Fakultät für Mathematik, Informatik und Statistik
Institut für Statistik
Ludwig-Maximilians-Universität München

Penalisierungsansätze in ordinalen Regressionsmodellen

BACHELORARBEIT

zur Erlangung des akademischen Grades eines Bachelor of Science (B. Sc.)

von
David Drießlein

Betreuer: Univ.-Prof. Dr. Gerhard Tutz,
Dipl. Stat. Wolfgang Pößnecker

München, 28.04.2013

Inhaltsverzeichnis

Kapitel 1

Einleitung und Überblick

1.1 Problemstellung

Kategoriale Regressionsmodelle eignen sich, um die Effekte metrischer oder kategorialer Kovariablen auf die Kategorien einer Zielgröße zu modellieren. Unter der Annahme, dass die Zielgrößenkategorien einer Ordnung unterliegen, lassen sich das kumulative und gegebenenfalls das sequentielle Modell anwenden.

Ein Vorteil dieser beiden Modelltypen besteht darin, dass sie mittels einer Erweiterung der Klasse der Generalisierten Linearen Modelle (GLM) auf multivariate Zielgrößen in das GLM-Rahmenwerk eingebunden werden können. Die Koeffizientenschätzer der Modellvariablen lassen sich folglich durch Maximum-Likelihood-Schätzung (ML-Schätzung) bestimmen. In einem allgemeinen Modellansatz ist je Zielgrößenkategorie und je erklärender Variable ein Regressionskoeffizient zu schätzen. Dies ermöglicht zwar eine sehr flexible Modellierung der Effekte der Einflussgrößen, allerdings kann dieser Modellansatz für große Anzahlen der Zielgrößenkategorien oder große Anzahlen an Einflussgrößen sehr schnell zu hochdimensionalen Parametrisierungen führen. Für den Fall, dass die Anzahl n der Beobachtungseinheiten geringer ist, als die Anzahl zu schätzender Parameter p, ist eine ML-Schätzung nicht mehr möglich. Ungeachtet einer Nicht-Existenz von Schätzern, ist es im Allgemeinen von Interesse, aus einer Vielzahl vorhandender Einflussgrößen diejenigen für das Modell selektieren zu können, die die stärksten Effekte aufweisen. Eine Selektion der stärksten Effekte verbessert in diesem Zusammenhang die Interpretierbarkeit des Modells.

Konkrete Ansätze, die sowohl im Fall $p > n$ einen ML-Schätzer bestimmen, als auch implizit eine Selektion von Variablen mit den stärksten Effekten durchführen können, beruhen auf einer Penalisierung der logarithmierten Likelihoodfunktion. Dabei wird der log-Likelihood ein Strafterm hinzugefügt, der auf einer Norm des Koeffizientenvektors beruht, es wird z.B. die Länge dieses Vektors bestraft. Je nach Stärke des Einflusses des Strafterms auf die ML-Schätzung, werden parametersparsame Modelle dadurch erzeugt, dass manche der Regressionskoeffizienten in ihrer Größe geschrumpft werden, andere auf null geschätzt werden, somit implizit eine Variablenselektion durchgeführt wird. Wirkt ein Prädiktor durch einen einzigen Koeffizienten auf eine univariate Zielgröße, entspricht die Selektion dieses Koeffizienten der Selektion dieser Variable für das Modell. Wird der Koeffizient auf null geschätzt, fällt dieser Prädiktor aus dem Modell heraus. Wirkt ein Prä-

diktor durch einen für jede Kategorie der Zielgröße spezifischen Koeffizienten auf diese, genügt es nicht, wenn lediglich einer dieser Koeffizienten auf null geschätzt wird, um diesen Prädiktor aus dem Modell entfernen zu können. Eine Variablenselektion tritt erst dann ein, wenn alle zu einem Prädiktor gehörigen Koeffizienten gleichzeitig auf null geschätzt werden. Erst dann kann diese Variable aus dem Modell entfernt werden.

In dieser Ausarbeitung werden Modelle betrachtet, in denen ein Prädiktor generell durch mehrere Koeffizienten vertreten ist. Dieses Charakteristikum ist für die Konstruktion und Wahl eines geeigneten Penalisierungsterms zu berücksichtigen. Es resultieren parametersparsame Modelle, die aufgrund einer Selektion der stärksten Effekte eine verbesserte Interpretierbarkeit besitzen. Gleichzeitig werden durch die Parameterschrumpfung zwar verzerrte Schätzer erzeugt, diese können allerdings eine geringere Varianz und im Sinne eines Bias-Varianz-Tradeoff einen geringeren MSE als der ML-Schätzer und verbesserte Prädiktionseigenschaften aufweisen.

1.2 Zielsetzung und Aufbau

Der erste Teil dieser Ausarbeitung behandelt in den Kapiteln 2 bis 4 die theoretischen Grundlagen, die die verwendeten Modelle und Penalisierungsansätze umfassend. Die in dieser Ausarbeitung betrachteten Penalisierungsansätze basieren auf einer Penalisierung der log-Likelihoodfunktion. Hierfür wird in Kapitel 2 dargestellt, wie sich ordinale Regressionsmodelle unter Verwendung einer verktorwertigen Responsefunktion und einer multivariaten Zielgrößenverteilung innerhalb der GLM-Klasse formulieren lassen. In diesem Zusammenhang werden die grundlegenden Komponenten Generalisierter Linearer Modelle im Fall univariater Zielgrößen betrachtet, anschließend diese auf den Fall multivariater Zielgrößen erweitert. Sowohl für univariaten, als auch multivariaten Fall, werden die Ansätze der ML-Schätzung aufgestellt.

In Kapitel 3 werden zwei Typen ordinaler Regressionsmodelle beschrieben: Das kumulative und das sequentielle Modell. Für beide Modelltypen wird die Idee, die dem jeweiligen Modellansatz zugrunde liegt, motiviert und die Modelltypen in verschiedenen Varianten skizziert. Der Vergleich dieser beiden Typen dient dazu, die Verwendung des sequentiellen Logit-Modells für den empirischen Teil der Ausarbeitung zu begründen.

Da die Idee zahlreicher Penalisierungsansätze auf Modellen für univariate Zielgrößen basiert, werden in Kapitel 4 mit der *Ridge Regression* und dem *Lasso-Verfahren* zunächst zwei klassiche Penalisierungsansätze vorgestellt. Unter Berücksichtigung der Charakteristika multivariater Modelle, werden mit dem *Group Lasso* und dem *Sparse Group Lasso* Penalisierungen gewählt, die der Verwendung für die Koeffizientenstruktur des allgemeinen sequentiellen Logit-Modells gerecht werden.

Im empirischen Teil dieser Ausarbeitung, der die Kapitel 5 und 6 umfasst, werden in einer Simulationsstudie (Kapitel 5) verschiedene Penalisierungsansätze für das sequentielle Logit-Modell, hinsichtlich der Güte ihrer Schätzer und der Fähigkeit zur Variablenselektion, miteinander verglichen. In Kapitel 6 werden ausgewählte Penalisierungsansätze beispielhaft auf zwei verschiedene Datensätze angewendet. Der Datensatz *Gründerstudie* befasst sich mit kategorisierten Zeitdauern

vom Gründungszeitpunkt eines Unternehmens, bis zu dessen Insolvenz oder Zensierung. Dieses Beispiel ist durch 1224 Beobachtungseinheiten und 14 kategoriale Prädiktoren mit je 2 bis 4 Kategorien charakterisiert. Zudem wird durch dieses Beispiel die Schnittstelle zwischen ordinalen Regressionsmodellen und Modellen zur Analyse von Lebensdauern deutlich. Der Datensatz *Gleason-Score* behandelt den Einfluss genetischer Disposition auf Prostatakarzinome. Dieses Beispiel ist mit einer, im Verhältnis zu 52 vorhandenen Beobachtungen, hohen Anzahl von fast 250 metrischen Einflussgrößen charakterisiert.

Abschließend werden in Kapitel 7 die zentralen Aspekte der Arbeit, Ergebnisse der Simulationsauswertungen und Datensatzanalysen zusammengefasst.

Teil I

Theoretische Grundlagen

Kapitel 2

Generalisierte Lineare Modelle

Innerhalb dieses Kapitels wird die Modellklasse der Generalisierten Linearen Modelle (GLM) dargestellt. Zunächst wird in Abschnitt 2.1 die allgemeine Struktur dieser Modellklasse, die aus einer stochastischen und einer strukturellen Komponente besteht, beschrieben. Es erfolgt eine Unterteilung in Modelle mit univariaten und mit multivariaten Zielgrößen, die dementsprechend univariate bzw. multivariate GLM bezeichnet werden. Innerhalb der univariaten GLM werden in Abschnitt 2.2 grundlegende Modelle für stetige und diskrete univariate Zielgrößenvariable aufgelistet. Der Schwerpunkt dabei, liegt entsprechend der Themenstellung auf diskreten Zielgrößen. Die Maximum-Likelihood-Schätzung, als grundlegendes Schätzkonzept für GLM, wird in Abschnitt 2.3 beschrieben. Um vollständig auf Modelle für mehrkategoriale Zielgrößen zugreifen zu können, wird in Abschnitt 2.4 diese Modellklasse auf multivariate Zielgrößen erweitert. Literarische Quellen dieser Darstellungen bilden das Kapitel 4 aus Tutz (2012) und Kapitel 3 aus Fahrmeir & Tutz (2001).

2.1 Struktur generalisierter Regressionsmodelle

Gegeben sei eine Datensituation $(y_i, \mathbf{x}_i)$ für $i = 1, \ldots, n$ Beobachtungen. Dabei bezeichne $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ den p-dimensionalen Vektor der Einflussgrößen (synonym Kovariablen, Prädiktoren genannt) $j = 1, \ldots, p$ und y_i die univariate Zielgröße (Responsevariable) für Beobachtungseinheit i.[1] Die Idee des klassischen linearen Regressionsmodells besteht darin, den Einfluss diverser erklärender Variablen $(x_{i1}, \ldots, x_{ip})$ auf den (bedingten) Erwartungswert einer metrischen Zielgröße y_i, mit Hilfe einer Funktion $f(x_{i1}, \ldots, x_{ip})$ zu modellieren. Diese Funktion sei eine Linearkombination der erklärenden Variablen:

$$\begin{aligned} \mathbb{E}(y_i|x_{i1}, \ldots, x_{ip}) &= f(x_{i1}, \ldots, x_{ip}) \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ip} \end{aligned} \tag{2.1}$$

[1] Zur Bezeichnung der Transponierten eines Vektors oder einer Matrix wird $()'$ verwendet.

Der Vektor der Regressionskoeffizienten $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)' \in \mathbb{R}^{p+1}$, inklusive Intercept β_0, dieser Funktion sei unbekannt und werde mit Hilfe beobachteter Daten der n Beobachtungseinheiten geschätzt. Charakteristisch für die klassische Regression ist, dass die Einflussvariablen direkt mit dem Erwartungswert der Zielgröße verknüpft sind und, dass im Rahmen der Normalregression für die, auf die Kovariablen bedingte, Verteilung der Zielgröße eine Normalverteilung angenommen wird. (Vgl. Fahrmeir et al. (2007), S. 60 ff.)

Eine Verallgemeinerung der Idee des klassischen linearen Regressionsmodells bildet die Modellklasse der Generalisierten Linearen Modelle. Diese basiert auf Nelder & Wedderburn (1972) und beschreibt ein Regressionsmodell anhand einer strukturellen und einer stochastischen Komponente.

Die **stochastische Komponente** (random component) bestimmt für die, gegeben die erklärenden Variablen $\mathbf{x}_i$ (bedingt) unabhängigen Beobachtungen y_i eine Wahrscheinlichkeitsdichte aus einer einfachen Exponentialfamilie. Eine derartige Dichtefunktion hat die allgemeine Form:

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right\} \tag{2.2}$$

Dabei bezeichnet θ_i den natürlichen Parameter der Exponentialfamilie, ϕ_i einen Skalenparameter, $b(\cdot)$ und $c(\cdot)$ charakteristische Funktionen der jeweiligen Verteilung. Verteilungen, die sich in dieser Form darstellen lassen sind z.B. die Binomialverteilung, die Poissionverteilung und die Normalverteilung.

Die **systematische Komponente** beinhaltet zwei strukturelle Spezifikationen. Zum einen die Struktur der erklärenden Variablen, zum anderen, wie diese auf den bedingten Erwartungswert der Zielgröße wirken. Die Linearkombination

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i'\boldsymbol{\beta}$$

wird als linearer Prädiktor bezeichnet und legt fest, dass die erklärenden Variablen x_j linear über ihre Koeffizienten β_j in das Regressionsmodell eingehen.

Die zweite strukturelle Spezifikation gibt an, mittels welcher bekannten, streng monoton steigenden, stetig differenzierbaren Funktion $h(\cdot)$ dieser lineare Prädiktor η_i mit dem bedingten Erwartungswert der Zielgröße $\mu_i = \mathbb{E}(y_i|\mathbf{x}_i)$ verknüpft wird:

$$\mu_i = h(\eta_i) \quad \Leftrightarrow \quad g(\mu_i) = \eta_i$$

Die Funktion $h(\cdot)$ trägt die Bezeichnung Responsefunktion und transformiert den linearen Prädiktor, die Funktion $g(\cdot) = h(\cdot)^{-1}$ wird als Linkfunktion bezeichnet und ist die Umkehrfunktion zu $h(\cdot)$. Diese gibt an, mit welcher Funktion μ_i transformiert wird, um den linearen Modellteil zu erhalten. Der Spezialfall, dass $h(\cdot)$ die Identitätsfunktion ist, führt zu einer direkten Verknüfung von Prädiktor und Erwartungswert der Zielgröße, wie sie im klassischen linearen Regressionsmodell zu finden ist.

2.2 GLM für stetige und diskrete univariate Responsevariablen

Die Klasse der GLM lässt sich entsprechend der Verteilung der Responsevariable in Modelle für stetigen und für diskreten Response untergliedern.

Für **stetige Responsevariablen** lassen sich neben der bereits genannten Normalverteilung beispielsweise die Exponentialverteilung, die Gamma-Verteilung oder die inverse Gauss-Verteiltung als Dichtefunktionen verwenden, die sich jeweils in der Form einer einfachen Exponentialfamilie parametrisieren lassen. Ohne Berücksichtigung einer konkreten Linkfunktion orientiert sich die Wahl der Responsedichte daran, ob bspw. nichtnegative Zielgrößen modelliert werden sollen (z.B. Exponentialverteilung) oder wie flexibel die Dichte der Zielgröße sein soll (z.B. Gamma-Verteilung). Da Modelle mit stetigem Response nicht Kern dieser Ausarbeitung sind, wird für eine vertiefte Darstellung auf Tutz (2012), Seite 53 ff. verwiesen.

Eine **diskrete Responsevariable** liegt dann vor, wenn die Zielgrößenvariable endlich oder abzählbar unendlich viele Ausprägungen annimmt. Dies kann z.B. in Form einer Zählvariable sein, für die die Zielgröße die Anzahl spezifischer Ereignisse (Versicherungsfälle, Arztbesuche in einem gegebenen Zeitraum) wiederspiegelt. Um Zählvariablen in das Rahmenwerk univariater GLM einzubinden, dienen bspw. die Poisson-Verteilung oder die Negative Binomialverteilung. Erstere zeichnet sich durch ihre Einfachheit und intuitive Interpretierbarkeit in der Verwendung einfacher Zähldaten aus. Zweitere ermöglicht flexible Modellierungsmöglichkeiten für Zähldaten mit der Fragestellung, wieviele Versuche bis zu einer gegebenen Anzahl von Erfolgen notwendig sind, sowie die Berücksichtigung von Dispersionsproblemen. (Vgl. Tutz (2012), Seite 56 ff.)

Eine weitere Möglichkeit diskreten Responses liegt in Form einer kategorialen Variable vor, die in ihrem einfachsten binären Fall entweder das Eintreten oder Nicht-Eintreten eines Ereignisses kodiert, z.B. ob in einem Haushalt ein Auto vorhanden ist oder nicht. Dabei dienen synonym die Begriffe Erfolg bzw. Misserfolg der Dichotomisierung einer binären Zielgröße. Da die Modellierung einer binären kategorialen Zielgröße grundlegend für das in Abschnitt 3.2 behandelte sequentielle Modell und dessen ML-Schätzung ist, wird dessen Modellierung ausführlicher dargestellt. Die binäre Zielgröße y_i nehme die beiden Ausprägungen 0 oder 1 an. Modelliert werde die Wahrscheinlichkeit eines Erfolgs $\pi_i = P(y_i = 1|\mathbf{x}_i)$. Die Bernoulli-Verteilung lässt sich für $y_i \in \{0, 1\}$ in der Form einer einfachen Exponentialfamilie darstellen mit der Wahrscheinlichksfunktion:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right\} \tag{2.3}$$

Unter Verwendung ihrer kanonischen Linkfunktion erhält man wegen $\theta(\pi_i) = \log(\pi_i/(1-\pi_i))$ das **binäre Logit-Modell**:

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}; \quad g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.4}$$

Ein Vorteil der Verwendung dieser kanonischen Linkfunktion, liegt in der Interpretierbarkeit der Regressionskoeffizienten: β_j lässt sich damit als additiver

Effekt, einer um eine Einheit veränderten Kovariable x_j, auf das erwartete logarithmierte Chancenverhältnis zwischen Erfolg $\pi_i = P(y_i = 1|\mathbf{x}_i)$ und Misserfolg $1 - \pi_i = 1 - P(y_i = 1|\mathbf{x}_i) = P(y_i = 0|\mathbf{x}_i)$ interpretieren, $\exp(\beta_j)$ als der multiplikative Effekt auf dieses Chancenverhältnis, unter der Bedingung, dass alle anderen Kovariablen unverändert bleiben..

Die binäre Modellierung ist nicht auf die Verwendung ihrer kanonischen Linkfunktion beschränkt. Es lässt sich jede streng monotone, steigende Verteilungsfunktion F als Verknüpfung zwischen π_i und dem linearen Prädiktor verwenden, sodass $\pi_i = F(\eta_i)$.

Da sich im Rahmen univariater GLM keine Zielgrößen mit mehr als zwei Kategorien modellieren lassen, wird in Abschnitt 2.3 eine Erweiterung auf mulivariate Responsevektoren vorgenommen.

2.3 Erweiterung der Modellklasse auf multivariate Responsevariablen

Die Darstellung der Modellklasse der GLM im vorangegangenen Abschnitt beschränkt sich auf Modelle mit univariater stetiger oder diskreter Responsevariable und einer Dichtefunktion aus einer einfachen univariaten Exponentialfamilie. Die einfachste Form einer kategorialen Zielgröße mit zwei möglichen Ausprägungen - einem binären Response - kann ebenfalls in diesen Rahmen eingebettet werden. Eine Verallgemeinerung der Modellklasse wird notwendig, sobald die Zielgröße als Realisation einer von mehr als zwei Kategorien auftreten kann. In diesem Fall lässt sich die Zielgröße nicht mehr wie ein univariater Response behandeln. Es wird notwendig für jede der Kategorien eine Dummyvariable einzuführen, wodurch eine multivariate Responsevariable resultiert.

Ziel dieses Abschnitts ist es, Modellformulierungen auch für mehr als zwei Kategorien aufzustellen. Der Struktur eines GLM entsprechend, wird dafür eine stochastische Komponente, d.h. eine multivariate Verteilung, die sich als (multivariate einfache) Exponentialfamilie parametrisieren lässt und eine strukturelle Komponente, d.h. eine vektorwertige Link- und Responsefunktion, benötigt. (Vgl. Fahrmeir & Tutz (2001), S. 69 ff.)

Datensituation

Die Datensituation verändert sich im Vergleich zum univariaten GLM nicht. Der Kovariablenvektor $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ beinhaltet die Werte der p Einflussgrößen für Beobachtung $i = 1, \ldots, n$ und die Variable $Y_i \in \{1, \ldots, k\}$ den Kategorieindex einer der k möglichen Kategorien, in die Beobachtung i fällt.

Der grundlegende Unterschied zum univariaten GLM besteht darin, dass die Responsevariable Y_i mit Hilfe einer Dummykodierung in eine vektorwertige Darstellung überführt wird. Mittels einer 0-1-Kodierung resultiert ein k-dimensionaler Responsevektor $\mathbf{y}_i = (y_{i1}, ..., y_{ik})'$, dessen r-ter Eintrag $(r = 1, \ldots, k)$ den Wert 1 annimmt, sofern die Zielgröße Y_i in diese Kategorie fällt und den Wert 0, falls die Zielgröße nicht in Kategorie r fällt:[2]

[2]Für die Notation wird festgelegt, dass die für Beobachtungseinheit i beobachtete Responsekategorie durch einen Großbuchstaben Y_i gekennzeichnet ist, der davon abgeleitete Responsevektor durch einen fettgedruckten Kleinbuchstaben $\mathbf{y}_i$.

$$y_{ir} = \begin{cases} 1 & \text{falls } Y_i = r; \quad r = 1, \ldots, k \\ 0 & \text{sonst} \end{cases}$$

Diese Darstellung lässt sich auf einen $q = (k-1)$ - dimensionalen Vektor $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})'$ reduzieren, sofern eine der k Kategorien als Referenzkategorie[3] gewählt wird. Fällt die Zielgröße in die Referenzkategorie c resultiert ein q-dimensionaler Nullvektor. Zwecks einer klaren Notation, dient im Folgenden die k-te Kategorie als Referenzkategorie. Für eine Beobachtung i, deren Zielgröße sich in Kategorie r $\in \{1, \ldots, q\}$ - einer anderen als der Referenzkategorie - realisiert, ergibt sich der Responsevektor mit einer 1 an r-ter Stelle:

$$Y_i = r \quad \Leftrightarrow \quad \mathbf{y}_i = (0, \ldots, 0, 1, 0, \ldots, 0); \quad r = 1, \ldots, q$$

Die verkürzte Darstellung mittels Referenzkategorie hat den Sinn, bei der Schätzung der Regressionskoeffizienten einem etwaigen Identifizierbarkeitsproblem entgegenzuwirken. Die Beobachtungseinheiten lassen sich in folgender Form als Matrizen darstellen:

$$\begin{pmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \begin{pmatrix} y_{11} & \ldots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \ldots & y_{nq} \end{pmatrix}; \quad \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{2.5}$$

Im Rahmen eines kategorialen Regressionsmodells ist von Interesse, die Wahrscheinlichkeit π_{ir} zu bestimmen, mit der eine Beobachtung für gegebene Einflussgrößen in eine der q Kategorien fällt:

$$\pi_{ir} = P(Y_i = r|\mathbf{x}_i) = P(y_{ir} = 1|\mathbf{x}_i); \quad r = 1, \ldots, q$$

Für die Basiskategorie c ergibt sich die Wahrscheinlichkeit $\pi_{ic} = P(Y_i = c|\mathbf{x}_i)$ als $1 - \sum_{r=1}^{q} \pi_{ir}$. Die zu bestimmenden q Wahrscheinlichkeiten lassen sich ebenfalls wie die Responsevariable in einem q-dimensionalen Vektor $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ darstellen.

Stochastische Komponente

Aufgrund dessen, dass der Response $\mathbf{y}_i$ ein q-dimensionaler Vektor ist, gilt dies auch für den bedingten Erwartungswert $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)$. Zur Bestimmung des Erwartungswertvektors dient als Verteilungsannahme für das multivariate GLM die Multinomialverteilung, die eine natürliche Verallgemeinerung der Binomialverteilung auf den mehrdimensionalen Fall ist.

