beinhaltet die erstellten Grafiken (im .pdf oder .png Format)

Genmix

MRSP_0.6.0.zip	Version 0.6.0 des R- Pakets MRSP, welches für die Lauffähigkeit der Funktionen installiert sein muss
genmix-helpers-2.R, genmix-gesamt-3.R, genmix_functions6.R, ...	Von Micha Schneider zur Verfügung gestellte Funktionen für CUB-, CUP und BETAMIX-Modelle

Anwendung

Ergebnisse_Code	R-Code der Anpassung und Speicherung unterschiedlicher Modelle für die in Kapitel 5 betrachteten Datensätze
Ergebnisse_WS	Enthält die von Ergebnisse_Code erstellten Modelle als .RData-Dateien

Simulation

Ergebnisse_Sim_Code	R-Code zur Erstellung der simulierten Datensätze. Die Szenarien 1-4 und das Szenario, in welchem der Responsestyle ignoriert wird, sind in separaten Daten abgespeichert. Der R-Code für Szenario 5 und 6 ist in einer Datei zu finden.
Ergebnisse_Sim_WS	Enthält die von Ergebnisse_Sim_Code erstellten Datensätze und die geschätzten Parameter als Dateien
Plots_Sim	R-Code für die grafische Darstellung der Simulationsergebnisse und für die Abbildungen der Wahrscheinlinkeitsfunktionen der Betabinomialverteilung (Abbildung 2.1 und 2.2)
Sim_BETAMIX.R	Enthält die Funktionen „sim.genmixT“ und „sample.genmixT“. Die Funktion sim.genmixT generiert einen Datensatz nach dem BETAMIX(c)-Modell-Annahmen. Mit sample.genmixT können mehrere Datensätze generiert werden. Zudem passt diese Funktion für alle generierten Datensätze das zur Datengenerierung verwendete Modell an.
Simulation_Alpha.R	Die Funktion „sim.genmixAlpha“ simuliert Datensätze nach dem BETAMIX(c)-Modell, in welchen der Parameter α mit Kovariablen verknüpft wird

Eidesstattliche Erklärung

Hiermit versichere ich, Cynthia Huber, die vorliegende Arbeit selbständig und lediglich unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst zu haben.

München, 3. März 2016

Cynthia Huber