Für eine einzelne Beobachtung i, mit Responsevektor $\mathbf{y}_i$ und dem Vektor der kategoriespezifischen Auftretenswahrscheinlichkeiten $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$, ist die Dichtefunktion $f(\mathbf{y}_i|\boldsymbol{\pi}_i)$ gegeben als:

$$f(\mathbf{y}_i|\boldsymbol{\pi}_i) = \pi_{i1}^{y_{i1}} \cdot \ldots \cdot \pi_{iq}^{y_{iq}} (1 - \pi_{i1} - \ldots - \pi_{iq})^{1 - y_{i1} - \ldots - y_{iq}} \tag{2.6}$$

[3]Als Referenzkategorie kann für ungeordnete Kategorien eine Beliebige gewählt werden, für geordnete Kategorien ermöglicht eine Wahl der ersten oder der letzten Kategorie eine sinnvolle Modellinterpretation.

Somit folgt die (gegeben eines Kovariablenvektors $\mathbf{x}_i$ bedingte) Verteilung des Responsevektors $\mathbf{y}_i$ einer Multinomialverteilung:

$$\mathbf{y}_i|\mathbf{x}_i \sim M(1, \boldsymbol{\pi}_i), \quad \boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})' \tag{2.7}$$

Die Intention und allgemeine Form der Multinomialverteilung ist in Anhang A.1 auf Seite 74 dargestellt. Für den q-dimensionalen Erwartungswertvektor $\boldsymbol{\mu}_i$ und die $q \times q$-dimensionale Varianz-Kovarianzmatrix $Cov(\mathbf{y}_i)$ des Responsevektors $\mathbf{y}_i$, erhält man:

$$\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{y}_i|\mathbf{x}_i) = \begin{pmatrix} \pi_{i1} \\ \vdots \\ \pi_{iq} \end{pmatrix} = \boldsymbol{\pi}_i; \quad Cov(\mathbf{y}_i) = \begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & \cdots & -\pi_{i1}\pi_{iq} \\ \vdots & \ddots & \vdots \\ -\pi_{iq}\pi_{i1} & \cdots & \pi_{iq}(1-\pi_{iq}) \end{pmatrix} \tag{2.8}$$

Anhand ihrer Darstellbarkeit als (einfache multivariate) Exponentialfamilie lässt sich die Multinomialverteilung $M(1, \boldsymbol{\pi}_i)$ in den GLM-Rahmen einfügen.

Strukturelle Komponente

Um Unterschiede im Einfluss der erklärenden Variablen auf verschiedene Kategorien bestimmen zu können, werden im multivariaten Fall kategoriespezifische lineare Prädiktoren η_{ir}, $r = 1, \ldots, q$ verwendet. Zur Verknüpfung des q-dimensionalen bedingten Erwartungswertvektors $\boldsymbol{\pi}_i (= \boldsymbol{\mu}_i)$ mit den kategoriespezifischen Prädiktoren ist eine q-dimensionale vektorwertige Responsefunktion $h(\boldsymbol{\eta}_i) = (h_1(\boldsymbol{\eta}_i), \ldots, h_q(\boldsymbol{\eta}_i))' : \mathbb{R}^q \to \mathbb{R}^q$ erforderlich. Dabei bezeichne $h_r(\boldsymbol{\eta}_i) = h_r(\eta_{i1}, \ldots, \eta_{iq})$, $r = 1, \ldots, q$ eine Funktion in Abhängigkeit der kategoriespezifischen Prädiktoren.

Die Linkfunktion $\mathbf{g}$ als Umkehrfunktion der Responsefunktion h ist ebenfalls eine q-dimensionale vektorwertige Funktion $g(\boldsymbol{\pi}_i) = (g_1(\boldsymbol{\pi}_i), \ldots, g_q(\boldsymbol{\pi}_i))' : \mathbb{R}^q \to \mathbb{R}^q$, mit $g_r(\boldsymbol{\pi}_i) = g_r(\pi_{i1}, \ldots, \pi_{iq})$. Eine konkrete Darstellung findet im Kontext des sequentiellen Logit-Modells in Abschnitt 3.2 statt.

Geordnete und ungeordnete Responsekategorien

In der bisherigen Darstellung wurde es noch nicht notwendig, auf die Responsekategorien genauer einzugehen. Um einen adäquaten Modelltyp für eine kategoriale Zielgröße zu bestimmen, wird eine Unterscheidung der Zielgröße hinsichtlich ihres Skalenniveaus vorgenommen.[4] Für kategoriale Daten existieren zwei Skalenniveaus: Die Zielgröße kann in geordneten Kategorien auftreten - sie ist ordinalskaliert - oder in ungeordneten Kategorien - sie ist nominal-skaliert. Beispiele für geordnete Kategorien sind ein empfundenes Schmerzlevel (keine, geringe, starke Schmerzen) oder die Häufigkeit von Arztbesuchen in einem Zeitraum (gar kein, seltener, regelmäßiger Besuch). Als Beispiel für ungeordnete Kategorien dient die Religionszugehörigkeit (katholisch, evangelisch, muslimisch, etc.) oder politische Parteipräferenz (CDU/CSU, SPD, Die Linke, Bündnis 90 - die Grünen, FDP). (Vgl. Agresti (2007), S. 2 f.)

[4]Eine ausführliche Darstellung der von Stevens (1946) vorgeschlagenen Skalenniveaueinteilung findet sich in Fahrmeir et al. (2007), S. 17 ff.

Für die Modellierung ungeordneter Responsekategorien eignet sich das multinomiale Modell. Mit der Wahl der logistischen Funktion als Responsefunktion, erhält man das multinomiale Logit-Modell, das als Verallgemeinerung des binären Logit-Modells aufgefasst werden kann. Für die Interpretation der Regressionskoeffizienten wird die Wahrscheinlichkeit für das Eintreten einer beliebigen Kategorie ins Verhältnis zum Eintreten einer (vorher gewählten) Referenzkategorie gesetzt. Die Interpretation eines Koeffizienten ist dabei äquivalent zu der des binären Logit-Modells. Das multinomiale Modell wird im Rahmen dieser Ausarbeitung nicht weiter betrachtet, da es sich nur eingeschränkt für die Modellierung geordneter Responsekategorien eignet. Die Berücksichtigung einer etwaigen Ordnungsstruktur im Modell hat den Vorteil einer weniger parameterintensive Modellierung. Das Nutzen der zusätzlichen Informationen aus der Ordnungsstruktur der Responsekategorien ermöglicht es, bei der Modellschätzung weniger Parameter bestimmen zu müssen, als wenn diese Kategorien wie ungeordnete im Modell behandelt werden. Dies ist dann günstig, wenn im Vergleich zu den möglichen Parametern relativ wenige Beobachtungen für die kategorialen Daten vorliegen. (Vgl. Fahrmeir & Tutz (2001), S. 81.)

Die beiden Modelltypen, die sich für geordnete Responsekategorien eignen, sind das kumulative und das sequentielle Modell. Beide Typen werden ausführlich im folgenden Kapitel dargestellt und ihre Verwendung gegenüber einander und gegenüber nominalen und metrischen Regressionsmodellen abgegrenzt.

2.4 Maximum-Likelihood Parameterschätzung

Sowohl für univariate, als auch für multivariate GLM ist das am häufigsten verwendete Schätzkonzept die Maximum-Likelihood Methode. Der Vorteil in der Verwendung der ML-Schätzung liegt in der Existenz einer allgemeinen Darstellung für Likelihood- und Scorefunktion. Diese hat ihre Grundlage darin, dass die bedingte Verteilung $f(y_i|\theta_i, \phi_i)$ der Zielgröße aus einer einfachen (multivariaten) Exponentialfamilie stammt. Entscheidend ist die Bestimmung des Erwartungwerts und der Varianz mit Hilfe des kanonischen Parameters θ_i:

$$\mu_i = \mathbb{E}(y_i) = \frac{\delta b(\theta_i)}{\delta \theta}; \quad \sigma_i^2 = \mathbb{V}(y_i) = \phi_i \frac{\delta^2(\theta_i)}{\delta \theta^2} \tag{2.9}$$

Zunächst wird die ML-Schätzung für univariate GLM betrachtet, anschließend die auf vektorwertige Funktionen erweiterte Schätzung für multivariate GLM.

2.4.1 ML-Schätzung für univariate GLM

Die Likelihoodfunktion $L(\boldsymbol{\beta})$ ist eine Funktion des unbekannten Vektors der Regressionsparameter in Abhängigkeit der gegebenen Daten. Diese lässt sich als Produkt der Dichten $f(y_i|\theta_i, \phi_i)$ bestimmen, da die Beobachtungen y_i als (bedingt) unabhängig angenommen werden können:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi_i) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \right\} \tag{2.10}$$

Der natürliche Parameter θ_i als eine Funktion des Erwartungswerts ist über die Responsefunktion mit dem linearen Prädiktor bzw. den Koeffizienten β_j verknüpft, d.h. $\theta_i = \theta(\mu_i)$ und $\mu_i = h(\eta_i) = h(\mathbf{x}_i'\boldsymbol{\beta})$, sodass $\theta_i = \theta(h(\mathbf{x}_i'\boldsymbol{\beta}))$. Zur Bestimmung der Koeffizientenschätzer, für die die Likelihoodfunktion das Maximum annimmt, wird zwecks einfacherer Berechenbarkeit die Likelihoodfunktion logarithmiert:

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \right\} \tag{2.11}$$

Aufgrund der Monotonie der Logarithmusfunktion bleibt das Maximum der Likelihoodfunktion erhalten. Um das Maximum zu bestimmen, wird die Scorefunktion $s(\boldsymbol{\beta})$ als erste Ableitung der log-Likelihood gebildet. Unter oben genannten Verknüpfungen ist die Scorefunktion gegeben als:

$$\begin{aligned} s(\boldsymbol{\beta}) = \frac{\delta l(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\delta l_i(\theta_i)}{\delta \theta} \frac{\delta \theta(\mu_i)}{\delta \mu} \frac{\delta h(\eta_i)}{\delta \eta} \frac{\delta \eta_i}{\delta \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i} \frac{\phi_i}{var(y_i)} \frac{\delta h(\eta_i)}{\delta \eta} \mathbf{x}_i \\ &= \sum_{i=1}^{n} \mathbf{x}_i \frac{\delta h(\eta_i)}{\delta \eta} \frac{y_i - \mu_i}{\phi_i v(\mu_i)} \end{aligned}$$

Dabei wird genutzt, dass $\phi_i v(\mu_i) = \mathbb{V}(y_i)$. Die Schätzgleichungen $s(\widehat{\boldsymbol{\beta}}) \stackrel{!}{=} \mathbf{0}$ haben die Form:

$$\sum_{i=1}^{n} \mathbf{x}_i \frac{\delta h(\eta_i)}{\delta \eta} \frac{y_i - \mu_i}{\phi_i v(\mu_i)} \stackrel{!}{=} \mathbf{0} \tag{2.12}$$

Unter Verwendung der kanonischen Linkfunktion, die den natürlichen Parameter direkt mit dem linearen Prädiktor - $\theta_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ - verknüpft, vereinfacht sich die Scorefunktion zu:

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \frac{(y_i - \mu_i)}{\phi_i} \tag{2.13}$$

Die numerische Bestimmung des ML-Schätzers $\hat{\boldsymbol{\beta}}$ aus Gleichung 2.12 erfolgt mittels iterativer Prozeduren, da für gewöhnlich der Schätzer in keiner geschlossenen Form angegeben werden kann. Iterative Methoden zur Lösung dieser nichtlinearen Gleichungen sind der Newton-Raphson-Algorithmus oder der Fisher-Scoring-Algorithmus. Ausgehend von einem Startwert für den Schätzvektor erfolgt eine sukzessive Verbesserung der geschätzten Lösung, bis die Veränderung des Schätzers zwischen zwei aufeinanderfolgenden Schritten kleiner ist, als eine vorgegebene Schranke.

Die asymptotische Varianz-Kovarianz-Matrix des Koeffizientenvektors $\boldsymbol{\beta}$ ergibt sich aus der erwarteten Fisher-Informationsmatrix $\mathbf{F}(\boldsymbol{\beta})$ als Erwartungswert der

beobachteten Fisherinformation $\mathbf{F}_{obs}(\boldsymbol{\beta})$:

$$\begin{aligned}\mathbf{F}(\boldsymbol{\beta}) &= \mathbb{E}[\mathbf{F}_{obs}(\boldsymbol{\beta})] = \mathbb{E}\left[-\frac{\delta^2 l(\boldsymbol{\beta})}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}'}\right] \\ &= \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' \left(\frac{\delta h(\eta_i)}{\delta\eta}\right)^2 / \sigma_i^2\end{aligned}$$

Durch Matrixnotation wird eine kompaktere Darstellung erreicht, die vor allem im multivariaten GLM hilfreich ist. Die Scorefunktion im univariaten GLM ergibt sich als $s(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{D}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$. Dabei bezeichnet $\mathbf{X}' = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ die Designmatrix der Einflussgrößen, $\mathbf{D} = \text{Diag}(\delta h(\eta_1)/\delta\eta, \ldots, \delta h(\eta_n)/\delta\eta)$ eine Diagonalmatrix der abgeleiteten linearen Prädiktoren und $\boldsymbol{\Sigma}^{-1} = \text{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$ die Varianzmatrix. $\mathbf{y} = (y_1, \ldots, y_n)'$ und $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ bezeichnen die Vektoren der Zielgrößen und Erwartungswerte. Eine Kombination von $\mathbf{D}$ und $\boldsymbol{\Sigma}$ in der Gewichtsmatrix $\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}'$ erzielt für die Scorefunkton $s(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ und für die erwartete Fishermatrix $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}$.

Ausgewertet an der Stelle des ML-Schätzers $\hat{\boldsymbol{\beta}}$, gibt die inverse Fisher-Informationsmatrix $\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})$ die asymptotische Varianz-Kovarianzmatrix des ML-Schätzers an. Unter Regularitätsbedingungen existiert ein eindeutiger und konsistenter ML-Schätzer, dessen Verteilung asymptotisch ($n \to \infty$) durch die einer Normalverteilung approximiert werden kann:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})) \tag{2.14}$$

Für eine ausführliche Darstellung des Schätz- und Inferenzkonzepts, sowie zur Bestimmung des ML-Schätzers mittels iterativer Algorithmen für univariate GLM wird auf Tutz (2012), Kapitel 3.6 ff. und Fahrmeir et al. (2009), Kapitel 4.4 verwiesen.

2.4.2 ML-Schätzung für multivariate GLM

Die Koeffizientenschätzung erfolgt für multivariate GLM, ebenso wie für die univariaten Modell, nach der Maximum-Likelihood Methode. Die Likelihoodfunktion ergibt sich als Produkt der als (bedingt) unabhängig angenommenen Dichten der Beobachtungen $\mathbf{y}_i$, für die eine Multinomialverteilung $M(1, \boldsymbol{\pi}_i)$ angenommen wurde:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\pi}_i) \tag{2.15}$$

Mit Hilfe der Matrixnotation lässt sich für alle kategorialen Regressionsmodelle das ML-Konzept einheitlich darstellen. Der Vektor der kategoriespezifischen Auftretenswahrscheinlichkeiten $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ ist dabei je nach Modell über die q-dimensionale Responsefunktion $h(\boldsymbol{\eta}_i) = (h_1(\boldsymbol{\eta}_i), \ldots, h_q(\boldsymbol{\eta}_i))'$ mit dem Vektor $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}$ der kategoriespezifischen linearen Prädiktoren verknüft. Dabei bezeichne $\mathbf{X}_i$ eine, in Abhängigkeit des konkreten Modells gestaltete, individuenspezifische Designmatrix und $\boldsymbol{\beta}$ den Vektor aller Regressionskoeffizienten. Erweitert auf den q-dimensionalen Fall hat dann die Scorefunktion $\mathbf{s}(\boldsymbol{\beta})$, als Vektor der ersten Ableitungen der logarithmierten Likelihoodfunktion nach den Koeffizienten,

ähnlich wie im univariaten Fall die Form:

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{D}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\pi}_i) \tag{2.16}$$

Dabei bezeichne $\mathbf{D}_i = \delta \mathbf{h}(\boldsymbol{\eta_i}) / \delta \boldsymbol{\eta}$ die Matrix der partiellen Ableitungen an der Stelle $\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta}$, $\boldsymbol{\Sigma}_i$ die Varianz-Kovarianzmatrix der Multinomialverteilung aus Gleichung 2.8 auf Seite 10. Die numerische Bestimmung des ML-Schätzers im multivariaten Fall, findet ebenfalls über iterative Prozeduren wie Fisher-Scoring statt.

Mit Hilfe der Gewichtsmatrix $\mathbf{W}_i = \mathbf{D}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i'$ ergibt sich die erwartete Fishermatrix $\mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i$. Analog zum univariaten Fall ist der eindeutige bestimmte konsistente ML-Schätzer $\hat{\boldsymbol{\beta}}$ asymptotisch normalverteilt mit:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})) \tag{2.17}$$

Ein Vergleich der matrixnotationellen Darstellungen zeigt die Ähnlichkeit der ML-Schätzung im univariaten und im multivariaten Fall, die auf die Annahme einer Exponentialfamilie für uni- wie multivariate Zielgrößenverteilungen im GLM zurückzuführen ist. Die Komplexität der einzelnen Komponenten wie der Designmatrix, des Koeffizientenvektors, der vektorwertigen Responsefunktion des multivariaten Falls wird deutlich, wenn diese im folgenden Kapitel, konkret für das sequentielle Logit-Modell, betrachtet werden.

2.5 Zusammenfassung

In diesem Kapitel wurde die Grundstruktur generalisierter linearer Modelle, die sich aus einer stochastischen und einer strukturellen Komponente zusammensetzt, dargestellt. Diese Modellklasse ermöglicht die Modellierung des Einflusses diverser Prädiktoren auf eine univariate Zielgröße. Dabei wird angenommen, dass die Zielgrößenverteilung einer einfachen Exponentialfamilie entstammt und der lineare Prädiktor mit dem Erwartungswert der Zielgröße über die sogenannte Responsefunktion verknüpft ist. Eine Erweiterung dieser Modellklasse auf multivariate Responsevariablen (nominale und ordinale Variablen) ist möglich, wenn für die multivariate Zielgröße eine Verteilung aus einer multivariaten Exponentialfamilie angenommen wird, wie die Multinomialverteilung. Zur Verknüpfung der kategoriespezifischen linearen Prädiktoren mit dem Erwartungswertvektor der Zielgröße, sind vektorwertige Funktionen zu wählen. Die Einbindung mehrkategorialer Responsevariablen in das GLM-Rahmenwerk erlaubt eine Maximum-Likelihood-Schätzung der Modellkoeffizienten, mit Hilfe der im Text beschriebenen Scorefunktion. Die Schätzung der Koeffizienten erfolgt über iterative Algorithmen.

Kapitel 3

Ordinale Regressionsmodelle

Innerhalb dieses Kapitels werden das kumulative und das sequentielle Modell als kategoriale Regressionsmodelle für geordnete Responsekategorien dargestellt. Zunächst werden in Abschnitt 3.1 ordinale Regressionsmodelle gegenüber nominalen und metrischen Regressionsmodellen abgegrenzt. In Abschnitt 3.1 wird das von McCullagh (1980) vorgeschlagene kumulative Modell, dessen Motivation anhand eines Schwellenwertansatzes,der Modellsansatz, sowie spezielle Modellvarianten erläutert. In Abschnitt 3.2 wird anhand selbigen Aufbaus das sequentielle Modell beschrieben. Dieses Modell wird in den multivariaten GLM-Rahmen eingebunden und diesbezüglich die Maximum Likelihood-Schätzung skizziert. In Abschnitt 3.4 wird das sequentielle Modelle mit zeitdiskreten Survivalmodellen verknüpft. Abschließend werden beide ordinalen Regressionsmodelle einander gegenübergestellt und erläutert, warum sich in dieser Arbeit auf das sequentielle Logit-Modell für Penalisierungsüberlegungen beschränkt wird. Da vornehmlich das sequentielle Modell im Fokus dieser Arbeit steht, wird dieses ausführlicher behandelt. Um Redundanzen in den Modellbeschreibungen des kumulativen und des sequentiellen Modells zu reduzieren, wird an gegebener Stelle auf Parallelitäten der beiden Modelle hingewiesen. Dieses Kapitel orientiert sich in seiner Darstellung an Fahrmeir & Tutz (2001), Kapitel 3.3 und Tutz (2012), Kapitel 9.

3.1 Abgrenzung ordinaler Regressionsmodelle

Ordinale Regressionsmodelle können mit Hilfe der den Kategorien zugrundeliegenden Ordnungsstruktur von nominalen und metrischen Regressionsmodellen abgegrenzt werden.

Kategorien mit ordinaler Skala lassen sich durch eine Ordnungsrelation miteinander vergleichen, indem alle Kategorien geordnet werden, sodass eine Kategorie als größer/kleiner eingestuft werden kann, als eine andere Kategorie. Für kategoriale Merkmale mit nominaler Skala ist eine Ordnung irrelevant. Aufgrund dieser schwächeren Annahme für nominale Kategorien, lassen sich nominale Modelle auch für ordinale Kategorien anwenden. Wie bereits in Abschnitt 2.3 erläutert wird allerdings zusätzliche Information, die sich aus der Ordnungsstruktur ergibt,

unbeachtet gelassen. Dadurch lässt sich mit zu schätzenden Parametern weniger ökonomisch umgehen, als dies bei expliziter Berücksichtigung der Ordnungsstruktur im Modellansatz möglich wäre. Zudem geht der genannte Informationsverlust mit einem Verlust an Aussagekraft einher. Aufgrund einer Berücksichtigung der Ordnungsstruktur lassen sich ordinale Modelle nicht auf nominale Zielgrößen anwenden, da das ordinale Skalenniveau eine strengere Anforderung an die kategoriale Variable impliziert. Sowohl der nominalen, als auch der ordinalen Skala ist gemeinsam, dass Abstände zwischen Kategorien nicht sinnvoll interpretiert werden können. Dies bleibt auch dann gültig, wenn die Kategorien in eine Ganzzahl-Kodierung transformiert werden. (Vgl. Agresti (2007), S. 2 f.)

Die Abgrenzung zu metrischen Regressionsmodellen basiert auf deren Annahme einer quantitativen Zielgröße, die auf einer Intervall- oder Verhältnisskala gemessen wird. Die stärkeren Annahmen hinsichtlich der Skala erlauben es im Allgemeinen nicht, mit Hilfe metrischer Modelle, den Einfluss der Kovariablen auf eine kategoriale Zielgröße darzustellen. Kategoriale Daten erfüllen gewöhnlich nicht die Annahmen für Fehler und Zielgrößenverteilungen eines metrischen Regressionsmodells. Für eine hinreichend große Anzahl an Responsekategorien lassen sich gegebenenfalls, aufgrund von Einfachheit und Schätzbarkeit, dennoch metrische Modelle verwenden. (Vgl. Tutz (2000), S. 208 f.)

Anderson (1984) unterscheidet mit gruppiert-stetigen (*grouped continuous*) und durch ordinale Beurteilung erlangten Variablen (*assessed ordered*) zwei Haupttypen kategorial-ordinaler Variablen:

Gruppiert-stetige Variablen werden generiert, indem eine zugrundeliegende stetige Variable, zwecks einer gröberen Klassifizierung, in Intervalle eingeteilt wird. Beispiele hierfür sind Einkommensklassen oder Arbeitslosigkeitsdauern (kurz-, mittel-, langfristig). Aus diesen Beispielen wird bereits ersichtlich, dass die Intervalle einer Ordnung folgen, aber nicht notwendigerweise gleich breit gewählt werden müssen. Die letzte Kategorie wird zumeist durch ein nach oben offenes Intervall gebildet, um etwaige Extremwerte einzubeziehen. Mit der Zeitdauer bis zur Insolvenz eines Unternehmens, wird eine gruppiert stetige Variable in Abschnitt 6.2 als kategoriale Zielgröße verwendet.

Kategorial-ordinale Merkmale, die aus **Beurteilungen** resultieren, treten häufig in Befragungen auf, in denen die befragte Person einen Sachverhalt auf einer gegebenen Skala verschiedener Ausprägungsgrade einstufen soll. Beispiele hierfür sind die Stärke von Schmerzen (kein, gering, stark), der Grad einer Behinderung (Skala von 20 bis 100 in 10er-Schritten) sowie der in Abschnitt 6.1 als Zielgröße verwendete Gleason-Score zur Beurteilung von Prostatakrebs. Ein Erklärungsansatz für diesen Variablentyp ist es, die kategorial-ordinalen Variable als eine, durch eine Beurteilung gewonnene, Realisation einer zugrundeliegenden unbeobachteten stetigen Variable aufzufassen.

Um diese Variablentypen als Zielgröße für ordinale Regressionsmodelle zu verwenden, eignen sich das im Folgenden dargestellte kumulative und sequentielle Modell. Letzteres unter der einschränkenden Annahme, wenn die Kategorien nur sukzessive erreicht werden können.

3.2 Das kumulative Modell

3.2.1 Motivation und Modellansatz

Das kumulative Modell ist das am häufigsten verwendete Modell für kategorial-ordinale Zielgrößen. Der Grund dafür liegt in seiner Einfachheit, sowie intuitiven Interpretierbarkeit der Regressionskoeffizienten. Es wurde von McCullagh (1980) aus dem *Proportional Hazards-* und dem *Proportional Odds-*Modell als multivariate Erweiterung von generalisierten linearen Modellen abgeleitet.

Die kategoriale Zielgröße $Y_i \in \{1, \ldots, k\}$ trete in k geordneten Kategorien auf. Der Modellansatz lässt sich dadurch motivieren, dass diese beobachtete Zufallsvariable Y_i, in Abhängigkeit eines Kovariablenvektors, die Realisation einer unbeobachteten stetigen Zufallsvariable $\widetilde{Y}_i$ ist. Die Verknüpfung der beobachteten und der latenten Variable folgt einem Schwellenwertmechanismus, der die Zielgröße genau dann Kategorie $r = 1, \ldots, k$ zuordnet, wenn die latente Variable zwischen zwei Schwellenwerten θ_{r-1} und θ_r ihres stetigen Wertebereichs liegt:

$$Y_i = r \quad \Leftrightarrow \quad \theta_{r-1} \; < \; \widetilde{Y}_i \; \leq \; \theta_r; \quad r = 1, \ldots, k \tag{3.1}$$

Die latente Variable wird durch die erklärenden Variablen in linearer Form bestimmt:

$$\widetilde{Y}_i = -\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

Dabei bezeichnet $\boldsymbol{\beta}$ den Koeffizientenvektor der erklärenden Variablen und ϵ_i einen Fehlerterm mit Verteilungsfunktion F. Das Minus vor dem Prädiktor dient weiteren rechnerischen Zwecken, kann aber auch in den Koeffizientenvektor integriert werden. Der Einfluss der erklärenden Variablen auf die latente metrische Zielgröße wirkt sich in einer Verschiebung dieser auf dem latenten Kontinuum aus, auf dem die Schwellenwerte $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$ angeordnet sind.

Die interessierende Wahrscheinlichkeit $P(Y_i = r|\mathbf{x}_i)$, dass Beobachtung i in Kategorie r fällt, lässt sich anhand der Schwellenwerte bestimmen als:

$$\begin{aligned} P(Y_i = r|\mathbf{x}_i) &= P(\theta_{r-1} < \widetilde{Y}_i \leq \theta_r) \\ &= P(\theta_{r-1} < -\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \leq \theta_r) \\ &= P(\theta_{r-1} + \mathbf{x}_i'\boldsymbol{\beta} < \epsilon_i \leq \theta_r + \mathbf{x}_i'\boldsymbol{\beta}) \\ &= F(\theta_r + \mathbf{x}_i'\boldsymbol{\beta}) - F(\theta_{r-1} + \mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

Es sei $\beta_{r0} = \theta_r$ die Parametrisierung des kategoriespezifischen Intercepts.[1] Die Wahrscheinlichkeit, dass Beobachtung i höchstens in Kategorie r fällt, bestimmt sich anhand:

$$P(Y_i \leq r|\mathbf{x}_i) = \sum_{s=1}^{r} P(Y_i = s|\mathbf{x}_i) = F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}); \quad r = 1, \ldots, k \tag{3.2}$$

Aus dieser Darstellung lässt sich der Begriff des „kumulativen“ Modells erschließen, da die kategoriespezifischen Wahrscheinlichkeiten bishin zu Kategorie r aufsummiert werden.

[1] Zwecks einer konsistenten Notation wird festgelegt, dass im Subskript eines Regressionskoeffizienten zuerst die Kategorie angezeigt wird, anschließend der zur j-ten Kovariable gehörige Effekt. Der zu Kategorie r gehörende Intercept besitzt das Subskript „r0“.

3.2.2 Modellvarianten

Je nach Wahl der Verteilungsfunktion F in Gleichung 3.2 ergeben sich verschiedene Varianten des kumulativen Modells. Aufgrund einer intuitiven und einfachen Interpretierbarkeit der geschätzten Modellkoeffizienten, fällt die häufigste Wahl der Verteilungsfunktion des Fehlerterms ϵ_i auf die logistische Verteilung $F(\epsilon_i) = \exp(\epsilon_i)/(1 + \exp(\epsilon_i))$. Dies führt zum sogenannten kumulativen Logit-Modell:

$$P(Y_i \leq r|\mathbf{x}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})} \quad \Leftrightarrow \quad \log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}\right) = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta} \tag{3.3}$$

Die Interpretation der Regressionsparameter ergibt sich für eine Individuum i aus dem Verhältnis der Wahrscheinlichkeiten in eine Kategorie kleinergleich der r-ten zu fallen, anstatt in eine Kategorie größer als der r-ten. Dieses Verhältnis wird als die kumulierten Chancen bezeichnet. Konkret bedeutet dies, dass für eine Veränderung der metrischen Kovariable x_j um eine Einheit, sich das eben genannte erwartete Chancenverhältnis multiplikativ um den Faktor $\exp(\beta_j)$ verändert, unter der Bedingung, dass alle übrigen Kovariablen gleich bleiben. In diesem Modell mit kategorieunspezifischen Steigungsparametern gilt diese Interpretation über alle Kategorien $r = 1, \ldots, k$ hinweg. Eine weitere Auffälligkeit besteht darin, dass das Verhältnis der Chancen bzgl. Kategorie r zum Verhältnis der Chancen einer anderen Kategorie s, unabhängig von Einflussgrößen ist:

$$\log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)/P(Y_i > r|\mathbf{x}_i)}{P(Y_i \leq s|\mathbf{x}_i)/P(Y_i > s|\mathbf{x}_i)}\right) = \log\left(\frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})}{\exp(\beta_{s0} + \mathbf{x}_i'\boldsymbol{\beta})}\right) = \beta_{r0} - \beta_{s0}$$

Da die kumulierten Chancen proportional zueinander und unabhängig von Einflussgrößen sind, wird das kummulative Modell auch als *Proportional-Odds-Model* bezeichnet. Anzumerken ist, dass, je nach Art der kategorialen Zielgröße, sich die Interpretation auf „Chancen“ oder „Risiken“ bezieht.

Andere Varianten des kumulativen Modells ergeben sich anhand der Wahl der Verteilungsfunktion F, z.B. kumulative Extremwertmodelle oder das kumulative Probit-Modell:

- für $F(\epsilon_i) = 1 - \exp(-\exp(\epsilon_i))$, der Minimum-Extremwert- /Gompertz-Verteilung, das kumulative Minimum-Extremwert-Modell, auch *Proportional-Hazards-Model* genannt:

  $P(Y_i \leq r|\mathbf{x}_i) = 1 - \exp(-\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}))$

- für $F(\epsilon_i) = \exp(-\exp(\epsilon_i))$, Maximum-Extremwert- /Gumbel-Verteilung, das kumulative Maximum-Extremwert-Modell:

  $P(Y_i \leq r|\mathbf{x}_i) = \exp(-\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}))$

- für $F(\epsilon_i) = \Phi(\epsilon_i)$, der Standardnormalverteilung, das kumulative Probit-Modell:

  $P(Y_i \leq r|\mathbf{x}_i) = \Phi(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})$

Dabei lassen sich zwar die beiden Extremwert-Verteilung über die Transformation der gompertzverteilten Zielgröße Y_i zu der gumbelverteilten Zielgröße $Y_r =$

$k + 1 - Y$ verknüpfen, wodurch die Ordnung der Kategorien umgekehrt wird, allerdings besteht zwischen den beiden Modellen kein Zusammenhang hinsichtlich ihrer Parameter, da beide Verteilungen nicht symmetrisch sind.

3.2.3 Verallgemeinertes kumulatives Modell

Bisher wurde angenommen, dass ausschließlich die Intercepts kategoriespezifisch bestimmt werden, sodass die Wahrscheinlichkeit in eine Kategorie zu fallen durch die Lage des Schwellenwertes auf dem latenten Kontinuum bestimmt wird. Eine Veränderung des linearen Prädiktors hatte eine Verschiebung der Zielgrößenverteilung auf diesem Kontinuum zur Folge.

In einer allgemeinen Form lässt sich das kumulative Modell auf kategoriespezifische Kovariableneffekte erweitern. Dies begründet sich damit, dass der Effekt einer Kovariable über die Kategorien hinweg variieren kann. Eine Kovariable x_j wirkt nun mit einem eigenen Koeffizienten für jede Kategorie auf die Zielgröße. Somit wird diese Kovariable durch q Koeffizienten repräsentiert. Das verallgemeinerte kumulative Modell besitzt die Darstellung:

$$P(Y_i \leq r|\mathbf{x}_i) = F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r), \quad r = 1, \ldots, q \tag{3.4}$$

Dabei bezeichnet $\boldsymbol{\beta}_r = (\beta_{r1}, \ldots, \beta_{rp})'$ den Vektor der kategoriespezifischen Effekte für Kategorie r.

Die einfache Herleitung anhand eines Schwellenwertansatzes ist hier zu modifizieren, da die Wahrscheinlichkeit, dass eine Beobachtung in eine Kategorie fällt, nicht mehr allein durch den Schwellenwert bestimmt wird. Eine Möglichkeit diese Herleitung zu erweitern besteht darin, die latente Variable auf die Störgröße zu reduzieren: $\widetilde{Y}_i = \epsilon_i$. Es wird im Weiteren angenommen, dass der lineare Prädiktor auf den Schwellenwert selbst in der linearen Form $\theta_r = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r$ wirkt, woraus o.g. Modell resultiert. Um zu garantieren, dass $P(Y_i \leq r - 1|\mathbf{x}_i) \leq P(Y_i \leq r|\mathbf{x}_i)$ gilt, muss die Bedingung $\beta_{r-1,0} + \mathbf{x}_i'\boldsymbol{\beta}_{r-1} \leq \beta_{r,0} + \mathbf{x}_i'\boldsymbol{\beta}_r$, $\forall r, \forall \mathbf{x}_i$ erfüllt sein.

Eine zweite Möglichkeit, das verallgemeinerte kumulative Modell zu motivieren, liegt in einer dichotomen Betrachtungsweise des Kategorienspektrums. Hierzu werden die Responsekategorien in zwei Gruppen $\{1, \ldots, r\}, \{r + 1, \ldots, k\}$ aufgespalten. Für diese $k - 1$ binären Splits werden voneinander abhängige binäre Regressionen, mit jeweils spezifischen Parametern, angenommen.

Eine Variante dieser Verallgemeinerung ist das verallgemeinerte kumulative Logit-Modell:

$$P(Y_i \leq r|\mathbf{x}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r)}{1 + \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r)} \quad \Leftrightarrow \quad \log\left(\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}\right) = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r \tag{3.5}$$

Die Modellparameter lassen sich äquivalent zum binären Logit-Modell interpretieren. Da im Allgemeinen $\beta_{rj} \neq \beta_{sj}, \forall r \neq s, \forall j = 1, \ldots, p$, ist die Effektinterpretation von x_j für jede Kategorie spezifisch.

Eine Eigenschaft dieses Modells ist die *Kollabierbarkeit* über Kategorien. Dies bedeutet, dass die Werte der geschätzten Parameter erhalten bleiben, auch wenn Kategorien zusammengefasst werden. Beispielsweise bleiben die Parameter die

selben, wenn statt der Kategorisierung $1, \ldots, k$, die ersten beiden Kategorien zusammengefasst werden, sodass $\{1, 2\}, 3, \ldots, k$. Diese Eigenschaft basiert auf der Dichotomisierbarkeit des Kategorienspektrums.

3.3 Das sequentielle Modell

3.3.1 Modellzweck und Motivation

Das sequentielle Modell ist ein weiteres multikategoriales Regressionsmodell für geordnete Responsekategorien $Y_i \in \{1, ..., k\}$. Im Unterschied zum kumulativen Modell wird bei der Modellierung der ordinalen Struktur ausgenutzt, dass die Kategorien - sofern eine inhaltliche Interpretation der Variable dies erlaubt - nur sukzessive erreicht werden können. Modelliert wird das Ende eines Prozesses, der in seinem Verlauf alle vorhergehenden Kategorien durchschritten hat. Das Ende dieses Prozesses realisiert sich in der letztlich beobachteten Kategorie. Zur Illustration dieses Prozesses werde angenommen, dass die beobachtete kategoriale Zielgröße die maximale Dauer der Arbeitslosigkeit eines Individuums in Monaten widerspiegelt. Ein Individuum kann beispielsweise nur dann drei Monate arbeitslos sein, wenn es zuvor bereits einen und zwei Monate arbeistslos war, sozusagen diese beiden Kategorien durchschritten hat.

Von Interesse ist demnach die Wahrscheinlichkeit $P(Y_i = r | Y_i \geq r, \mathbf{x}_i)$, $r = 1, \ldots, k$, dass ein bestimmter Prozess für eine Beobachtungseinheit $i = 1, ..., n$ in Kategorie r endet, vorher allerdings die Kategorien $1, \ldots, r-1$ durchlaufen hat. Diese Wahrscheinlichkeit ist abhängig von dem Vektor $\mathbf{x}_i$, der diverse Einflussgrößen für Beobachtungeinheit i enthält.

Die Idee des sequentiellen Modells, lässt sich ebenfalls durch die Annahme latenter Variablen motivieren. Diese metrischen latenten Variablen U_r, $r = 1, \ldots, q$ stehen hinter dem sukzessiven Prozess und initieren den Übergang zur jeweils nächst höheren Kategorie. Dieser Übergang findet genau dann statt, wenn die latente Variable einen Schwellenwert θ_r eines zugrunde liegenden stetigen Wertebereichs überschreitet. Im Gegensatz zu den Schwellenwerten des kumulativen Modells, wird für die Schwellenwerte des sequentiellen Modells keine Ordnung benötigt. Dieser schrittweise Prozess wird nachfolgend beschrieben. Hierfür wird angenommen, dass die latente metrische Variable sich in linearer Form $U_r = -\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_r$ darstellt. Dabei sei ϵ_r eine unabhängige Störgröße mit Verteilungsfunktion F, $\mathbf{x}_i$ ein Einflussgrößenvektor und $\boldsymbol{\beta}$ der Koeffizientenvektor.

Um die beobachtete Variable Y_i mit den latenten Variablen zu verknüpfen, wird ein sequentieller Mechanismus betrachtet, der Y_i dann eine Kategorie r zuordnet, sofern die latente Variable unterhalb eines Schwellenwerts verbleibt, vorausgesetzt, die vorangegangenen $r-1$ Kategorien wurden bereits erreicht. Der Mechanismus startet in der ersten Kategorie mit der binären Entscheidung, ob Y_i in dieser verbleibt oder einer höheren Kategorie zugeordnet wird:

$$Y_i = 1 \quad \Leftrightarrow \quad U_1 \leq \theta_1 \qquad \text{oder} \qquad Y_i > 1 \quad \Leftrightarrow \quad U_1 > \theta_1$$

Unter der Voraussetzung, dass der Prozess nicht in Kategorie eins geendet hat, also Y_i mindestens die zweite Kategorie erreicht ($Y_i \geq 2$), ergibt sich erneut die

binäre Entscheidung, ob der Prozess in Kategorie zwei endet oder weiter andauert:

$$Y_i = 2|Y_i \geq 2 \quad \Leftrightarrow \quad U_2 \leq \theta_2 \qquad \text{oder} \qquad Y_i > 2|Y_i \geq 2 \quad \Leftrightarrow \quad U_2 > \theta_2$$

Der Prozess binärer Übergänge setzt sich solange fort, bis ein Schwellenwert θ_r nicht mehr überschritten wird, gegeben, dass alle vorherigen Schwellen überschritten wurden. In einer allgemeinen Notation lässt sich dies formulieren als:

$$Y_i = r|Y_i \geq r \quad \Leftrightarrow \quad U_r \leq \theta_r \qquad \text{oder} \qquad Y_i > r|Y_i \geq r \quad \Leftrightarrow \quad U_r > \theta_r \tag{3.6}$$

Die bedingte Wahrscheinlichkeit $P(Y_i = r|Y_i \geq r, \mathbf{x}_i)$, dass eine Beobachtungseinheit in Kategorie r fällt, bzw. ein Prozess in Kategorie r endet, lässt sich mit Hilfe der Verteilungsfunktion des Störterms bestimmen:

$$\begin{aligned} P(Y_i = r|Y_i \geq r, \mathbf{x}_i) &= P(U_r \leq \theta_r) = P(-\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_r \leq \theta_r) = P(\epsilon_r \leq \theta_r + \mathbf{x}_i'\boldsymbol{\beta}) \\ &= F(\epsilon_r + \mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

Parametrisiert man den Schwellenwert als einen kategoriespezifischen Koeffizienten für die Konstante des linearen Prädiktors mit $\beta_{r0} = \theta_r$ erhält man für Beobachtungseinheit i:

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}), \quad r = 1, ..., q \tag{3.7}$$

Die ausschließlich auf den Kovariablenvektor bedingte Wahrscheinlichkeit $\pi_{ir} = P(Y_i = r|\mathbf{x}_i)$ für eine Kategorie r, berechnet sich als das Produkt der Wahrscheinlichkeiten den Übergang in eine höhere als Kategorie r nicht zu vollziehen und der Wahrscheinlichkeit die Übergänge bis hin zu Kategorie r vollzogen zu haben:

$$P(Y_i = r|\mathbf{x}_i) = P(Y_i = r|Y_i \geq r, \mathbf{x}_i) \cdot P(Y_i \geq r|\mathbf{x}_i) \tag{3.8}$$

$$= P(Y_i = r|Y_i \geq r, \mathbf{x}_i) \prod_{s=1}^{r-1} P(Y_i > s|Y_i \geq s, \mathbf{x}_i) \tag{3.9}$$

$$= F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}) \prod_{s=1}^{r-1} (1 - F(\beta_{s0} + \mathbf{x}_i'\boldsymbol{\beta})), \quad r = 1, ..., k \tag{3.10}$$

3.3.2 Modellvarianten

Ebenso wie im kumulativen Modell, ergeben sich Varianten des sequentiellen Modells durch die Wahl der Verteilungsfunktion F des Störterms ϵ_r. Wird für die Verteilungsfunktion die logistische Verteilung $F(\epsilon_r) = \exp(\epsilon_r)/(1 + \exp(\epsilon_r))$ angenommen, erhält man das logistische sequentielle Modell (sequentielles Logit-Modell). Für die bedingte Wahrscheinlichkeit des Verbleibs in Kategorie r ergibt sich :

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta})}, \quad r = 1, \ldots, q \tag{3.11}$$

Zwecks einer intuitiven Interpretation eignet sich die Darstellung des sequentiellen Logit-Modells durch das logarithmierte Verhältnis der bedingten Chancen eines Verbleibens in Kategorie r, statt eines Übergangs in eine höhere Kategorie:

$$\log\left(\frac{P(Y_i = r|Y_i \geq r, \mathbf{x}_i)}{1 - P(Y_i = r|Y_i \geq r, \mathbf{x}_i)}\right) = \log\left(\frac{P(Y_i = r|Y_i \geq r, \mathbf{x}_i)}{P(Y_i > r|Y_i \geq r, \mathbf{x}_i)}\right) = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta} \tag{3.12}$$

Für die Veränderung einer Kovariable x_j um eine Einheit, lässt sich der geschätzte Werte eines Regressionskoeffizienten β_j als die additive Veränderung der logarithmierten Chancen in Kategorie r zu verbleiben, statt in eine höhere Kategorie überzugehen, interpretierten (unter der Bedingung, dass diese Kategorie bereits erreicht wurde und alle übrigen Kovariablen ihren Wert beibehalten). Eine äquivalente Darstellung ist gegeben durch:

$$\frac{P(Y_i = r|Y_i \geq r, x_i)}{P(Y_i > r|Y_i \geq r, x_i)} = \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}) = \exp(\beta_{r0}) \cdot \exp(x_1\beta_1) \cdots \exp(x_p\beta_p) \tag{3.13}$$

In dieser Darstellung besitzt die Veränderung einer Kovariable einen multiplikativen Effekt des exponentierten Regressionskoeffizienten auf das bedingte Chancenverhältnis unter sonst identischen Einflussgrößen. Da dieser Regressionskoeffizient kategorie**un**spezifisch ist, ist für die Interpretation irrelevant, um welchen Übergang es sich handelt. Die Kovariable besitzt im einfachen Modell also einen globalen Effekt. Eine Verallgemeinerung auf kategoriespezifische Effekte wird im folgenden Abschnitt vorgenommen.

Wird eine andere Verteilung für F gewählt, ergeben sich weitere Varianten des sequentiellen Modells:

- für die Gleichverteilung von ϵ_r das lineare sequentielle Modell:

  $P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}$

  Die Einflussgrößen wirken unmittelbar linear auf die Chancen des Verbleibs in Kategorie r, statt eines Übergangs in die nächst höhere Kategorie.

- für $F(\epsilon_r) = 1 - \exp(-\epsilon_r)$ (Exponentialverteilung) das exponentielle sequentielle Modell:

  $P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = 1 - \exp(-(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}))$

- für $F(\epsilon_r) = 1 - \exp(-\exp(\epsilon_r))$ (Minimum-Extremwert-/Gompertz-Verteilung) das sequentielle Minimum-Extremwert-Modell, auch *Proportional-Hazards-Model* genannt:

  $P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = 1 - \exp(-\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}))$

  Im Fall der Minimum-Extremwertverteilung ist das sequentielle Modell äquivalent zum kumulativen Modell. Es findet lediglich eine Umparametrisierung des kategoriespezifischen Intercepts statt: $\beta_{r0} = \log(\exp(\tilde{\beta}_{r,0}) - \exp(\tilde{\beta}_{r-1,0}))$, $r = 1, \ldots, k-1$. Dabei bezeichnet $\tilde{\beta}_{r0}$ den Intercept des kumulativen Modells.

3.3.3 Verallgemeinerung des Modells

Unter der Annahme, dass der Effekt einer Kovariable nicht für alle Übergänge gleich auf die kategoriespezifische Wahrscheinlichkeit wirkt, lässt sich entsprechend der Verallgemeinerung des kumulativen Modells auch das verallgemeinerte

sequentielle Modell mit kategoriespezifischen Kovariableneffekten formulieren:

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = F(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r), \quad r = 1, ..., q \tag{3.14}$$

Zur Herleitung des Modellansatzes wird angenommen, dass die Kovarialen einen, für den jeweiligen Übergang spezifischen, linearen Einfluss auf die latente Variable U_r besitzen, sodass $U_r = -\mathbf{x}'\boldsymbol{\beta}_r + \epsilon_r$.

Mit der Verallgemeinerung des Modells wird wiederum eine parametersparsame Modellierung zugunsten einer höheren Flexibilität aufgegeben. Sofern beispielsweise die Zielgröße die Monate der Arbeitslosigkeit bezeichnet, lässt sich ein variierender Effekt des Alters oder des Geschlechts, über die Dauer der Arbeistlosigkeit hinweg, spezifizieren. Diese Verallgemeinerung erfordert es nicht zwingend, für alle Kovariablen kategoriespezifische Effekte annehmen zu müssen. Dies gilt äquivalent auch für das verallgemeinerte kumulative Modell. Für Kovariablen können sowohl kategorieunspezifische (globale), als auch kategoriespezifische Effekte modelliert werden. Der kategoriespezifische lineare Prädiktor weist dann eine Mischung dieser Effekten auf: $F(\eta_{ir}) = F(\beta_{r0} + \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\beta}_r)$. $\mathbf{z}_i$ bezeichne dabei den Vektor der Kovariablen mit kategorieunspezifischen Effekten $\boldsymbol{\gamma}$ und $\mathbf{x}_i$ den Vektor der Kovariablen mit kategoriespezifischen Effekten $\boldsymbol{\beta}_r$. Wiederum führt die Annahme von kategoriespezifschen Effekten für eine Kovariable dazu, dass nicht mehr nur ein Koeffizient für diese Kovariable geschätzt werden muss, sondern q Effekte. Die starke Zunahme von zu schätzenden Regressionskoeffizienten kann dazu führen, dass eine ML-Schätzung zunehmend instabil wird. Für den Fall $p > n$, dass mehr Koeffizienten zu schätzen, als Beobachtungen im Modell vorhanden sind, existiert gar kein ML-Schätzer. Für eine übersichtlichere Schreibweise werden im Folgenden ausschließlich kategoriespezifische Effekte verwendet.

Für die Wahl der logistischen Verteilung resultiert das verallgemeinerte sequentielle Logit-Modell:

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r)}{1 + \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r)}, \quad r = 1, \ldots, q \tag{3.15}$$

Eine einfache rechnerische Umformung ergibt äquivalent das logarithmierte Chancenverhältnis

$$\log\left(\frac{P(Y_i = r|Y_i \geq r, \mathbf{x}_i)}{P(Y_i > r|Y_i \geq r, \mathbf{x}_i)}\right) = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r, \tag{3.16}$$

aus dem die Interpretation der Regressionskoeffizienten deutlich wird. Im Unterschied zum einfachen sequentiellen Logit-Modell, ist in diesem Fall die Interpretation eines Effekts einer Kovariable daran gebunden, für welche Kategorie das Chancenverhältnis betrachtet wird.

Ebenso wie das sequentielle Logit-Modell lassen, sich die Modellvarianten der anderen genannten Verteilungsfunktionen verallgemeinern, indem der lineare Prädiktor auf kategoriespezifische Effekte erweitert wird.

3.3.4 Darstellung als multivariates GLM

Für die Darstellung des sequentiellen Modells als multivariates GLM wird - wie in Abschnitt 2.3 beschrieben - die Multinomialverteilung als stochastische Komponente verwendet. Das konkrete Aussehen vektorwertiger Response- und Linkfunktionen wird im Folgenden dargestellt:

Sei $\eta_{ir} = \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r$, $r = 1, \ldots, q$ der lineare Prädiktor der r-ten Kategorie mit kategoriespezifischem Intercept und kategoriespezifischen Kovariableneffekten. Da, wie in Gleichung 3.10 ersichtlich, die betrachtete unbedingte Wahrscheinlichkeit $\pi_{ir} = P(Y_i = r|\mathbf{x}_i)$ von allen linearen Prädiktoren bis einschließlich dem r-ten Prädiktor abhängt, lässt sich für π_{ir} formulieren:

$$\pi_{ir} = h_r(\boldsymbol{\eta}_i) = h_r(\eta_{i1}, \ldots, \eta_{ir}) = F(\eta_{ir}) \prod_{s=1}^{r-1} (1 - F(\eta_{is})). \tag{3.17}$$

Für die Verknüpfung der linearen Prädiktoren mit dem q-dimensionalen Wahrscheinlichkeitsvektor $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ ist eine q-dimensionale multivariate Responsefunktion $h = (h_1, \ldots, h_q)' : \mathbb{R}^q \to \mathbb{R}^q$ notwendig, sodass $\boldsymbol{\pi}_i = h(\boldsymbol{\eta}_i) = h(\mathbf{X}_i'\boldsymbol{\beta})$. Mit Hilfe der individuenspezifischen Designmatrix $\mathbf{X}_i$ der Dimension $q \times (q + q \cdot p)$ und dem Vektor $\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{q0}, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$ aller Koeffizienten, stellt sich das sequentielle Logit-Modell mit kategoriespezifischen Koeffizienten als multivariates GLM wie folgt dar:

$$\begin{pmatrix} \pi_{i1} \\ \vdots \\ \pi_{iq} \end{pmatrix} = h \left\{ \begin{pmatrix} 1 & & & \mathbf{x}_i' & & & \\ & 1 & & & \mathbf{x}_i' & & \\ & & \ddots & & & \ddots & \\ & & & 1 & & & \mathbf{x}_i' \end{pmatrix} \begin{pmatrix} \beta_{10} \\ \vdots \\ \beta_{q0} \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_0 \end{pmatrix} \right\} \tag{3.18}$$

Diese Darstellung gilt sowohl für das verallgemeinerte sequentielle, also auch kumulative Modell. Die einfachere Version des Modells ohne kategoriespezifische Kovariableneffekte hat für sequentielles und kumulatives Modell die Gestaltung:

$$\begin{pmatrix} \pi_{i1} \\ \vdots \\ \pi_{iq} \end{pmatrix} = h \left\{ \begin{pmatrix} 1 & & & \mathbf{x}_i' \\ & 1 & & \mathbf{x}_i' \\ & & \ddots & \vdots \\ & & & 1 \; \mathbf{x}_i' \end{pmatrix} \begin{pmatrix} \beta_{10} \\ \vdots \\ \beta_{q0} \\ \boldsymbol{\beta} \end{pmatrix} \right\} \tag{3.19}$$

Die äquivalente Formulierung mittels Linkfunktion $g = h^{-1} = (g_1, \ldots, g_q) : \mathbb{R}^q \to \mathbb{R}^q$ ergibt für die r-te Komponente der Linkfunktion:

$$g_r(\pi_{i1}, \ldots, \pi_{iq}) = F(\pi_{ir}/(1 - \pi_{i1} - \cdots - \pi_{i,r-1})) \tag{3.20}$$

3.3.5 Schätzung der Modellparameter

Zur Bestimmung der Regressionskoeffizienten mittels Maximum-Likelihood Methode wird ebenso, wie zur Motivation des Modellansatzes auf die sukzessiven binären Übergänge zurückgegriffen. Betrachtet wird zunächst der Likelihoodbeitrag einer Beobachtungseinheit i, dessen Responsevariable Y_i in Kategorie r_i fällt. Statt dem q=(k-1) - dimensionalen 0-1-Vektor, wird die verkürzte Variante eines r-dimensionalen Vektors $(y_{i1}, \ldots, y_{ir_i}) = (0, \ldots, 1)$, dessen Einträge nach der „1“ entfernt wurden, betrachtet. Der Likelihoodbeitrag L_i und der Log-Likelihoodbeitrag l_i dieser Beobachtungseinheit ergeben sich anhand Gleichung

3.10 und des verkürzten Responsevektors als:

$$
\begin{aligned}
L_i = P(Y_i = r_i|\mathbf{x}_i) &= F(\eta_{ir_i}) \prod_{j=1}^{r_i-1} (1 - F(\eta_{ij})) \\
&= \prod_{j=1}^{r_i} F(\eta_{ij})^{y_{ij}} (1 - F(\eta_{ij}))^{1-y_{ij}} \\
l_i = \log(L_i) &= \sum_{j=1}^{r_i} [y_{ij} \log(F(\eta_{ij})) + (1 - y_{ij}) \log(1 - F(\eta_{ij}))]
\end{aligned}
$$

Aus der zweiten Zeile des Likelihoodbeitrags wird ersichtlich, dass die 0-1-Einträge y_{ij} der verkürzten Version des Responsevektors steuern, welcher lineare Prädiktor aktiviert wird, da $F(\eta_{ij})$ bleibt, wenn $y_{ij} = 1$, und $(1 - F(\eta_{ij}))$, wenn $y_{ij} = 0$. Mit $L = \prod_{i=1}^{n} L_i$ und $l = \log(L)$ erhält man den Likelihood- und den log-Likelihoodbeitrag der gesamten Beobachtungen:

$$
L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \prod_{j=1}^{r_i} F(\eta_{ij})^{y_{ij}} (1 - F(\eta_{ij}))^{1-y_{ij}} \tag{3.21}
$$

$$
\mathrm{l} = \log(\mathrm{L}) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} [y_{ij} \log(F(\eta_{ij})) + (1 - y_{ij}) \log(1 - F(\eta_{ij}))] \tag{3.22}
$$

Zum Vergleich wird die Likelihood und die log-Likelihood für das binäre Modell mit $y_i \in \{0, 1\}$ aus Abschnitt 2.2 angegeben:

$$
L = \prod_{i=1}^{n} F(\eta_i)^{y_i} (1 - F(\eta_i))^{1-y_i} \tag{3.23}
$$

$$
l = \log(L) = \sum_{j=1}^{n} [y_i \log(F(\eta_i)) + (1 - y_i) \log(1 - F(\eta_i))] \tag{3.24}
$$

Die Ähnlichkeit zwischen der Likelihood/log-Likelihood des sequentiellen Modells $P(Y_i = r|Y_i \geq r, x_{ir}) = F(\eta_{ir})$ aus Gleichung 3.22 zu der des binären Modells $P(y_{ir} = 1|x_{ir}) = F(\eta_{ir})$ aus Gleichung 3.24 ist das Ergebnis der sukzessiven binären Übergänge je Beobachtungseinheit. Dabei weist die Likelihood des sequentiellen Modells der n Beobachtunseinheiten $r_1 + \cdots + r_n$ binäre Übergänge auf. Es ist zu berücksichtigen, dass die r_i binären Übergänge, die mit Hilfe der trunkierten Version des vollständigen multinomialen Responsevektors für jede Beobachtungseinheit erzeugt werden, keine unabhängigen Beobachtungen sind. Somit lassen sich keine Inferenzmethoden für binäre Modelle anwenden. Es muss auf Inferenzmethoden für multivariate Verteilungen zurückgegriffen werden.(Vgl. Tutz (2012), S. 264 f.)

3.4 Beziehung zur Survival-Analyse

Wie bereits aus der ursprünglichen Herleitung des kumulativen Modells von McCullagh (1980) ersichtlich wird, besteht ein enger Zusammenhang zwischen ordinalen Regressionsmodellen und zeitdiskreten Survivalmodellen.

In der Survival-Analyse (Synonym: Lebensdauer-, Verweildaueranalyse) ist die **Zeit** von Interesse, die **bis zum Eintritt eines Ereignisses**, z.B. dem Tod oder Ausfall einer Beobachtungseinheit, verstreicht. Die entsprechende Benennung dieser Zeitdauer (Überlebenszeit/Verweildauer/Ausfallzeit) ergibt sich aus dem Kontext der konkreten Datengrundlage und Fragestellung. Da sich das Merkmal Zeit je nach Datengrundlage, sowohl als eine stetige metrische Variable, als auch eine diskrete Variable auffassen lässt, ist zwischen Modellen für stetige und für **diskrete Survival-Zeiten** zu unterscheiden. Dem Zusammenhang zu ordinalen Regressionsmodellen entsprechend, findet eine Fokussierug auf diskrete Lebenszeiten statt. Einem Ereigniszeitpunkt wird die diskrete Zeit $T \in \{1, \ldots, k\}$ mit $T = t$ zugewiesen, wenn das Ereignis im Zeitintervall $[a_{t-1}, a_t)$ stattgefunden hat. Hierzu wird das Zeitintervall aller Beobachtungen in k Teilintervalle $[a_0, a_1), [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, a_\infty)$ mit $q = k - 1$ zerlegt.[2] Gewöhnlich wird der Beginn der Zeitmessung $a_0 = 0$ gesetzt. Diese Diskretisierung der Zeitskala eignet sich dann, wenn der Zeitpunkt des Ereigniseintritts nicht exakt beobachtet wurde, sondern lediglich bekannt ist, dass das Ereignis zwischen zwei bekannten Zeitpunkten stattgefunden hat. Ein derartiges Ereignis wird als intervallzensiert bezeichnet. Die konstruierten Zeitintervalle lassen sich als Kategorien interpretieren, womit eine Verbindung zu ordinalen Regressionsmodellen hergestellt wird, insbesondere zum sequentiellen Modell, da diese Zeitintervalle bis zum Ereigniseintritt sukzessive durchschritten werden.

In der Lebensdaueranalyse wird, zur Charakterisierung der Verteilung der nichtnegativen Zufallsgröße T, die die Zeit bis zu dem vorher spezifizierten Ereignis (gegeben eines Kovariablenvektors) angibt, neben der Wahrscheinlichkeitsdichte $f_T(t|\mathbf{x})$ und der Verteilungsfunktion $F_T(t|\mathbf{x})$ die Hazardfunktion $\lambda_T(t|\mathbf{x})$ und die Survivalfunktion $S_T(t|\mathbf{x})$ verwendet. Alle vier Funktionen lassen sich ineinander umformen und eignen sich dazu, verschiedene Aspekte der Verteilung von T zu illustrieren.

Die diskrete Hazardfunktion $\lambda_T(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x})$, $t = 1, \ldots, q$ gibt die bedingte Wahrscheinlichkeit an, dass eine Beobachtungseinheit den Ereigniseintritt in Intervall $[a_{t-1}, a_t)$ erlangt, gegeben, dass dieses Intervall erreicht wurde. Die Interpretation der diskrete Hazardfunktion entspricht somit der Wahrscheinlichkeit in Gleichung 3.7, dass eine Beobachtungseinheit in Kategorie t bzw. r verbleibt, gegeben, dass alle vorherigen t-1 bzw. r-1 Kategorien durchschritten wurden. Die Survivalfunktion gibt die Wahrscheinlichkeit an, dass der Zeitpunkt t erreicht wird, ehe ein Ereignis eintritt: $S_T(t|\mathbf{x}) = 1 - F_T(t|\mathbf{x}) = P(T > t|\mathbf{x}) = \prod_{s=1}^{t} (1 - \lambda_T(s|\mathbf{x}))$. Mit $\widetilde{S}_T(t|\mathbf{x}) = P(T \geq t|\mathbf{x}) = S_T(t-1|\mathbf{x})$ ergibt sich die unbedingte Wahrscheinlichkeit in Gleichung 3.10 als:

$$P(T = t|\mathbf{x}) = \lambda_T(t|\mathbf{x}) \prod_{s=1}^{t-1} (1 - \lambda_T(s|\mathbf{x})) = \lambda_T(t|\mathbf{x}) \cdot \widetilde{S}_T(t|\mathbf{x}), \quad t = 1, ..., k \quad (3.25)$$

[2]Da im Folgenden keine Modellbildung stattfindet, wird auf einen Beobachtungseinheitenindex i (T_i) verzichtet.

Parametrische Regressionsmodelle zur Schätzung der diskreten Hazardfunktion in Abhängigkeit von Kovariablen erhält man aus

$$\lambda_T(t|\mathbf{x}) = F(\beta_{t0} + \mathbf{x}'\boldsymbol{\beta}) \tag{3.26}$$

für geeignete Wahl der Verteilungsfunktion F. Für die logistische Verteilungsfunktion resultiert das sequentielle Logit-Modell. Dieses konvergiert gegen das zeitstetige Cox-Modell, wenn die Intervallbreiten gegen Null konvergieren. Für die Wahl der Minimum-Extremwert-Verteilung resultiert das Gruppierte Cox-Modell.

Ist die Anzahl der Intervalle/Kategorien sehr hoch (im Vergleich zu der Anzahl an Beobachtungseinheiten), ergibt sich eine großen Anzahl zu schätzender Parameter $\beta_{10}, \ldots, \beta_{q0}$. Die Problematik hochdimensionaler Modelle verbleibt wie für die ordinalen Regressionsmodelle auch hier: Instabilität bzw. Nichtexistenz von ML-Schätzern. Eine Lösung ergibt sich, indem man die Baseline-Hazardrate, die durch die intervallspezifischen Intercepts gegeben ist, durch Polynom- oder Regressions-Splines schätzt. Ist eine ML-Schätzung möglich, lässt sich die Likelihood identisch zu der des sequentiellen Modells der binären Übergänge herleiten und somit die ML-Schätzer auf dem selben Weg wie für multivariate GLM bestimmen. Es ist allerdings notwendig für zensierte Daten ein Random Censoring anzunehmen.(Vgl. Fahrmeir & Tutz (2001), S. 396 ff.) Generell versteht man unter Zensierung ein Charakteristikum, das oftmals in Daten, die die Zeitdauer bis zu einem Ereigniseintritt beschreiben, auftritt. Eine Beobachtungseinheit wird dann als zensiert bezeichnet, wenn zwar ein Intervall bekannt ist, in dem das Ereignis eingetreten ist, nicht aber der exakte Zeitpunkt. Von einer rechts-zensierten Beobachtung spricht man, wenn bekannt ist, dass das Ereignis nach einem Zeitpunkt t eingetreten ist, dieser selbst aber nicht bekannt ist. Rechts-Zensierungen treten auf, wenn in einer Studie von Beginn an eine feste Anzahl an Einheiten beobachtet wird und neben Ereigniszeiten anderweitige Abgänge (Zensierungen) aus der Studie zu verzeichnen sind. Beispiele hierfür sind der zufällige Tod, der Wegzug eines Individuums oder eine Reduktion der beobachteten Einheiten aufgrund des Studiendesigns. Ist der Zeitpunkt der Zensierung einer Beobachtung unabhängig von dessen Ereigniszeitpunkt, spricht man von Random Censoring. Dies bedeutet, dass aus dem Zensierungszeitpunkt kein Rückschluss auf den unbekannten Zeitpunkt des Ereignisses gezogen und somit keine zusätzliche Information gewonnen werden kann. Eine Darstellung verschiedener Zensierungsmechanismen findet sich in Klein & Moeschberger (2003), Kapitel 3.

In Abschnitt 6.2 wird ein Datensatz aus der Münchner Gründerstudie zur Anwendung verschiedener Penalisierungsansätze im sequentiellen Logit-Modell herangezogen. Dieser enthält Daten von neu gegründeten Unternehmen, sowie deren Zeitdauer in Monaten bis zu einer eventuellen Insolvenz und lässt sich somit in das Gebiet der Survival-Anlayse einordnen. Anhand der Verknüfung von Survival-Modellen für diskrete Zeitpunkte mit dem sequentiellen Logit-Modell, ist mit Hilfe der Penalisierungsansätze eine parameterintensivere Modellierung auch für Survival-Daten möglich, sofern diese mittels Maximum-Likelihood geschätzt werden. Eine penalisierte ML-Schätzung erlaubt es, auch im Fall $p > n$ Schätzer für die Kovariablenkoeffizienten zu finden, wodurch parameterintensivere Modelle mit kategoriespezifischen Effekten aufgestellt werden können.

3.5 Gegenüberstellung der beiden Modelltypen und Zusammenfassung

Sowohl das kumulative, als auch das sequentielle Modell dienen als Regressionsmodelle für mehrkategoriale Zielgrößen mit geordneten Kategorien. Dabei lässt sich das sequentielle Modell nur dann verwenden, wenn eine höhere Kategorie erst erreicht wird, wenn alle vorangegangenen Kategorien sukzessive durchschritten worden sind. Ist eine derartige Interpretation der Kategoriestruktur möglich, wird für gewöhnlich das sequentielle dem kumulativen Modell vorgezogen. Um dies zu begründen, werden im Folgenden die beiden Modelltypen, hinsichtlich der Interpretier- und Schätzbarkeit ihrer Parameter und der Flexibilität ihrer Modellierung, miteinander verglichen. (Vgl. Tutz (2012), S. 257.)

Die **Interpretation der Koeffizienten** hängt von dem jeweiligen Modelltyp und der Modellvariante ab. Eine plausible Interpretation der Modellkoeffizienten ergibt sich, wenn im kumulative Modell das kumulative Chancenverhältnis $\frac{P(Y_i \leq r|\mathbf{x}_i)}{P(Y_i > r|\mathbf{x}_i)}$, im sequentiellen Modell die Chancen des Übergangs zur nächst höheren Kategorie $\frac{P(Y_i = r|Y_i \geq r, x_i)}{P(Y_i > r|Y_i \geq r, x_i)}$ betrachtet werden. Die intuitive Interpretation, dass eine Veränderung der Variable x_j um eine Einheit einen multiplikativen Effekt $\exp(\beta_j)$ bzw. $\exp(\beta_{rj})$ auf die genannten Chancenverhältnisse besitzt, beschränkt sich auf die Modellvarianten mit logistischer Verteilungsfunktion.

Für das einfache kumulative Modell wird vorausgesetzt, dass die dem Modellansatz zugrundeliegenden Schwellenwerte auf dem latenten Kontinuum geordnet sind, sodass $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$. Für das verallgemeinerte kumulative Modell muss, damit für alle Kategorien $P(Y_i \leq r - 1|\mathbf{x}_i) \leq P(Y_i \leq r|\mathbf{x}_i)$ gilt, die Bedingung $\beta_{r-1,0} + \mathbf{x}_i'\boldsymbol{\beta}_{r-1} \leq \beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r$, $\forall r, \forall \mathbf{x}_i$ erfüllt sein. Werden diese Restriktionen an die Parameter bei der Konstruktion des Schätzalgorithmus nicht berücksichtigt, konvergieren die iterativen Schätzalgorithmen gegebenenfalls nicht. Die **Existenz eines Schätzers** im kumulativen Modell ist somit nicht gewährleistet. Da diese Restriktionen für die Parameter des sequentiellen Modells nicht gelten, ist die Schätzung auch komplexerer und somit flexiblerer Modelle einfacher als im kumulativen Modell.

Die **Modellkomplexität** im Sinne der Anzahl zu schätzender Parameter bestimmt sich in den beiden Modelltypen durch die Anzahl der Zielgrößenkategorien und durch die Anzahl in das Modell einbezogener Prädiktoren. Unter Verwendung ausschließlich kategoriespezifischer Kovariablen sind für ein verallgemeinertes Modell inklusive kategoriespezifischer Intercepts $q \times (p + 1)$ Regressionskoeffizienten zu schätzen. Pro zusätzlicher Kategorie müssen $p + 1$ zusätzliche Koeffizienten, pro zusätzlichem kategoriespezifischen Prädiktor q zusätzliche Koeffizienten geschätzt werden. Dabei wurde bisher implizit angenommen, dass ein Prädiktor durch maximal einen Koeffizienten pro Kategorie in das Modell eingeht. Dies ist für metrische und binäre Prädiktoren der Fall. Tritt ein kategorialer Prädiktor mit mehr als zwei Kategorien auf, vervielfacht sich die Anzahl zu schätzender Parameter in Abhängigkeit der Anzahl seiner Kategorien. Dies wird explizit im nächsten Abschnitt berücksichtigt.

Da für den Fall $p > n$ keine ML-Schätzer existieren, muss sich auf parametersparsame, weniger flexible Modelle beschränkt werden, wenn die Anzahl der

Beobachtungseinheiten im Vergleich zur Kategorieanzahl oder zu der Zahl der Prädiktoren gering ist. Ansätze, die auch in diesen Fällen eine Schätzung ermöglichen und gleichzeitig Kovariablen mit schwachen Effekten aus dem Modell entfernen, sind die im folgenden Kapitel dargestellten, auf einer Penalisierung der log-Likelihood beruhenden, Penalisierungskonzepte.

Kapitel 4

Penalisierungsansätze

Im vorangegangenen Kapitel wurde anhand der verallgemeinerten Modellansätze bereits der hohe Grad an Parameterintensitivität erörtert, der durch eine Vielzahl von Responsekategorien oder kategoriespezifischen Prädiktoren erzeugt wird. Um den Auswirkungen dieser Modellkomplexität auf die Parameterschätzung und -interpretation Rechnung zu tragen, wird eine Regularisierung notwendig. Die in dieser Arbeit betrachteten Regularisierungsansätze basieren auf einer Penalisierung der log-Likelihoodfunktion. Die dahinter steckende Intention sowie dessen Effekt, werden in Abschnitt 4.1 dargelegt. In Abschnitt 4.2 werden mit der *Ridge Regression* und dem *Lasso-Verfahren* grundlegende Penalisierungsansätze, mit dem *Group Lasso* und dem *Sparse Group Lasso* erweiterte Ansätze skizziert. Mittels *Adaptive Lasso* und *Refitting* werden in Abschnitt 4.3 zwei Methoden zur Erweiterung der Penalisierungsansätze vorgeschlagen, die die Schätz- und Selektionseigenschaften optimieren sollen. Abschließend werden die Ergebnisse in Abschnitt 4.4 zusammengefasst. Die Darstellung orientiert sich an Tutz, Pößnecker & Uhlmann (2012), sowie Tutz (2012), Kapitel 6.

4.1 Intention und Grundlagen

4.1.1 Problemstellung und Lösungsansätze

Abhängig von der Anzahl der Responsekategorien und Prädiktoren, die in ein (verallgemeinertes) sequentielles Modell aufgenommen werden, resultieren Modelle, deren zu schätzende Parameteranzahl p die Anzahl verfügbarer Beobachtungen n, des zu untersuchenden Datensatzes, (deutlich) überschreitet. In den Fällen $p > n$ oder $p \gg n$ existieren für die ML-Schätzung keine Schätzer mehr, da diese Schätzmethode mehr Beobachtungen benötigt, als Koeffizienten zu schätzen sind. Selbst in Situationen, in denen p im Vergleich zu n groß ist, oder Kollinearität in der Designmatrix der Einflussgrößen auftritt, sind die **ML-Schätzer instabil bzw. nicht existent**.

Sind die Koeffizienten des Modells schätzbar, ist man zusätzlich daran interessiert, hinsichtlich deren Interpretation, nur die Prädiktoren herauszustellen, die die stärksten Effekte aufweisen. Liegt eine Vielzahl an Einflussgrößen vor, die durch deren gemeinsame Parameterschätzung kleine Effekte erhalten, ist eine klare **Interpretation des Modells beeinträchtigt**. In diesem Fall wäre eine

Selektion der Variablen mit den stärksten Effekten wünschenswert.

Subset Selection

Eine Möglichkeit, sowohl die Parameterschätzbarkeit, als auch die Selektion von Prädiktoren zu gewährleisten, ist durch Variablenselektionsverfahren gegeben. *Schrittweise* Selektion, wie die *Vorwärtsselektion* gehen von einem Interceptmodell aus und fügen diesem Startmodell diejenigen Variablen hinzu, die die Anpassungsgüte im Sinne eines Kriteriums (AIC, BIC) verbessern. Unter Verwendung der Vorwärtsselektion ist eine Parameterschätzbarkeit gegeben, solange $p < n$ ist. In der *Rückwärtsselektion* werden, von dem voll parametrisierten Modell ausgehend, schrittweise diejenigen Prädiktoren entfernt, die die Anpassungsgüte am geringsten verbessern. Das maximale Modell lässt sich allerdings nur im Fall $p < n$ bestimmen. Beide Seklektionsprozesse lassen sich zudem kombinieren, indem in jedem Schritt nicht nur eine Variable hinzugefügt, sondern auch wieder entfernt werden kann. Eine weitere Möglichkeit, die *Best Subset Selection*, besteht darin, alle möglichen Teilmengen von Prädiktoren hinsichtlich ihrer Anpassungsgüte miteinander zu vergleichen und dasjenige Modell mit der besten Prognosegüte auszuwählen. Generell eignen sich diese, als diskrete Variablenselektion bezeichneten Verfahren nur eingeschränkt für kategoriale Regressionsmodelle, da eine Schätzbarkeit der Modelle nicht in allen Fällen möglich ist und Selektionsprozesse bei entsprechend hochdimensionaler Parametrisierung, aufgrund der Vielzahl zu schätzender Modelle, sehr rechenaufwändig sind. Ein weiterer Nachteil diskreter Selektionsverfahren liegt in ihrer Sensibilität gegenüber Veränderungen der Datengrundlage. (Vgl. Hastie, Tibshirani & Friedman (2011), S. 57 ff.)

Penalisierung der log-Likelihood

Die in dieser Arbeit betrachtete Alternative sind Verfahren, die auf einer Penalisierung der log-Likelihood beruhen und somit den Variablenselektionsprozess bereits bei der Koeffizientenschätzung ansetzen. Als Basis dient, diesen verschiedenen Penalisierungsansätzen gemeinsam, die penalisierte log-Likelihoodfunktion $l_p(\boldsymbol{\beta})$, des im jeweiligen Modell zu bestimmenden Parametervektors $\boldsymbol{\beta}$:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}) \tag{4.1}$$

Dabei bezeichne $l(\boldsymbol{\beta})$ die gewöhnliche log-Likelihood des entsprechenden Modells und $J(\boldsymbol{\beta})$ ein Funktional, das eine Norm des Parametervektors penalisiert, z.B. die Länge des Parametervektors bzw. die Größe der geschätzten Koeffizienten bestraft. Der Parameter λ wird als Penalisierungs- oder Tuningparameter bezeichnet und bestimmt, wie stark der Bestrafungsterm auf die log-Likelihood wirken soll. Für $\lambda = 0$ fällt der Bestrafungsterm weg und es resultieren - sofern existent - die gewöhnlichen Maximum-Likelihood-Schätzwerte für die Regressionsparameter. Für steigendes λ nimmt der Einfluss des Penalisierungsterms zu und es resultieren penalisierte Koeffizienten, deren Eigenschaften durch die spezielle Form des Funktionals bestimmt werden. Wie sich die exakte Form des Funktionals auf die geschätzten Koeffizienten auswirkt, wird innerhalb der folgenden Darstellung spezieller Penalisierungsmethoden deutlich.

Den geschätzten Koeffizienten ist gemeinsam, dass sie im Vergleich zum ML-Schätzer, für steigenden Einfluss des Penalisierungsterms, gegen null geschrumpft

werden. Dabei werden Koeffizienten mit schwächeren Effekten schneller gegen null geschrumpt, somit die stärkeren Effekte selektiert. Aufgrund einer stetigen Schrumpfung gegen null, wird dieses Verfahren auch als stetige Variablenselektion bezeichnet. Die Gestaltung des Penalisierungsterms mit einer Norm des Parametervektors lässt sich damit begründen, dass für die Schätzer eine, im Vergleich zu der des ML-Schätzers, reduzierte Varianz erlangt werden kann.

Der mögliche Koeffizientenvektor $\boldsymbol{\beta}$ ist durch die Wahl der Kovariablen für das Regressionsmodell bestimmt. Die Form des Funktionals $J(\boldsymbol{\beta})$ wird ebenfalls vor der Parameterschätzung festgelegt. Für die Bestimmung des optimalen Penalisierungsparameters λ ist ein Auswahlkriterium notwendig, dessen Wahl gewöhnlich auf Akaikes Informationskriterium (AIC), das Bayessche Informationskriterium (BIC) fällt oder sich an bestmöglicher Prognosequalität orientiert, die für diesen optimalen Penalisierungsparameter mit dem Modell erreicht wird. Zur Bestimmung des erwarteten Prognosefehlers lassen sich Kreuzvalidierungsmethoden heranziehen. Für eine k-fache Kreuzvalidierung wird der verwendete Datensatz in k Teildatensätze zerlegt, die annähernd gleich groß sein sollen. Dabei wird für k meist 5 oder 10 festgelegt. Aus den Daten von k-1 Teilen werden die Modellparameter geschätzt, die dazu dienen die n_k Zielgrößen des k-ten Teils zu prognostizieren. Der erwartete Prognosefehler PE berechnet sich als arithmetisches Mittel der quadrierten Abweichungen zwischen tatsächlichen Zielgrößenwerten y_i und prognostizierten Werten $\hat{y}_i$, $i = 1, \ldots, n_k$: $PE = 1/n_k \sum_{i=1}^{n_k} (y_i - \hat{y}_i)^2$. Dieses Vorgehen wird für alle k Teildatensätze wiederholt und anschließen das Mittel der Prognosefehler der Teildatensätze gebildet. Um den optimalen Penalisierungsparameter zu erhalten, wird genanntes Vorgehen für ein Raster von möglichen λ-Werten wiederholt und dasjenige λ, für das der geringste erwartete Prognosefehler vorliegt, gewählt. (Vgl. Hastie, Tibshirani & Friedman (2011), S. 241 ff.)

Folgende Vorteile ergeben sich für die, mit einer Norm des Koeffizientenvektors, penalisierte Schätzung gegenüber der ML-Schätzung bzw. diskreter Selektionsverfahren:(Vgl. Tutz (2012), S. 143)

- Selbst für $p > n$ lassen sich Werte für die Regressionskoeffizienten schätzen, sodass eine **Existenz von Schätzern** gewährleistet ist.
- Eine Selektion der Prädiktoren mit den stärksten Effekten findet implizit dadurch statt, dass manche der Koeffizienten für optimales λ auf null geschätzt werden. Hierdurch wird die **Interpretierbarkeit** durch das parametersparsamere Modell verbessert. Verglichen mit diskreter Selektion ist eine stetige Variablenselektion weniger empfindlich gegenüber Veränderungen in der Datengrundlage.
- Im Vergleich zu ML-Schätzern, können Schätzer mit geringerer Varianz und Modelle mit höherer **Prädiktionsgenauigkeit** generiert werden.

4.1.2 Überblick über Penalisierungsansätze

In den letzten beiden Jahrzehnten wurde eine Vielzahl von möglichen Penalisierungsansätzen vorgeschlagen und bestehende Ansätze weiter entwickelt. Dabei ist die Form des Penalisierungsansatzes, genauer gesagt die Form des Funktionals

und somit die Eigenschaften der resultierenden Koeffizientenschätzer auf eine konkrete Fragestellung bzw. einen konkreten Modelltyp ausgerichtet. Die eventuelle Eignung eines Penalisierungsansatzes zur Verwendung für das sequentielle Logit-Modell hängt davon ab, inwieweit Charakteristika eines multikategorialen Regressionsmodells in der Konstruktion des Funktionals berücksichtigt werden.

Zahlreiche bisherige Penalisierungsansätze sind auf Modelle mit univariaten Zielgrößen ausgerichtet. Hierzu zählen die beiden klassischen Verfahren Ridge Regression von Hoerl & Kennard (1970) und Lasso von Tibshirani (1996), deren Funktional sich verallgemeinert darstellen lässt als:

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{p} |\alpha_j|^{\tau}; \quad \tau > 0 \tag{4.2}$$

Dabei bezeichne $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$ einen p-dimensionalen Vektor von p Einflussgrößen eines univariaten Modells, inklusive Interceptparamter. Dieser dient der übersichtlicheren Darstellung der Idee eines Penalisierungsansatzes und grenzt sich durch seine Einfachheit von dem komplexeren Vektor $\boldsymbol{\beta}$ des sequentiellen Modells ab, der nachfolgend eingeführt wird. Für $\tau = 2$ resultiert die Ridge Regression, dessen Penalisierungsnorm für den Koeffizientenvektor des Modells die ℓ_2-Norm ist, somit die Länge (euklidische Norm) des Vektors. Für $\tau = 1$ resultiert das Lasso-Verfahren, das mit der ℓ_1-Norm penalisiert. Während die Ridge Regression nicht in der Lage ist Prädiktoren zu selektieren, führt Lasso eine Selektion dann durch, wenn ein Prädiktor durch einen einzigen Koeffizienten (metrische oder binäre Variable) im Modell vertreten ist und dieser auf null geschätzt wird, für gegebenen Penalisierungsgrad. Alle Variablen, deren Koeffizient als von null verschieden geschätzt wird, werden somit für das Modell selektiert. Beide grundlegenden Ansätze werden im folgenden Abschnitt betrachtet.

Für das allgemeine sequentielle Logit-Modell aus Gleichung 3.15 auf Seite 23 setzt sich der Koeffizientenvektor $\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{q0}, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$ aus den kategoriespezifischen Intercepts und den kategoriespezifischen Koeffizienten zusammen. Die q kategoriespezifischen Koeffizienten einer metrischen oder binären Einflussgröße x_j lassen sich in dem partiellen Vektor $\boldsymbol{\beta}_{\bullet j} = (\beta_{1j}, \ldots, \beta_{qj})'$ zusammenfassen. Wie bereits festgestellt wurde, findet die Selektion einer Variable, die durch q Koeffizienten im Modell vertreten ist, nur dann statt, wenn diese q Koeffizienten simultan aus dem Modell entfernt werden. Das heißt, das Funktional muss die Eigenschaft aufweisen können, diese q Koeffizienten gleichzeitig als irrelevant zu schätzen, somit diese als eine zusammengehörige Gruppe zu erkennen. Ein Ansatz, der in der Lage ist, mehrere Gruppen von Koeffizienten einzubeziehen wurde mit dem *Group Lasso* von Huan & Lin (2006) vorgeschlagen. Allerdings zielt deren Intention auf Gruppen von Variablen, die den Effekt eines kategorialen Prädiktors auf eine univariate Zielgröße wiederspiegeln. Tutz, Pößnecker & Uhlmann (2012) wenden mit ihrem *Categorically Structured Lasso* (CATS-Lasso) diesen Gruppen-Ansatz auf das multinomiale Logit-Modell an, der den Charakteristika eines multikategorialen Regressionsmodells gerecht wird. Daran anknüpfend, werden in dieser Arbeit Penalisierungsansätze für das sequentielle Logit-Modell miteinander verglichen.

Eine Erweiterung des Group Lasso wurde mit dem *Sparse Group Lasso* von Simon et al. (2012) vorgenommen. Mit dieser Erweiterung lassen sich zusätzlich

auch einzelne Koeffizienten innerhalb einer selektierten Gruppe auf null schätzen, sodass innerhalb einer selektierten Gruppe eine weitere Selektion der stärksten Effekte stattfindet.

Group Lasso und Sparse Group Lasso werden als diejenigen Penalisierungsansätze, die den Erfordernissen des sequentiellen Logit-Modells gerecht werden, ebenfalls im nächsten Abschnitt betrachtet.

4.2 Penalisierungsansätze

4.2.1 Ridge Regression

Hoerl & Kennard (1970) schlagen mit der Ridge Regression einen Penalisierungsansatz vor, der alle Variablenkoeffizienten eines univariaten linearen Regressionsmodells gleichmäßig gegen null (und sich selbst) schrumpft, für zunehmenden Grad der Bestrafung. Hierbei wird die quadrierte ℓ_2-Norm eines Koeffizientenvektors als Penalisierungsfunktional verwendet:

$$J(\boldsymbol{\alpha}) = ||\boldsymbol{\alpha}||_2^2 = \sum_{j=1}^{p} \alpha_j^2 = \alpha_1^2 + \cdots + \alpha_p^2 \tag{4.3}$$

Die Maximierung der penalisierten log-Likelihood

$$\hat{\alpha}^{ridge} = \underset{\alpha}{\operatorname{argmax}} \left\{ l(\boldsymbol{\alpha}) - \lambda \sum_{j=1}^{p} \alpha_j^2 \right\} \tag{4.4}$$

liefert den Koeffizientenschätzer $\hat{\boldsymbol{\alpha}}^{ridge}$. Aufgrund der quadratischen Koeffizienten, lässt sich die Lösung des Maximierungsproblems in Matrixnotation als lineare Funktion in y angeben:

$$\hat{\alpha}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda P)^{-1}\mathbf{X}'y \tag{4.5}$$

P bezeichnet dabei eine $(p+1) \times (p+1)$-Matrix, die einer Einheitsmatrix gleicher Dimensionalität enstpricht, mit dem Unterschied, dass das erste Diagonalelement eine Null ist. $\mathbf{X}$ bezeichnet die $n \times (p+1)$-dimensionale Designmatrix mit Interceptspalte. Der Ridge-Schätzer unterscheidet sich lediglich durch den Term λP von dem ML-Schätzer $\hat{\alpha}^{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$. Dieser Term spiegelt die ursprüngliche Intention von Hoerl & Kennard (1970) wieder, kleine Beträge auf die Diagonale der Produktsummenmatrix $\mathbf{X}'\mathbf{X}$ zu addieren, um diese invertierbar machen zu können. Somit kann auch im Fall einer singulären Produktsummenmatrix, die sich bspw. durch Kollinearität in den Kovariablen ergibt, ein Schätzer bestimmt werden. Diese Behandlung einer singulären Matrix wird als Regularisierung bezeichnet. Durch einfache Berechnung wird deutlich, dass der Ridge-Schätzer nicht erwartungstreu ist. Allerdings kann gezeigt werden, dass der Ridge-Schätzer für bestimmte λ-Werte kleinere Varianz besitzt als der ML-Schätzer. Im Sinne eines Bias-Varianz-Tradeoff kann der verzerrte Ridge-Schätzer somit einen geringeren MSE besitzen als der erwartungstreue ML-Schätzer.

Dass die Ridge Regression simultan alle Koeffizienten für steigenden Grad der Penalisierung stetig gegen null schrumpft, macht diesen Penalisierungsterm ungeeinet für die Verwendung im sequentiellen Logit-Modell. Der Grund liegt darin,

dass keine Koeffizienten auf null geschätzt werden können, die nicht durch den ML-Schätzer bereits einen Null-Koeffizienten erhalten würden. Somit verbleiben für gewöhnlich fast alle Koeffizienten im Modell und es werden keine Variablen aus dem Modell entfernt.

Frank & Friedman (1993) formulieren als Verallgemeinerung der Ridge Regression das Penalisierungsfunktional

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{p} |\alpha_j|^{\tau} \tag{4.6}$$

Der Parameter $\tau > 0$ drückt dabei eine Präferenz bzgl. der Penalisierung der einzelnen Koeffizienten aus. Für $\tau = 2$ resultiert der Ridge Penalisierungsterm, für $\tau = 1$ der Lasso Penalisierungsterm.

4.2.2 Lasso

Eine der zentralen Grundlagen, für die im folgenden dargestellten Penalisierungsansätze, bildet das von Tibshirani (1996) vorgeschlagene Lasso-Verfahren. Das Akronym Lasso steht für *Least Absolute Shrinkage and Selection Operator*. Die Bestimmung des penalisierten Parametervektores $\hat{\boldsymbol{\alpha}}^{lasso}$ erfolgt durch Maximierung der, mittels ℓ_1-Norm penalisierten, log-Likelihood bezüglich $\boldsymbol{\alpha}$:

$$\hat{\alpha}^{lasso} = \underset{\alpha}{\operatorname{argmax}} \left\{ l(\boldsymbol{\alpha}) - \lambda \sum_{j=1}^{p} |\alpha_j| \right\} \tag{4.7}$$

Aufgrund der Betragsfunktion ist die Lösung der Maximierung eine nichtlineare Funktion in y und lässt sich somit nicht in geschlossener Form darstellen. Für die numerische Optimierung existieren effiziente Algorithmen, wie der von Efron et al. (2004) vorgeschlagene LARS-Algorithmus. Da es für den Lasso-Schätzer keine explizite Darstellung gibt, existieren auch keine Standardfehler. Diese können allerdings mittels einer iterativen Ridge-Regression approximativ bestimmt werden.

Der Vorteil des Lasso-Verfahrens besteht darin, dass aufgrund der Penalisierung mit der ℓ_1-Norm für den mittels Kreuzvalidierung bestimmten optimalen Penalisierungsparameter λ, sowohl Regressionskoeffizienten direkt auf null gesetzt werden - implizit also selektiert werden-, als auch geschrumpft werden.´Es resultieren parametersparsame Modell, die somit eine verbesserte Interpretiebarkeit und Prädiktionsfähigkeit aufweisen. (Vgl. Tibshirani (1996))

Allerdings ist die geeignete Verwendung des Lasso-Penalisierungsansatzes auf Modelle mit univariaten Zielgrößen beschränkt, deren Prädiktoren entweder metrisch oder binär sind. In beiden Fällen wird die Zielgröße ausschließlich durch einen einzigen Regressionskoeffizienten beeinflusst.

Für das sequentielle Logit-Modell erhält man unter Verwendung kategoriespezifischer Kovariableneffekte den Parametervektor $\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{q0}, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$. Ein Prädiktor wird also durch mehrere Koeffizienten im Modell vertreten. Der Lasso-Penalisierungsterm hätte für das sequentielle Logit-Modell die Form:

$$J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 = \sum_{r=1}^{q} ||\boldsymbol{\beta}_r||_1 = \sum_{r=1}^{q}\sum_{j=1}^{p} |\beta_{rj}| \tag{4.8}$$

Mit dem *elastic net* wird von Zou & Hastie (2005) ein Ansatz vorgeschlagen, der durch das Penalisierungsfunktional

$$J(\boldsymbol{\alpha}) = \lambda \sum_{j=1}^{p} (\nu\alpha_j^2 + (1-\nu)|\alpha_j|), \quad \nu \in [0,1] \tag{4.9}$$

einen Kompromis zwischen Ridge Regression und Lasso findet. Dieser Ansatz teilt den Variablenselektionsprozess des Lasso, als auch die Fähigkeit der Ridge Regression, Koeffizienten zueinander hin zu schrumpfen. Zou & Hastie (2005) zeigen, dass dem Lasso ähnlich sparsame Modell erzeugt werden können, gleichzeitig auch die Prognosegüte erhöht werden kann. Ein weiterer Vorteil des *elastic net* liegt darin, Gruppen korrelierter Variablen entweder gemeinsam zu selektieren oder komplett aus dem Modell zu entfernen, wie es in Gen-Studien wünschenswert ist. Ähnlich dem Lasso, ist das elastic net, aufgrund fehlender Eigenschaft eine Gruppe von Koeffizienten zu selektieren, ungeeignet für das sequentielle Modell und wird deswegen nicht weiter betrachtet.

4.2.3 Group Lasso

Yuan & Lin (2006) schlagen mit dem Group Lasso einen Penalisierungsansatz vor, der es erlaubt Gruppen von Koeffizienten gemeinsam zu selektieren. Um diesen Ansatz darzustellen, wird zunächst ein Regressionsmodell mit univariater Zielgröße und mehreren kategorialen Prädiktoren (Faktorvariablen) betrachtet. Ein kategorialer Prädiktor mit ℓ Kategorien werde anhand einer Dummykodierung mit Hilfe von ℓ-1 Dummyvariablen in das Modell aufgenommen. Somit tritt diese Faktorvariable durch ℓ-1 Koeffizienten im Modell auf. Eine Verwendung des klassischen Lasso-Verfahrens würde dazu führen, eventuell einzelne dieser ℓ-1 Koeffizienten auf null zu schätzen, andere hingegen nicht. Des Weiteren wäre diese Selektionslösung davon abhängig, mit welcher Referenzkategorie die Faktorvariable kodiert wurde. Die Idee des Group Lasso besteht darin, die zu einem Faktor gehörige Koeffizientengruppe entweder gemeinsam aus dem Modell zu entfernen, d.h. alle Koeffizienten simultan auf null zu schätzen, oder alle Koeffizienten dieser Gruppe gemeinsam im Modell zu behalten.

Für eine formale Darstellung dieses Penalisierungsfunktionals werde angenommen, dass $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iG}')$, $j = 1, \ldots, G$ ein p-dimensionaler Parametervektor sei, dessen Einträge G Gruppen der dummykodierten Repräsentanten der Faktorvariablen sind. Die $df_j = \ell_j - 1$ Einträge des Vektors $\mathbf{x}_{ij}$ repräsentieren folglich die Kodierung des Faktors j, der ℓ_j Kategorien besitzt. Des Weiteren kann eine derartige 'Gruppe' auch aus einer metrischen Variable mit $df_j = 1$, also einem Koeffizienten bestehen. Die Dimension p bestimmt sich als $df_1 + \cdots + df_G$. Der dazugehörige Parametervektor sei $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_G')$ mit $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{j,df_j})'$. Das Group Lasso-Funktional hat dann die Darstellung:

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{G} \sqrt{df_j}||\boldsymbol{\alpha}_j||_2 = \sum_{j=1}^{G} \sqrt{df_j}(\alpha_{j1}^2 + \cdots + \alpha_{j,df_j}^2)^{1/2} \tag{4.10}$$

Das Group Lasso wendet die ℓ_2-Norm auf die j-te Koeffizientengruppe an, mit dem Ziel entweder $\hat{\boldsymbol{\alpha}}_j = \mathbf{0}$ oder $\alpha_{js} \neq 0,\ \forall\ s = 1, \ldots, df_j$ zu erreichen. Der Faktor $\sqrt{df_j}$ weist dabei unterschiedlichen Koeffizientengruppen unterschiedliche Gewichte anhand der Größe der Gruppe zu.

Die Ähnlichkeit des Group Lasso zu einer Variablenselektion im sequentiellen Logit-Modell besteht darin, Gruppen von Koeffizienten gemeinsam aus dem Modell zu entfernen oder gemeinsam beizubehalten. Allerdings besteht ein Unterschied in der Intention des jeweilgen Penalisierungsansatzes. Für das Group Lasso werden eine univariate Zielgröße und vornehmlich kategoriale Prädiktoren angenommen. Für das sequentielle Logit-Modell ist die Zielgröße eine kategoriale Variable, die für die Koeffizientengruppen ursächlich ist. Eine Koeffizientengruppe tritt im sequentiellen Modell auf, da metrische oder binäre Prädiktoren mit jeweils einem Koeffizienten pro Zielgrößenkategorie in das Modell eingehen. Die vektorielle Formulierung der Koeffizientengruppe $\boldsymbol{\beta}_{\bullet j} = (\beta_{1j}, \ldots, \beta_{qj})'$ eines Prädiktors j ermöglicht es, das Group Lasso-Funktional für eine korrespondierende Penalisierung im sequentiellen Logit-Modell umzuformulieren:

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{df_j} ||\boldsymbol{\beta}_{\bullet j}||_2 \overset{df_j = df}{=} \sqrt{df} \sum_{j=1}^{p} (\beta_{1,j}^2 + \cdots + \beta_{q,j}^2)^{1/2} \tag{4.11}$$

Da die Größe jeder Koeffizientengruppe durch die Anzahl der Zielgrößenkategorien q bestimmt wird, sind die, denen des Group Lasso entsprechenden Gewichte für jede Koeffizientengruppe, unter obigen Annahmen, gleich. Diese Formulierung entspricht der Basisvariante, des von Tutz, Pößnecker & Uhlmann (2012) bezeichneten CATS-Lasso.

Bisher wurden für die Penalisierungsüberlegungen im sequentiellen Logit-Modell ausschließlich metrische und binäre Kovariablen berücksichtigt, die mit jeweils einem Koeffizienten pro Zielgrößenkategorie in das Modell einfließen. Werden zusätzlich auch **kategoriale Prädiktoren** berücksichtigt, ist es notwendig, neben dem für die multivariate Zielgröße modifizierten Group Lasso, zusätzlich auf die ursprüngliche Intention des Group Lasso zurückzugreifen. Ein kategorialer Prädiktor x_j trete mit ℓ_j Kategorien auf, somit mit $m_j = \ell_j - 1$ Koeffizienten für jede der q Kategorien der Zielgröße. Sei $\boldsymbol{\beta}_{rj_\bullet} = (\beta_{rj_1}, \ldots, \beta_{rj_{m_j}})'$ der Koeffizientenvektor dieses kategorialen Prädiktors für Kategorie r. Um den kategorialen Prädiktor x_j vollständig aus dem Modell zu entfernen, müssen alle m_j Koeffizienten für alle der q Zielgrößenkategorien simultan auf null gesetzt werden. Der Vektor dieser Parametergruppe sei durch $\boldsymbol{\beta}_{\bullet j\bullet} = (\boldsymbol{\beta}'_{1j_\bullet}, \ldots, \boldsymbol{\beta}'_{qj_\bullet})'$ gegeben. Für einen metrischen Prädiktor sei $m_j = 1$. Das erweiterte Group Lasso-Funktional für die Anwendung auf das sequentielle Logit-Modell hat dann die Form:

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \phi_j ||\boldsymbol{\beta}_{\bullet j\bullet}||_2 \tag{4.12}$$

Dabei bezeichne $\phi_j = \sqrt{q \cdot m_j}$ das Penalisierungsgewicht der einzelnen Koeffizientengruppen. Diese sind nun nicht mehr identisch für alle j Koeffizientengruppen.

4.2.4 Sparse Group Lasso

Das Group Lasso ermöglicht ein sparsames Modell, indem Parametergruppen vollständig selektiert werden. Wird eine Gruppe in das Modell aufgenommen, dann besitzen alle ihrer Koeffizienten einen von null verschiedenen Wert. Betrachtet man ein Regressionsszenario mit einer größeren Anzahl kategorialer Prädiktoren, die jeweils wenige Kategorien aufweisen, ist das Group Lasso eine geeignete Methode die wichtigsten dieser Prädiktoren auszuwählen. Ein anderes Szenario enthalte wenige kategoriale Prädiktoren, die allerdings eine Vielzahl von Kategorien aufweisen. Wird mit dem GL ein Prädiktor ausgewählt, gehen alle dessen Kategorien in das Modell ein, auch wenn einzelne Kategorien irrelevant sind. In diesem Fall scheint das klassische Lasso eine geeignetere Selektion zu vollziehen.

Mit dem Sparse Group Lasso schlagen Simon et al. (2012) einen Penalisierungsansatz für univariate Zielgrößen vor, der für oben genannte Szenarien einen Kompromis zwischen einer Sparsamkeit in der Auswahl ganzer Parametergruppen und einer Sparsamkeit in der Auswahl der Parameter innerhalb einer Gruppe findet. Dieser Kompromis wird durch eine Kombination von Group Lasso und klassischem Lasso erreicht, indem das Penalisierungsfunktional die Form

$$J(\boldsymbol{\alpha}) = (1-\nu)\sum_{j=1}^{G}\sqrt{df_j}||\boldsymbol{\alpha}_j||_2 + \nu||\boldsymbol{\alpha}||_1 \tag{4.13}$$

annimmt. Dabei bezeichne $\boldsymbol{\alpha}$ den im vorangegangenen Unterabschnitt 4.2.3 definierten Vektor. Der erste Summand beschreibt oben dargestelltes Group Lasso-Funktional, der zweite Summand das Lasso-Funktional. Durch $\nu \in [0,1]$ wird eine konvexe Kombination von Group Lasso und Lasso erreicht, wobei für $\nu = 1$ der Group Lasso Summand entfällt, für $\nu = 0$ der Lasso Summand. Beide Penalisierungsansätze sind also Spezialfälle des Sparse Group Lasso. Dieser Penalisierungsterm ähnelt zwar dem von Zou & Hastie (2005) vorgeschlagenen *elastic net*, unterscheidet sich allerdings darin, dass $||\cdot||_2$ nicht in $\mathbf{0}$ differenzierbar ist und somit Parametergruppen vollständig auf null geschätzt werden. Es lässt sich zeigen, dass innerhalb der Parametergruppen eine *elastic net* Penalisierung erfolgt, wodurch auch einzelne der Koeffizienten auf null geschätzt werden können. Das Sparse Group Lasso ermöglicht es, die Anzahl der Gruppen zu reduzieren, in denen mindestens ein Koeffizient von null verschieden ist (Sparsamkeit in den Gruppen) und die Anzahl der von null verschiedenen Koeffizienten innerhalb einer ausgewählten Gruppe zu reduzieren (Sparsamkeit innerhalb einer Gruppe). Die Effektivität und Effizienz des Sparse Group Lasso im Vergleich zu Group Lasso und Lasso zeigen Friedman et al. (2010) in einem Simulationsbeispiel. Eine Implementierung in der Statistik-Software R findet sich im Package SGL.

Ähnlich der Anwendung des Group Lasso, lässt sich auch das Sparse Group Lasso für das sequentielle Logit-Modell adaptieren. Dies ist zum Beispiel in einem Szenario sinnvoll, in dem die multivariate Zielgröße eine große Anzahl von Kategorien aufweist, womit jeder selektierte Prädiktor von null verschiedene Werte für jeden Koeffizienten seiner Gruppe erhält. Unter der für das Group Lasso genannten Argumentation, dass eine Gruppierung der Parameter anhand der Kategorien der Zielgröße stattfindet, stellt sich das angepasste Funktional, ohne kategoriale

Prädiktoren mit mehr als zwei Kategorien, dar als:

$$J(\boldsymbol{\beta}) = (1-\nu)\sum_{j=1}^{p}\sqrt{df_j}||\boldsymbol{\beta}_{\bullet j}||_2 + \nu||\boldsymbol{\beta}||_1 \tag{4.14}$$

Das Funktional der Erweiterung des Sparse Group Lasso auf kategoriale Prädiktoren hat die Form:

$$J(\boldsymbol{\beta}) = (1-\nu)\sum_{j=1}^{p}\sqrt{q\cdot df_j}||\boldsymbol{\beta}_{\bullet j\bullet}||_2 + \nu||\boldsymbol{\beta}||_1 \tag{4.15}$$

$\boldsymbol{\beta}$ bezeichne wiederum den vollständigen Koeffizientenvektor, mit Koeffizientengruppen für kategoriale Prädiktoren.

Sowohl Group Lasso, als auch Sparse Group Lasso, dienen in ihren Varianten für multikategoriale Responsevariablen einer adäquaten Variablenselektion im sequentiellen Logit-Modell.

4.3 Verbesserung der Variablenselektion

4.3.1 Adaptives Lasso

Das Lasso-Verfahren kann in bestimmten Szenarien zu einer inkonsistenten Variablenselektion führen, das heißt asymptotisch wird nicht die richtige Teilmenge an Koeffizienten selektiert. Eine Selektionsprozedur wird als konsistent bezeichnet, wenn asymptotisch die richtigen Koeffizienten in das Modell aufgenommen werden, alle übrigen auf null geschätzt werden. Dies wird als Orakel-Eigenschaft bezeichnet. Um diese wünschenswerte Eigenschaft dem Lasso zugänglich zu machen, schlägt Zou (2006) mit dem *adaptive Lasso* eine Korrektur für den Lasso-Strafterm vor, indem Gewichte w_j auf die einzelnen Koeffizienten gelegt werden:

$$J(\boldsymbol{\alpha}) = \sum_{r=1}^{p} w_j||\alpha_j||_1 \tag{4.16}$$

$\boldsymbol{\alpha}$ bezeichne wiederum den einfachen Koeffizientenvektor. Die Variablenselektion des Lasso lässt sich dadurch verbessern, dass auf schwache Prädiktoren stärkeres Penalisierungsgewicht gelegt wird, wohingegen der Grad der Bestrafung für stärkere Prädiktoren gering sein sollte. Eine Wahl für geeignete Gewichte $w_j = 1/|\tilde{\alpha}_j|^{\delta}$ für festes $\delta > 0$, kann mittels der ML-Schätzer $\hat{\alpha}^{ML}$ für $\tilde{\alpha}_j = \hat{\alpha}_j^{ML}$ erreicht werden. Die Verwendung adaptiver Gewichte kann sowohl die Selektionsfähigkeit des Lasso verbessern, als auch die Prädiktionsgenauigkeit des resultierenden Modells. Derartige Gewichte lassen sich geeignet adaptiert, für einen Penalisierungsterm im sequentiellen Logit-Modell verwenden, indem beispielsweise die Gewichte $\sqrt{df_j}$ aus Gleichung 4.14 durch $\sqrt{df_j}/||\hat{\boldsymbol{\beta}}_j^{ML}||_2$ ersetzt werden. Dabei bezeichne $\hat{\boldsymbol{\beta}}_{\bullet j}^{ML}$ den ML-Schätzer der jeweiligen Koeffizientengruppe.

4.3.2 Refitting

Eine weitere Möglichkeit, die Selektionseigenschaften und die Güte der geschätzten Parameter zu verbessern, besteht darin, dass die penalisierte Variablenselek-

tion und die letztendliche Schätzung der Parameter voneinander entkoppelt werden. Dieses Verfahren verwenden Efron et al. (2004) unter der Bezeichnung *LARS-OLS hybrid* und Candes & Tao (2007) als *Gauss-Dantzig-Selector*. Im ersten Schritt wird ein gewählter Penalisierungsansatz ausschließlich dazu verwendet, Variablen zu selektieren. Im zweiten Schritt findet unter Verwendung dieser selektierten Variablen eine erneute Modellanpassung (*Refitting*) statt. Wird im zweiten Schritt unpenalisiert geschätzt, spielt der Bias der den Parametern durch die penalisierte Schätzung des ersten Schritts auferlegt wird keine Rolle für die letztlich geschätzten Parameter. Im ersten Schritt kann somit eine stärkere Variablenselektion mittels eines größeren Penalisierungsparameters λ duchgeführt werden. Im Fall, dass der Refit ebenfalls penalisiert durchgeführt wird, wird für die bereits selektierten Parametern eine weitere Selektion durchgeführt. Es können wiederum Variablen mit schwachen Effekten aus dem Modell entfernt werden, gleichzeitig erhöht sich allerdings die Gefahr relevante Variablen auf null zu schätzen. Es lässt sich beobachten, dass im Vergleich zu Schätzungen ohne Refit, stärkere Penalisierungen durchgeführt werden. (Vgl. Tutz, Pößnecker & Uhlmann (2012))

4.4 Zusammenfassung

Mit Regularisierungsansätzen, die auf einer Penalisierung der Log-Likelihoodfunktion beruhen, wurden in diesem Abschnitt Verfahren beschrieben, die es ermöglichen in ordinalen Regressionsmodellen aus einer Vielzahl an Kovariableneffekten diejenigen metrischen und kategorialen Prädiktoren zu selektieren, die die stärksten Effekte aufweisen. Dies ermöglicht zum einen eine verbesserte Interpretierbarkeit des Modells, zum anderen werden durch die Verringerung ihrer Werte zwar verzerrte Koeffizienten erzeugt, allerdings können diese eine geringere Varianz aufweisen, als die korrespondierenden ML-Schätzer. Im Gegensatz zur ML-Schätzung können auch dann hochdimensionale Modelle geschätzt werden, wenn die Anzahl verfügbarer Beobachtungen geringer ist, als die Anzahl zu schätzender Parameter. Während Sparse Group Lasso und Group Lasso Gruppen von Koeffizienten simultan selektieren und somit den Anforderungen eines kategorialen Regressionsmodells gerecht werden, ist diese Eigenschaft dem klassischen Lasso vorenthalten. Diese drei Verfahren werden in der Simulationsstudie in Kapitel 5 miteinander, hinsichtlich ihrer Selektionseigenschaften und der Güteeigenschaften ihrer geschätzten Koeffizienten, verglichen. Zusätzlich werden diese drei Verfahren jeweils mit einer adaptiven ML-Gewichtung oder einem penalisierten Refit dem unmodifizierten Ansatz gegenübergestellt. Aufgrund der vollständig fehlenden Selektionseigenschaft, wird die Ridge Regression für diese Simulationsstudie nicht berücksichtigt.

Teil II

Simulationen und Anwendungsbeispiele

Kapitel 5

Simulationsstudie

In diesem Kapitel wird in mehreren Szenarien die Anwendung verschiedener Penalisierungsansätze auf das sequentielle Logit-Modell miteinander verglichen. Es werden Situation betrachtet, in denen mehr Beobachtungen, als zu schätzende Parameter vorhanden sind, sowie der umgekehrte Fall. Weiterhin wird ein Szenario untersucht, in dem die Effekte des Modells im Vergleich zu den wahren Effekten fehlspezifizert sind. Die verschiedenen Szenarien, sowie der Vergleich der Penalisierungsansätze hinsichtlich ihrer Selektionsfähigkeit, Schätz- und Prognosegüte, werden in Abschnit 5.1 vorbereitet. Die Ergebnisse der einzelnen Szenarien, werden in Abschnitt 5.2 dargestellt und erörtert und in Abschnitt 5.3 zusammengefasst.

5.1 Simulationssetup

5.1.1 Modell und Szenarien

Modell

Aus den in Abschnitt 3.5 dargestellten Gründen, entsprechend der Aufgabenstellung, wird sich für die Simulationsstudie und die Datenauswertungen auf das sequentielle Logit-Modell beschränkt. Die Modellgleichung ergibt sich aus dem allgemeinen sequentiellen Modell unter Verwendung der logistischen Linkfunktion, wie in Abschnitt 3.3 hergeleitet, als:

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i, \mathbf{z}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r + \mathbf{z}_i'\boldsymbol{\gamma})}{1 + \exp(\beta_{r0} + \mathbf{x}_i'\boldsymbol{\beta}_r + \mathbf{z}_i'\boldsymbol{\gamma})}, \quad r = 1, ..., q \tag{5.1}$$

Der kategoriespezifische lineare Prädiktor setzt sich in allgemeiner Variante, als Kombination aus kategoriespezifischen Effekten $\boldsymbol{\beta}_r$ der kategoriespezifischen Kovariablen $\mathbf{x}_i$ und globalen Effekten $\boldsymbol{\gamma}$ der globalen Kovariablen $\mathbf{z}_i$ zusammen. Je nach Szenario wird der lineare Prädiktor auf rein kategoriespezifische, rein globale oder auf eine Mischung der Effekte variiert. In den Szenarien werden ausschließlich Kovariablen betrachtet, die mit einem Koeffizienten pro Kategorie in das Modell eingehen.

Simulationsszenarien

Betrachtet werden vier Simulationsszenarien. Die kleineren Modelle der ersten drei Szenarien besitzen jeweils fünf Responsekategorien und maximal 15 Prädiktoren, das größere Modell des vierten Szenarios besitzt zehn Responsekategorien und 60 Prädiktoren. Um die Variablenselektionsfähigkeit der verschiedenen Penalisierungsansätze beurteilen zu können, werden jedem Modell sowohl relevante, als auch irrelevante Prädiktoren zugrundegelegt. Die Koeffizienten der relevanten globalen Prädiktoren werden zufällig, unabhängig voneinander aus der Menge $\{-3; -2,5; -2; -1,5; -1; -0,5; 0,5; 1; 1,5, 2; 2,5; 3\}$ gezogen. Für kategoriespezifische Koeffizienten besteht zudem die Möglichkeit, den Wert 0 mit einer Wahrscheinlichkeit von 2/14 anzunehmen. Der Koeffizientenwert der Prädiktoren, die keinen Einfluss auf die Zielgröße haben, beträgt stets 0.

Für die Varianz-Kovarianz-Matrix der Prädiktoren, die aus einer multivariaten Gauss-Verteilung gezogen werden, wird eine Equikorrelation von 0,2 oder 0,6 angenommen. Anhand der wahren Koeffizienten und zufälligen Kovariablenwerte, errechnen sich die wahren kategoriespezifischen Auftretenswahrscheinlichkeiten. Diese bilden die Grundlage die Responsekategorien aus einer Multinomialverteilung zu ziehen. Der Vektor der wahren Modellkoeffizienten wird einmal für jedes Modell gezogen, die Anzahl der Simulationsiterationen liegt zwischen 60 und 100.

In **Szenario 1** mit fünf Responsekategorien wird untersucht, welchen Effekt verschiedene Penalisierungsansätze auf die, weiter unten im Text, genannten Vergleichskriterien haben, wenn die Koeffizientenstruktur des Modells fehlspezifiziert ist. Hierzu werden zwei Varianten einer möglichen Fehlspezifikation betrachtet: Für Modell 1.1 wird angenommen, dass die wahren Effekte der Prädiktoren kategoriespezifisch, für Modell 1.2, dass die wahren Effekte der Prädiktoren kategorieunspezifisch (global) auf die Kategorien der Responsevariable wirken. Für diese Modelle werden jeweils zehn aktive (relevante) und fünf inaktive (irrelevante) Prädiktoren angenommen, die mit einer Equikorrelation von 0,2 schwach positiv korreliert sind. Beide Modelleschätzungen werden in $n_{rep} = 80$ Simulationsdurchläufen für je $n = 200$ Beobachtungseinheiten wiederholt. Für das Modell mit wahren kategoriespezifischen Koeffizienten beträgt die Anzahl zu schätzender Parameter mit kategoriespezifischen Intercepts 64, was der Anzahl der wahren Parameter entspricht. Die Anzahl der Parameter des wahren Modells mit globalen Effekten beträgt 19.

Szenario 2 beschreibt für ein Modell mit ebenfalls fünf Responsekategorien eine Situation, in der mehr Beobachtungen vorhanden sind, als zu schätzende Parameter und umgekehrt. In diesem Szenario wird eine wahre Prädiktorstruktur von fünf kategoriespezifischen, fünf globalen und vier irrelevanten Prädiktoren herangezogen. Die Anzahl wahrer Parameter, einschließlich der kategoriespezifischen Intercepts, beträgt somit 45. Da allerdings für die Schätzung nicht bekannt ist, welcher Prädiktor mit globalen Effekten auftritt, werden in diesem Fall ebenso kategoriespezifische Effekte geschätzt, sodass die Anzahl der zu schätzenden Parameter bei 60 liegt. Für den Fall, dass fälschlicherweise für alle Prädiktoren globale Effekte angenommen werden, sind 18 Parameter zu bestimmen. Die Korrelation zwischen den Prädiktoren beträgt 0,2. In Modell 2.1, wird eine datenreiche Situation mit $n = 200$ verfügbaren Beobachtungen angenommen. In Modell 2.2, einer datenarmen Situation, sind $n = 40$ Beobachtungen vorhanden, sodass $p > n$ gilt.

Das Szenario wird für beide Modelle 80 Mal wiederholt.

In **Szenario 3** wird das Prädiktorsetting von Szenario 2 aufgeriffen und verglichen, inwiefern eine Veränderung der Beobachtungsanzahlen von 200 im vorherigen Szenario auf 1000 in diesem Szenario, sowie einer Veränderung der Korrelation der Prädiktoren Auswirkungen auf die Penalisierungsansätze hat. Hierfür wird in Modell 3.1 eine schwache Equikorrelation von 0,2 und in Modell 3.2 eine starke von 0,6 angenommen. Die Anzahl der Simulationswiederholungen beträgt für beide Modelle 100.

Das Modell in **Szenario 4** besitzt zehn Responsekategorien und 30 Prädiktoren mit kategoriespezifischen Effekten, 10 Prädiktoren mit globalen Effekten und 20 irrelevante Prädiktoren mit einer Equikorrelation von 0,6. Die Anzahl wahrer Effekte beträgt einschließlich kategoriespezifischer Intercepts 469. Werden kategoriespezifische Effekte geschätzt, sind 549 Koeffizienten zu schätzen. Werden globale Effekte geschätzt, sind 69 Koeffizienten zu bestimmen. Mit der Wahl von 500 verfügbaren Beobachtungseinheiten, sind in der Variante kategoriespezifisch geschätzer Koeffizienten mehr Koeffizienten zu schätzen, als Beobachtungen vorhanden sind. Das Szenario wird 60 Mal wiederholt.

Penalisierungsansätze

Es werden für alle Szenarien folgende **unpenalisierte und penalisierte ML-Schätzungen** durchgeführt: Eine unpenalisierte ML-Schätzung mit kategoriespezifischen Effekten (*ML*), eine unpenalisierte ML-Schätzung mit globalen Effekten (*ML glob*), eine Group Lasso penalisierte ML-Schätzung in der klassischen Variante (*GL*), mit adaptiven ML-Gewichten (*ada GL*), mit Refit (*rf GL*), eine Lasso-penalisierte ML-Schätzung in den eben genannten Varianten (*Lasso, ada Lasso, rf Lasso*), eine Sparse Group Lasso-penalisierte ML-Schätzung (*SGL, ada SGL, rf SGL*), sowie eine Lasso-penalisierte ML-Schätzung globaler Effekte (*glob, ada glob, rf glob*). Sowohl Group Lasso, als auch Sparse Group Lasso benötigen kategoriespezifische Kovariableneffekte, da sie sich sonst auf das klassische Lasso für globale Effekte vereinfachen. Die grafischen Darstellungen erfolgen in der genannten Reihenfolge mit den in Klammern angegebenen Kurzformen.

Die Simulationen werden in der Software R (R Developement Core Team (2012)) mit Hilfe des in Tutz, Pößnecker & Uhlman (2012) verwendeten *Fast Iterative Shrinkage-Thresholding Algorithmus* (FISTA) auf der Grundlage des Algorithmus von Beck & Teboulle (2009) durchgeführt. Dieser Algorithmus wurde für die Koeffizientenschätzung des sequentielle Modells adaptiert und zusammen mit Simulationsfunktionen, von Wolfgang Pößnecker für diese Auswertungen zur Verfügung gestellt.

5.1.2 Vergleichsmethoden

Die verschiedenenen Penalisierungsansätze, die innerhalb eines Simulations- szenarios Verwendung finden, werden anhand von vier Methoden miteinander verglichen. Hierfür werden die Güte, der durch die jeweiligen Penalisierungsansätze gewonnenen Schätzer, die Genauigkeit bei der Auswahl der Koeffizienten/-Variablenselektion und die Prognosequalität der geschätzten Modelle bestimmt.

Um die Güte der Schätzer miteinander zu vergleichen, wird der **Squared Error der geschätzten Koeffizienten** für jede Simulationsiteration berechnet. Es

bezeichne $\hat{\boldsymbol{\theta}}^{(s)}$ den geschätzten Koeffizientenvektor der s-ten Replikation eines Modellszenarios. Für jede Replikation wird der quadratische Fehler $(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*)'(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*)/d$ berechnet. Dabei bezeichne $\boldsymbol{\theta}^*$ den wahren Koeffizientenvektor und d die Anzahl zu schätzender Parameter des jeweiligen Szenarios. Die quadratischen Fehler jedes betrachteten Penalisierungsansatzes werden mit Hilfe von Boxplots, die die Variabilität und das Auftreten von Ausreißern ersichtlich machen, dargestellt. Diese Abbildungen werden mit *MSE of coefficients* betitelt. Für eine approximative Berechnung des MSE der Koeffizienten $MSE_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)/d]$ lässt sich das arithmetische Mittel über alle n_{rep} Iterationen berechnen:

$$MSE_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}) = \frac{1}{n_{rep}} \sum_{s=1}^{n_{rep}} (\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*)'(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*)/d \tag{5.2}$$

Mit Hilfe der geschätzten Modellkoeffizienten, berechnet sich als weiteres Gütekriterium der Penalisierungsansätze der **Mean Squared Error** der geschätzten kategoriespezifischen **Wahrscheinlichkeiten** für eine Iteration, als:

$$MSE_{\boldsymbol{\pi}}(\hat{\boldsymbol{\pi}}^{(s)}) = \frac{1}{n}\frac{1}{q} \sum_{i=1}^{n} \sum_{r=1}^{q} (\hat{\pi}_{ir} - \pi_{ir})^2 \tag{5.3}$$

Mit Gleichung 3.17 wird, unter Verwendung der logistischen Verteilungsfunktion, π_{ir} mit den wahren Modellkoeffizienten und $\hat{\pi}_{ir}$ mit den geschätzten Koeffizienten berechnet. Die Wahrscheinlichkeiten-MSE werden ebenfalls durch Boxplots mit dem Titel *MSE of probabilities* abgebildet.

Ein drittes Vergleichskriterium betrachtet die Prognosegüte der verschiedenen Ansätze. Zur Beurteilung der Prognosegüte wird die **Prädiktionsdevianz** bestimmt. Hierfür werden aus dem wahren Modell $3 \cdot n$ neue Beobachtungen gezogen und die Diskrepanz zwischen diesen zusätzlichen Beobachtungen und den durch das Modell vorhergesagten Werten bestimmt. Die Prädiktionsdevianzen für jede Iteration, werden ebenfalls in Boxplots unter dem Titel *Predictive Deviance* dargestellt.

Um die Güte der Variablenselektion eines Penalisierungsansatzes zu beurteilen, werden zwei Relationen betrachtet. Zum einen wird die Anzahl inaktiver Variablen, die fälschlicherweise als aktiv geschätzt wurden, also mindestens einen von null verschiedenen Koeffizienten innerhalb der Koeffizientengruppe dieser Variable erhalten haben, ins Verhältnis zur Gesamtanzahl der tatsächlich inaktiven Variablen gesetzt. Dieses Verhältnis wird als **Falsch-Positiv-Rate** (FPR) bezeichnet. Die andere Relation berücksichtigt die relevanten (aktiven) Variablen, die fälschlicherweise als irrelevant für das Modell erachtet wurden, deren Koeffizientengruppe demnach vollständig auf null geschätzt wurde. Das Verhältnis der Anzahl fehleingeschätzter relevanter Variablen zur Gesamtzahl der tatsächlich relevanten Variablen wird als **Falsch-Negativ-Rate** (FNR) bezeichnet. Die beiden Relationen werden als Balken, jeweils für einen Penalisierungsansatz nebeneinander, dargestellt.

5.2 Auswertung der Szenarien

5.2.1 Szenario 1

Die Simualtionsergebnisse für Szenario 1 sind in Abbildung 5.1 dargestellt. In der linken Spalte der Vergleich der Penalisierungsansätze für Modell 1.1 mit wahren kategoriespezifischen Effekte, in der rechten Spalte für Modell 1.2 mit wahren globalen Effekten. In der ersten Zeile werden die Koeffizienten-MSE, in der zweiten Zeile die Wahrscheinlichkeiten-MSE, in der dritten Zeile die prädiktiven Devianzen und in der vierten Zeile die Selektionseigenschaften mittels Falsch-Positiv-Raten (grau) und Falsch-Negativ-Raten (schwarz) miteinander verglichen.

Zunächst werden die Ergebnisse der Koeffizientenschätzung und Variablenselektion des Modells 1.1 in der **linken Spalte** analysiert. Auffällig in den Boxplots der Koeffizienten-MSE ist die große Variabilität der geschätzten Effekte für die unpenalisierte kategoriespezifische ML-Schätzung (*ML*) sowie für Lasso-Ansätze (*Lasso, ada Lasso, rf Lasso*). Dies deutet zum einen auf die Instabilität der ML-Schätzer hin, zum anderen darauf, dass unter Verwendung des klassischen Lasso (einschließlich adaptiver und Refit-Modifikationen) dieser Instabilität, durch Penalisierung der einzelnen ungruppierten Effekte, nicht entgegengewirkt werden kann, sondern im Gegenteil die Instabilität der Effekte verstärkt wird.[1] Eine ähnliche, wenn auch geringere Instabilität tritt für *rf GL* und *rf SGL* auf, da beide Ansätze einen unpenalisierten ML-Refit erhalten. Erwartungsgemäß gelingt Group Lasso und Sparse Group Lasso, ohne und mit adaptiver Gewichtung, die beste MSE-Performance. Auffällig sind des Weiteren der unpenalisierte und die penalisierten globalen Ansätze (*ML glob, glob, ada glob, rf glob*) mit einem ähnlichen Medianniveau und einer sehr geringen Variabilität. Deren Wahrscheinlichkeiten-MSE weisen hingegen eine deutlich höhere Abweichung, sowie größere Variabilität auf, sodass die Schätzung globaler Effekte einen starken Einfluss auf die daraus berechneten Wahrscheinlichkeiten hat. Für eine geeignete Darstellung, wurde der mittlere Teil der Achse entfernt, ohne dabei den Interpretationsgehalt der Abbildung zu beeinträchtigen. Für die Lasso-Ansätze ergibt sich hingegen ein umgekehrtes Bild. Trotz der Variabilität in der Schätzung der einzelnen Koeffizienten, reduziert sich die Variabilität der daraus berechneten Wahrscheinlichkeiten. Gleiches gilt für den kategoriespezifische ML-Schätzer. GL- und SGL-Varianten schneiden ebenfalls für den Wahrscheinlichkeiten-MSE am Besten ab, wobei ohne und mit adaptiver Gewichtung ähnliche Ergebnisse erzielt werden, geringfügig vor den Ergebnissen mit Refit.

[1] Vereinzelt werden Boxen in der Graphik bis auf den Median nicht vollständig berücksichtigt, da dies die Darstellung kürzerer Boxen vereinfacht.

(a) wahre kategoriespezifische Effekte **(b)** wahre globale Effekte

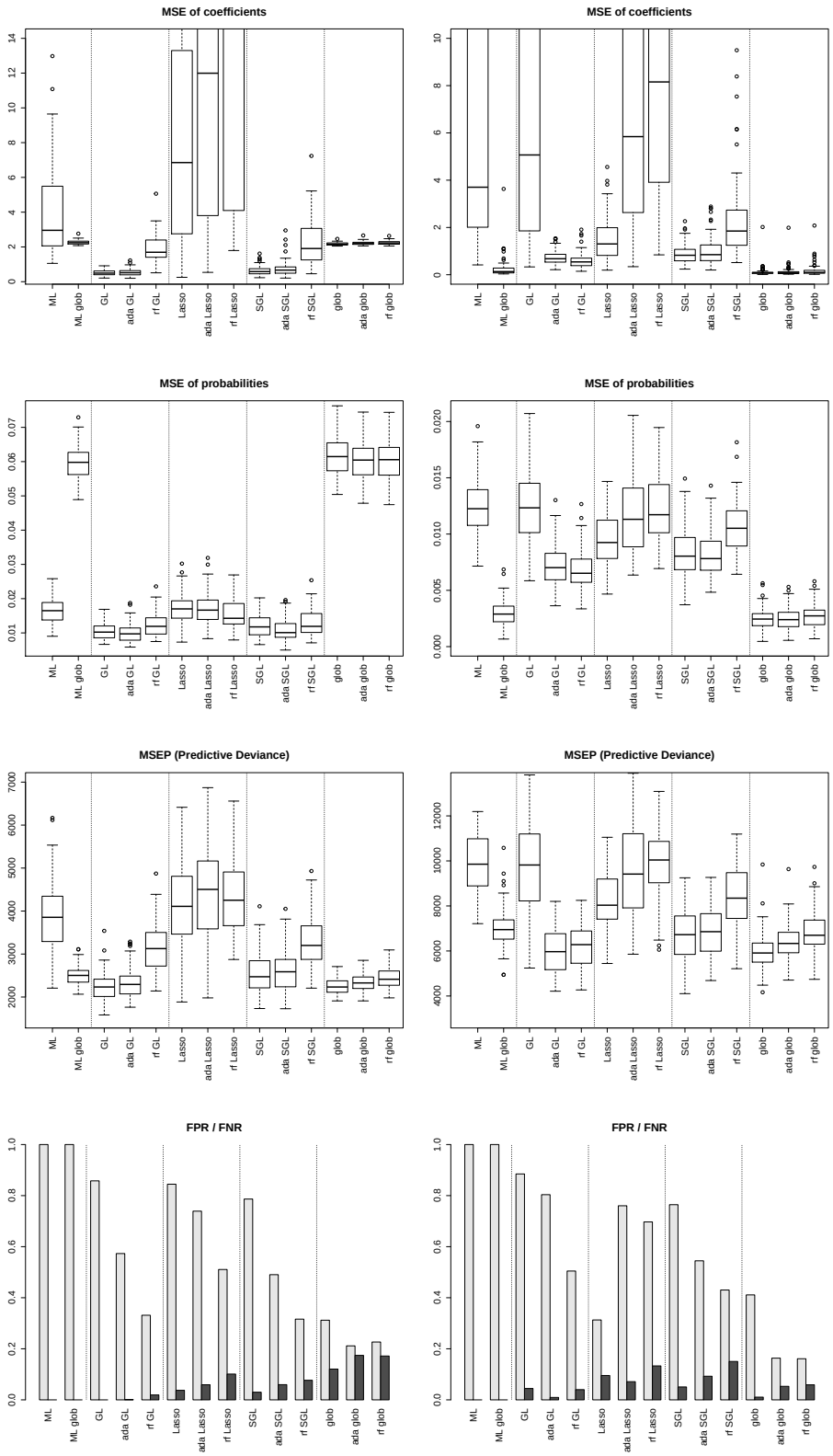

Abbildung 5.1: Ergebnisse Szenario 1: 5 Responsekategorien; 15 kategoriespezifische Prädiktoren (linke Spalte), 15 globale Prädiktoren (rechte Spalte); jeweils 200 Beobachtungen, Korrelation 0,2 und 80 Iterationen.

Für die prädiktive Devianz besitzen globale Effekte leichte Vorteile vor Group Lasso und Sparse Group Lasso (mit und ohne Gewichtung). Ebenso gelingt es den penalisierten globalen Effekten, die geringsten FPR aufzuweisen. Dieses geringe Risiko fälschlicherweise irrelevante Variablen einzubeziehen, geht mit einem erhöhten Risiko einher, relevante Variablen fälschlicherweise zu entfernen (hohe FNR). Wie bereits im vorangegangenen Kapitel erörtert, eignen sich unpenalisierte Effekte nicht zu einer Variablenselektion, da alle Koeffizienten einen von null verschiedenen Effekt erhalten. Dies betrifft ebenso alle irrelevanten Koeffizienten, sodass die FPR einen Wert von eins annimmt. Äquivalent dazu, kann kein relevanter Effekt fälschlicherweise als irrelevant erachtet werden, sodass die FNR einen Wert von Null annimmt. Generell führt eine adaptive/Refit-Optimierung zu einem verbesserten Ausschluss irrelevanter Variablen. Gleichzeitig steigt allerdings die Gefahr, relevante Effekte fälschlicherweise auszuschließen. Somit weisen die Refit-Varianten zwar geringste FPR, aber auch höchste FNR auf. Der Anstieg letzterer ist jedoch geringer, als die Reduktion der FPR.

Besitzen die wahren Koeffizienten einen globalen Einfluss auf die Zielgrößenkategorien, wird anhand der **rechten Spalte** der Abbildung 5.1 deutlich, dass in allen Vergleichskriterien die Varianten penalisierter globaler Effekte am besten abschneiden. Adaptive Gewichte und Refit führen zu Boxplots mit geringfügig höherem Medianwert und stärkerer Streuung, besitzen gegenüber der unmodifizierten Variante jedoch deutliche Vorteile hinsichtlich der FPR. Für die kategoriespezifische ML-Schätzung, sowie Lasso-Varianten ergibt sich ein ähnlich instabiles Bild, wie in der linken Spalte. Auffällig ist zudem das sehr instabile Group Lasso, das sich hinsichtlich aller Kriterien nicht eignet. Tendenziell positiv schneidet neben den penalisierten globalen Effekten, Group Lasso in adaptiver und Refit-Variante ab, sowie Sparse Group Lasso mit und ohne adaptive Gewichtung.

Anhand diesen Szenarios wird deutlich, wie stark die Auswirkungen sowohl auf Schätzgüte, Prädiktionsgüte, als auch Selektionsfähigkeit sind, wenn die wahren Effekte fehlspezifiziert in die Modellgleichung aufgenommen werden. Ist die Prognosegüte oder Variablenselektionfähigkeit in einem Modell von Interesse, tendiert die Wahl eines Verfahrens zu penalisierten globalen Effekten, sofern die wahren Koeffizienten entweder vollständig kategoriespezifische oder globale Effekte aufweisen. In Bezug auf die Güte der Schätzung scheint es weniger gravierend zu sein, wahre globale Effekte fälschlicherweise kategoriespezifisch zu schätzen, als wahre kategoriespezifische Effekte als global. Hinsichtlich aller Schätzverfahren tendiert die Variante mit adaptiver Gewichtung dazu, für Schätz- und Prognosegüte die besten Ergebnisse zu erzeugen. In Bezug auf Variablenselektionseigenschaften sind Refit-Varianten zu bevorzugen, sofern für eine stärkere Reduktion der FPR, ein geringfügiger Anstieg der FNR akzeptiert wird. Es wird in diesem Szenario bestätigt, dass eine etwaige kategoriespezifische ML-Schätzung ungeeignet ist, deren Instabilität aufgrund der ungruppierten Penalisierung auch nicht mit Varianten des klassischen Lasso behoben werden kann.

Die Interpretation der beiden Modelle ist darauf beschränkt, dass für alle wahren Koeffizienten eine einheitliche Struktur angenommen wurde. Szenarien, die ein Mischung globaler und kategoriespezifischer Effekte annehmen, werden nachfolgend betrachtet.

5.2.2 Szenario 2

Die Ergebnisse des Szenario 2 zeigt Abbildung 5.2. Wie bereits in der Szenariobeschreibung deutlich wird, setzen sich die wahren Effekte aus fünf Variablen mit kategoriespezifischen und fünf Variablen mit globalen Effekten zusammen. Es wird davon ausgegangen, dass kein Hinweis für eine korrekte Spezifikation der Variableneffekte existiert, sodass die Effekte aller Variablen einheitlich innerhalb eines Penalisierungsansatzes behandelt werden. Entweder werden alle Effekte kategoriespezifisch oder global geschätzt. Die **linke Spalte**, die Modell 2.1 mit 200 verfügbaren Beobachtungen beschreibt, zeigt ähnliche Resultate, wie Modell 1.1 mit ausschließlich kategoriespezifischen wahren Effekten (linke Spalte in Abbildung 5.1). Dies impliziert, dass die Ergebnisse der globalen Penalisierungsansätze hinsichtlich aller Vergleichskriterien durch die fehlspezifizierten kategoriespezifischen Effekte dominiert werden. Allerdings fällt die Differenz zu den kategoriespezifischen Ansätzen, durch die vorhandenen wahren globalen Effekte, geringer aus. Es wird hingegen kein Effekt aus den fehlspezifizierten globalen Effekten in kategoriespezifisch gruppierten Ansätzen sichtbar.

Die **rechte Spalte** gibt die Vergleichskriterien bezüglich der mit $n = 40$ beobachtungsarmen Modellsituation wieder. Da im Fall einer ML-Schätzung mit kategoriespezifischen Effekten weniger Beobachtungen, als zu schätzende Parameter vorhanden sind, wird, um einen ML-Schätzer zu generieren, eine geringfügige Ridgekorrektur hinzugefügt. Während die Koeffizienten-MSE für *ML* und *rf Lasso* erfahrungsgemäß instabil sind, findet für Lasso und adaptives Lasso eine Stabilisierung statt. Auffällig sind die penalisierten globalen Verfahren, die eine Vielzahl extremer Ausreißer aufweisen, während der Median auf dem Niveau der penalisierten kategoriespezifischen Verfahren liegt. Hinsichtlich der Wahrscheinlichkeiten-MSE, sowie der prädiktiven Devianzen, wird eine große Variablilität durch die langezogenen Boxen deutlich. Für beide Vergleichskriterien schneiden in dieser datenarmen Modellsituation *GL*, *SGL*, sowie beide adaptive Varianten am besten ab. Die Variabilität in der Schätzung und Prädiktion spiegelt sich ebenfalls in der Variablenselektionsfähigkeit wieder. Zwar sind die Falsch-Positiv-Raten für manche der Ansätze geringer als im datenreichen Modell, allerdings wird der Wert 0,2 nicht unterschritten. Die Falsch-Negativ-Raten liegen zwischen ca. 0,2 und ca. 0,5 und somit deutlich höher als in allen bisherigen Modellen. Ein Kompromis zwischen FPR und FNR wird für die adaptiven kategoriespezifischen Varianten *ada GL* und *ada SGL* gefunden. Es lässt sich festhalten, dass in diesem Szenario eines eher kleinen Modells und einer datenarmen Schätzsituation in allen Vergleichskriterien *(ada) GL* und *(ada) SGL* die beste Performance aufweisen. Allerdings lässt sich mit keinem Verfahren mehr eine zufriedenstellende Variablenselektion durchführen.

(a) 200 Beobachtungen **(b)** 40 Beobachtungen

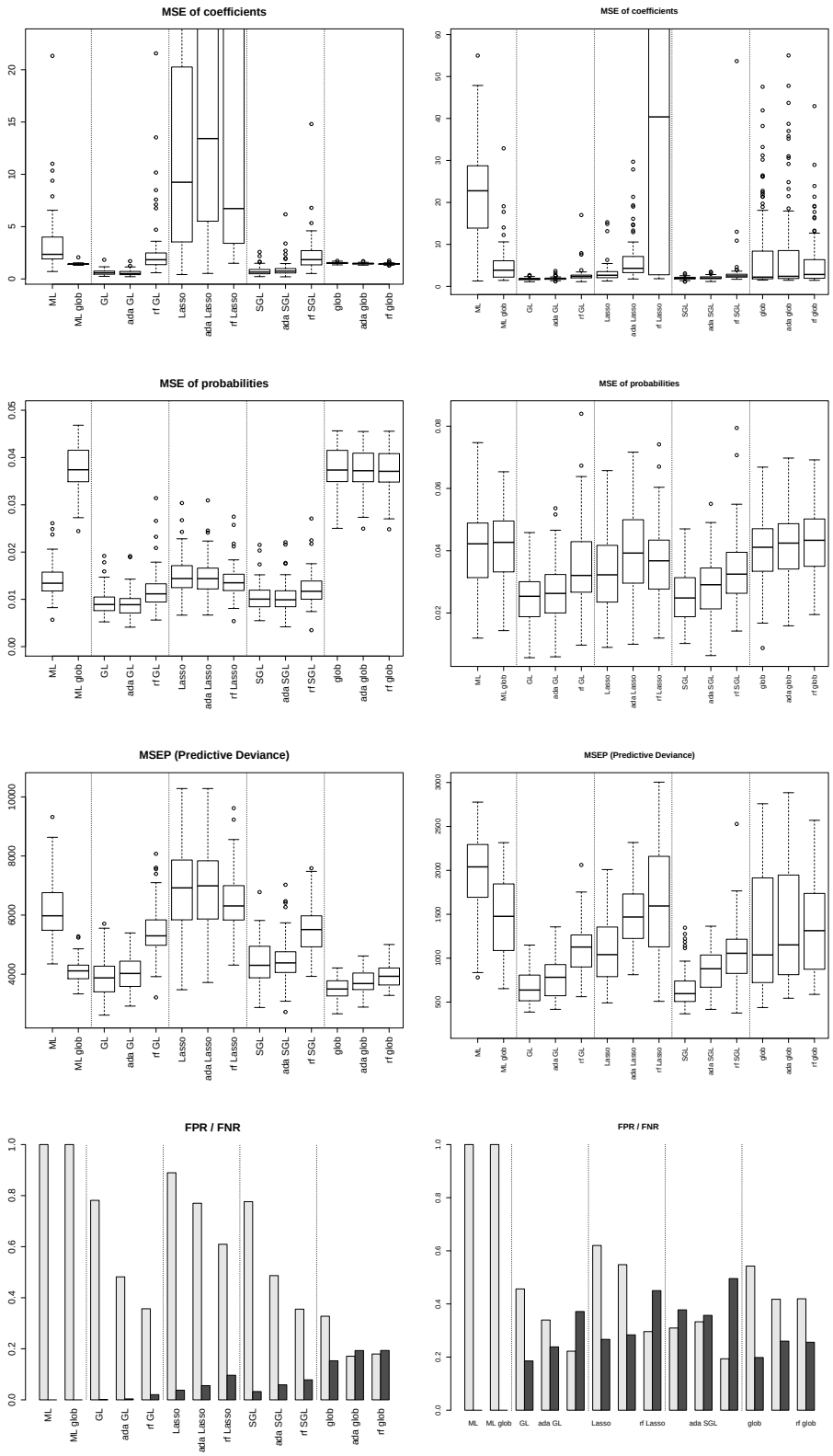

Abbildung 5.2: Ergebnisse Szenario 2: 5 Responsekategorien; 5 kategoriespezifische, 5 globale, 5 irrelevante Prädiktoren; 200 Beobachtungen (linke Spalte), 40 Beobachtungen (rechte Spalte); jeweils Korrelation 0,2; 80 Iterationen.

5.2.3 Szenario 3

Für Szenario 3 wurde im Vergleich zu vorherigem Setting die Zahl der Beobachtungseinheiten auf 1000 erhöht. In der **rechten Spalte** von Abbildung 5.3 sind die Ergebnisse für Modell 3.1 mit einer Equikorrelation von 0,2 dargestellt. Es wird deutlich, dass aufgrund der deutlich verbesserten Datengrundlage die Variabilität der Koeffizienten-MSE in den Lasso-Varianten vollständig verschwindet. Die Ansätze, deren finales Modell mit Hilfe einer ML-Schätzung erzeugt wird (*ML, rf GL, rf Lasso, rf SGL*) weisen die bekannte Variablilität mit extremen Ausreißern auf. Tendenziell schneiden die gruppiert-penalisierten Ansätze mit adaptiver Gewichtung am besten ab. In dieser datenreichen Situation wird der Unterschied globaler zu kategoriespezifischen Verfahren hinsichtlich der Koeffizienten- und Wahrscheinlichkeiten-MSE deutlich. Allerdings wird für penalisierte globale Effekte (*glob*) eine geringfügig bessere prädiktive Devianz erzeugt, gefolgt von *GL*, *ada GL* und *SGL*. Ansätze mit Refit-Modifikation schneiden zwar hinsichtlich Koeffizientenschätzung und prädiktiver Devianz am schlechtesten ab, erreichen allerdings mit dem *rf GL* (FPR bei ca. 0,1) und *rf SGL* eine sehr starke Performance hinsichtlich ihrer Selektionsfähigkeit.

Über die Situation, dass unter sonst gleichen Bedingungen, eine Korrelation von 0.6 zwischen den Prädiktoren vorliegt, gibt die **rechte Spalte** von Abbildung 5.3 Auskunft. Bezüglich aller Vergleichskriterien ergeben sich bei erhöhter Korrelation keine Veränderungen gegenüber der Situation mit schwächerer Korrelation. Einzig auffällig ist eine erhöhte Variabilität der ML-Schätzer und der Refit-Schätzungen.

5.2.4 Szenario 4

In diesem Szenario mit zehn Responsekategorien, 60 Prädiktoren und einer Beobachtungsanzahl von 500, tritt für die Schätzung kategoriespezifischer Effekte erneut eine Situation auf, in der mehr Parameter zu bestimmen, als Beobachtunseinheiten vorhanden sind. Die Ergebnisse sind in Abbildung 5.4 dargestellt. Für die Koeffizienten-MSE der ML-Schätzungen, die mit Hilfe einer Ridgekorrektur erzeugt wurden, sowie der Refit-Varianten tritt die bekannt hohe Variablilität auf. Dies gilt ebenfalls für alle Lasso-Varianten. Die Box des Lasso mit Refit konnte hierbei nicht mehr in die Grafik eingefügt werden, ohne den Interpretationsgehalt aller anderen Boxen erheblich einzuschränken. Die globalen Schätzvarianten, ebenso wie die kategoriespezifischen mit und ohne adaptiver Gewichtung weisen eine geringe Variabilität und ein ähnliches Medianniveau auf. Hinsichtlich der Wahrscheinlichkeiten-MSE und der prädiktiven Devianzen werden für GL und SGL mit und ohne adaptiver Gewichtung die besten Resultate erreicht. Globale Penalisierungsansätze erzeugen bzgl. der Prädiktionsgüte vergleichbare Ergebnisse wie GL und SGL in adapitver Variante.

Die Selektionsfähigkeit der verschiedenen Penalisierungsansätze nimmt in diesem Simulationssetting deutlich ab. Zwar werden weniger irrelevante Variablen als relevant erachtet (geringere FPR), allerdings ist das Risiko in fast allen Varianten sehr hoch, relevante Variablen als irrelevant zu schätzen. Dies ist auf den, im Verhältnis zur Anzahl zu schätzender Parameter, geringen Datenumfang und die hohe Anzahl Responsekategorien zurückzuführen. Anders als in Modell 2.1 (40 Beobachtungen, 5 Responsekategorien), wird hier durch adaptives Group Lasso eine verhältnismäßig starke Selektionsfähigkeit erreicht. Die FPR liegt bei ca. 25 %, die FNR unter 10 %.

(a) Korrelation 0.2 **(b)** Korrelation 0.6

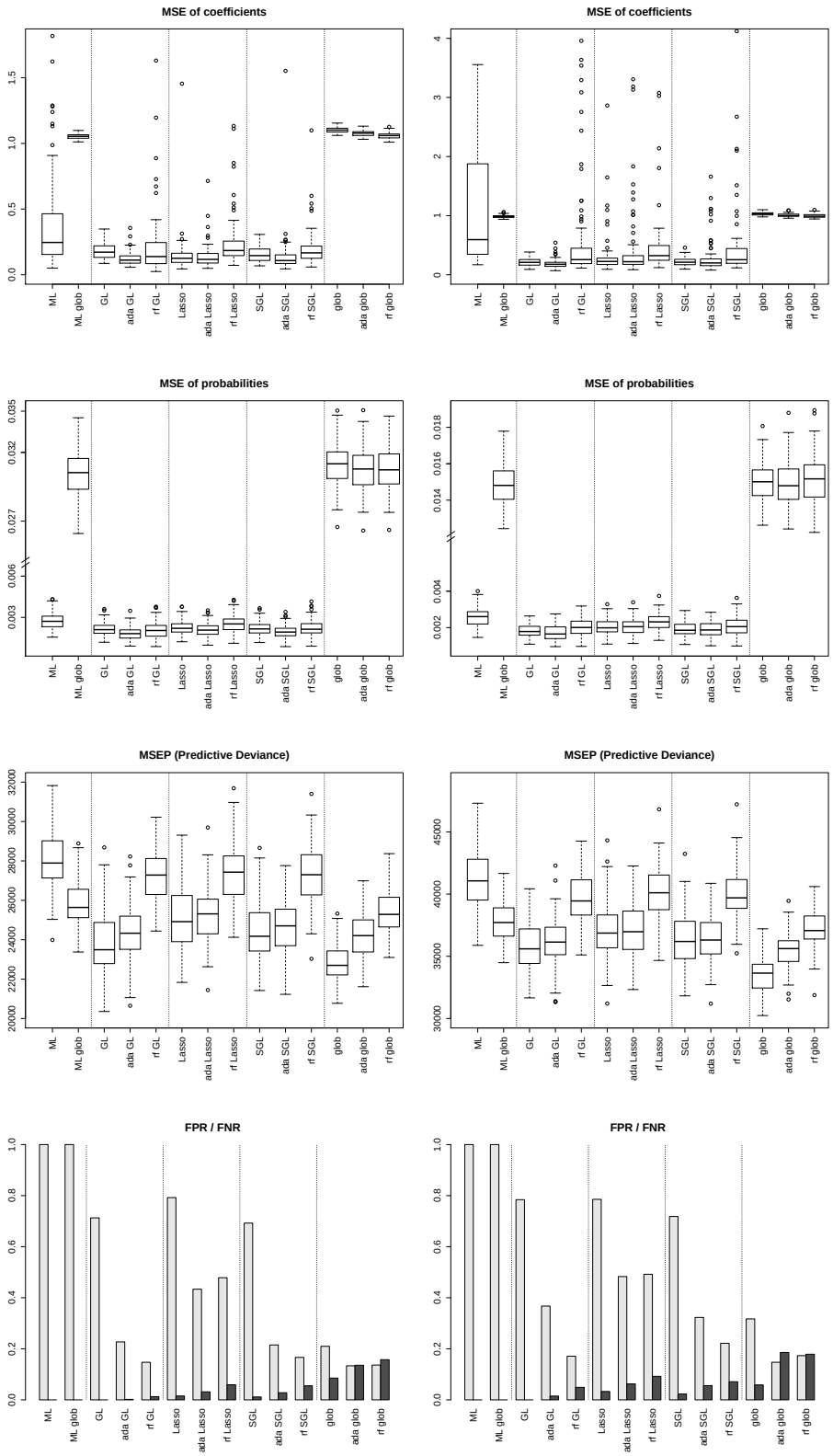

Abbildung 5.3: Ergebnisse Szenario 3: 5 Responsekategorien; 5 kategoriespezifische, 5 globale, 5 irrelevante Prädiktoren; Korrelation: 0,2 (linke Spalte), 0,6 (rechte Spalte); jeweils 1000 Beobachtungen und 100 Iterationen.

5.3 Zusammenfassung

Wie erwartet gelingt es den gruppiert penalisierenden Varianten für kategoriespezifische Effekte, eine Stabilisierung der Schätzer im Vergleich zu denen der ML-Schätzung herbeizuführen. Dies gelingt dem klassischen Lasso nicht, ebenso nicht den Varianten, die ihre finalen Koeffizientenwerte durch einen Refit erhalten. Unter der Annahme, dass sich die wahre Prädiktorstruktur aus Prädiktoren mit kategoriespezifischen und globalen Effekten zusammensetzt, werden für GL- und SGL-Penalisierung mit und ohne adaptive Gewichtung, hinsichtlich der Koeffizienten- und Wahrscheinlichkeiten-MSE, die besten Ergebnisse erlangt. Dies gilt ebenso für die prädiktive Devianz in Fällen, in denen die Anzahl zu schätzender Parameter im Vergleich zur Anzahl vorhandener Beobachtungen relativ groß ist. In sehr datenreichen Schätzsituationen besitzen globale Schätzer eine tendenziell bessere Prognosegüte.

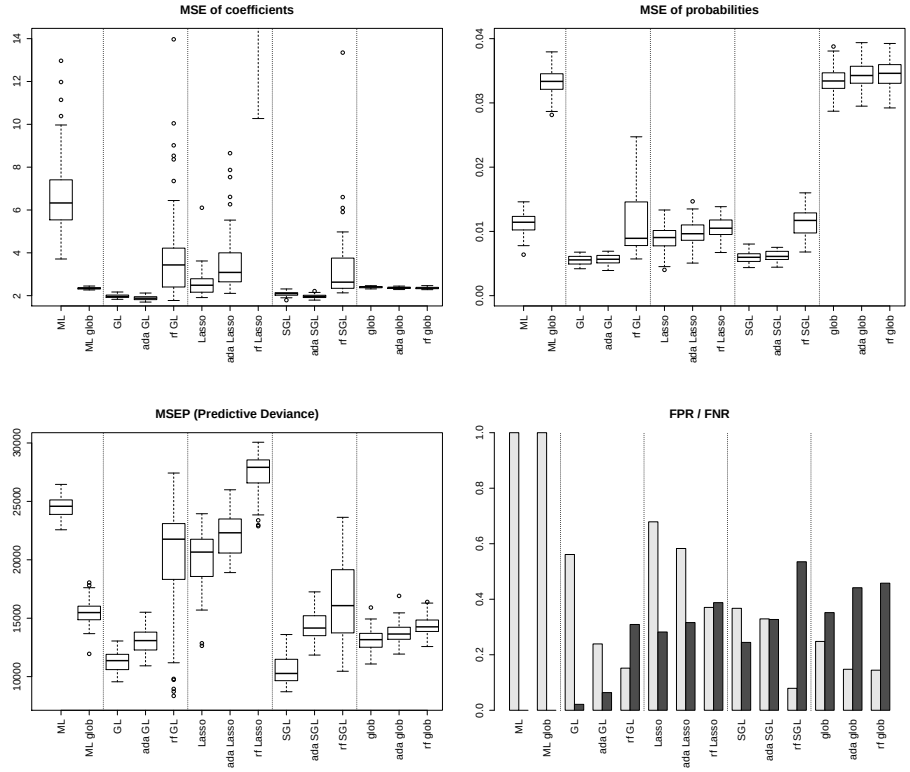

Abbildung 5.4: Ergebnisse Szenario 4: 10 Responsekategorien; 30 kategoriespezifische, 10 globale, 20 irrelevante Prädiktoren; Korrelation: 0,6; 500 Beobachtungen; 60 Iterationen.

Vergleicht man die Penalisierungsansätze hinsichtlich ihrer Selektionsraten ist festzuhalten, dass durch Modifikationen (adaptive Gewichte, Refit) die FPR sinken, da die fälschlicherweise als aktiv geschätzte Variablen durch zweistufige Schätzung bzw. Koeffizientengewichtgung reduziert werden. Gleichzeitig steigt

durch die schärfere Selektion die Gefahr, relevante Variablen fälschlicherweise als irrelevant einzustufen, was sich in einem Anstieg der FNR wiederspiegelt. Je besser die Datenlage, desto geringer ist der Unterschied der FPR einer adaptiven Gewichtung im Vergleich zu der mit Refit. Im Fall $p > n$ steigen, aufgrund der schlechten Datensituation, die FNR extrem stark an, parallel dazu fällt die FPR. Da ein fälschliches missachten relevanter Variablen eher nicht wünschenswert ist, gelingt es keinem der Penalisierungsansätze in einer datenarmen Schätzsituation zufriedenstellende Selektionsergebnisse zu generieren.

Insgesamt erzielen Group und Sparse Group Lasso ähnliche Resultate in den Vergleichskriterien. Beide Verfahren werden mit adaptiven Gewichten, denen ein besserer Kompromis zwischen Schätz- und Selekionsgüte gelingt als Refit-Varianten, in den Datenbeispielen in Kapitel 6 angewendet.

Kapitel 6

Anwendungsbeispiele

Anhand von zwei Datensätzen werden innerhalb dieses Kapitels Anwendungsbeispiele für verschiedenen Penalisierungsansätze im sequentiellen Logit-Modell gegeben. Der in Abschnitt 6.1 analysierte Datensatz „Gründerstudie“ betrachtet den Einfluss betriebswirtschaftlicher Unternehmensmerkmale auf die Zeitdauer bis zum eventuellen Konkurs. Die Zahl der beobachteten Unternehmen liegt hierbei mit 1224 deutlich über der Anzahl zu schätzender Koeffizienten von 150. Dieses datenreiche Beispiel dient gleichzeitig der Verknüpfung von Survivaldaten mit sequentiellen Modellen. Der in Abschnitt 6.2 untersuchte Datensatz „Gleason-Score“ betrachtet den Einfluss von Genexpositionen auf das Level des Gleason-Score. Mit Daten von 52 Beobachtungseinheiten und 490 zu schätzenden Kategoriekoeffizienten ist dies eine Koeffizientensituation, die durch eine gewöhnliche ML-Schätzung nicht gelöst werden kann, weshalb penalisierte Ansätze verwendet werden müssen.

6.1 Datensatz: Gründerstudie

6.1.1 Beschreibung

Dieser Datensatz basiert auf der Studie „Berufserfahrungen und Erfolgschancen von Unternehmensgründern“ (Münchner Gründerstudie), die im Jahr 1990 von der Universität München in Zusammenarbeit mit der Industrie- und Handelskammer für München und Oberbayern durchgeführt wurde. Brüderl, Preisendörfer & Ziegler (1992) untersuchen in diesem Rahmen die Überlebenschancen von, in den Jahren 1985-1986 in Oberbayern, neu gegründeten Unternehmen, vor dem Hintergrund diverser betriebswirtschaftlicher Determinanten. Hierzu wurden von 1849 Unternehmen Daten bezüglich Unternehmenscharakteristika, z.B. Anzahl Angestellter, Kapitalausstattung, Rechtsform, Branche und Charakterstika des Günders, z.B. Arbeitserfahrung, Schulbildung erhoben.

In dieser Ausarbeitung wird auf den in Tutz (2000) verwendeten Datensatz zurückgegriffen, der die Überlebensdauern - Anzahl der Monate bis zum Konkurs - von 1224 beobachtete Unternehmen, 14 kategoriale Merkmale und mit dem Alter des Unternehmensgründers, ein metrisches Merkmal umfasst. Die erhobenen Merkmale sind in Tabelle 6.1 mit Beschreibung und entsprechender Kodierung der kategorialen Merkmale aufgelistet.

Variable	Beschreibung	Kodierung
wirt	Wirtschafts- bereich	1 Industrie, verarbeitendes und Baugewerbe 2 Handel 3 Dienstleistungen
recht	Rechtsform	1 Kleingewerbe ohne Handelsregistereintrag 2 Einzelfirma oder Vollkaufmann 3 GmbH, GmbH & CoKG 4 GbR, KG, OHG
stdort	Standort	1 Wohngebiet 2 Geschäftsgegend, Industrieviertel, Mischviertel
neu	Neugründung oder Firmenübernahme	1 vollständige Neugründung, 2 teilweise Übernahme, Firmenübernahme, sonst
ezweck	Erwerbszweck	1 Vollerwebszweck 2 Nebenerwerbszweck
stkap	Gesamtes Startkapital	1 kein Startkapital 2 0 DM $<$ Startkapital $\leq$ 25000 DM 3 25000 DM $<$ Startkapital $\leq$ 75000 DM 4 75000 DM $<$ Startkapital
ek	Eigenkapital	1 kein Eigenkapital 2 Eigenkapital $<$ 20000 DM 3 20000 DM $\leq$ Eigenkapital $<$ 50000 DM 4 50000 DM $<$ Eigenkapital
fk	Fremdkapital in DM	1 kein Fremdkapital 2 Fremdkapital $>$ 0
zielm	Zielmarkt	1 lokaler Markt 2 überregionaler Markt
kart	Kreis der Kunden	1 breit gestreut 2 kleine Zahl großer Kunden, ein großer Kunde
schab	Schulabschluss	1 Volks-, Hauptschule 2 mittlere Reife 3 Fachhochschulreife, Abitur
sex	Geschlecht	1 Frau 2 Mann
berf	Berufserfahrung vor Gründung	1 unter zehn Jahre 2 zwischen zehn und zwanzig Jahre 3 länger als zwanzig Jahre
be	Anzahl der Beschäftigten im Gründungsjahr	1 kein oder ein Beschäftigter 2 zwei oder drei Beschäftigte 3 mehr als drei Beschäftigte
age	Alter des Unternehmensgründers im Zeitpunkt der Gründung (metrisch)	

Tabelle 6.1: Abkürzungen, Beschreibungen und Kategorien der Merkmale des Gründerdatensatzes

Die Verweildauern der 1224 Unternehmen, in Monaten, sind in Abbildung 6.1 dargestellt. Die minimale Beobachtungsdauer beträgt einen Monat, die maximale 66 Monate. 865 Unternehmen wurden (rechts-)zensiert, d.h. der Gründungszeitpunkt dieser Unternehmen ist zwar bekannt, allerdings sind diese Unternehmen im Verlauf der Studie - nicht durch Konkurs - aus dieser herausgefallen und konnten nicht weiter beobachtet werden. Keines der Unternehmen in diesem Datensatz wurde innerhalb der ersten 36 Monate zensiert. Für 83 % der 359 Unternehmen, deren Konkurszeitpunkt bekannt ist, fand der Konkurs innerhalb der ersten 36 Monate statt. Aufgrund dieser Aspekte kann bei geeigneter Kategorienbildung der Zensierungsindikator außer Acht gelassen werden. Für die Verwendung der Überlebensdauer als kategoriale Zielgröße im sequentiellen Logit-Modell wird eine Einteilung in 7 Kategorien vorgenommen. In Kategorie 1 fallen Unternehmen mit einer Überlebensdauer von höchstens 6 Monaten, d.h. Unternehmen, die in den ersten sechs Monaten Konkurs anmelden mussten. Kategorie 2 umfasst die Monate 7 bis einschließlich 12, Kategorie 3 die Monate 13 bis einschließlich 18, Kategorie 4 die Monate 19 bis einschließlich 24, Kategorie 5 die Monate 25 bis einschließlich 30, Kategorie 6 die Monate 31 bis einschließlich 36 und Kategorie 7 die Monate größer als 36, d.h. alle Unternehmen, die mindestens 36 Monate bestanden haben. Die Einteilung wird durch die vertikalen gestrichelten Linien in Abbildung 6.1 verdeutlicht. Diese Diskretisierung der Variable Überlebenszeit zeigt die in Abschnitt 3.4 angesprochene Verknüpfung diskreter Survivalmodelle mit dem sequentiellen Modell. Dieser Datensatz lässt sich ebenso mit Hilfe von Survivalmodellen für diskrete und stetige Zeit analysieren.

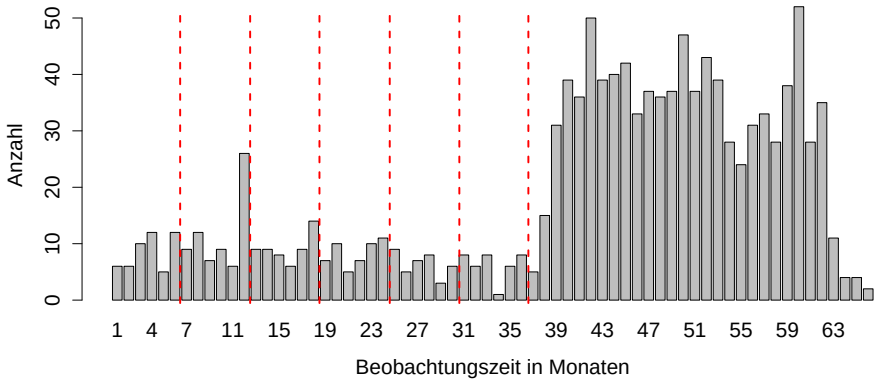

Abbildung 6.1: Beobachtungsdauern der 1224 Unternehmen - Konkurse und Zensierungen (ab dem 37. Monat).

Als Prädiktoren gehen alle in Tabelle 6.1 genannten Variablen mit kategoriespezifischen Effekten in das sequentielle Logit-Modell ein. Die Faktorvariablen werden mit der jeweils erstgenannten Kategorie als Referenzkategorie dummycodiert. Für die Analyse der Modellergebnisse, setzt sich die Bezeichnung des Regressionskoeffizienten der Kategorie eines Faktors aus dem Variablennamen und dem Kategorieindex zusammen, z.B. bezeichnet *ek4* den Koeffizienten für die Kategorie Eigenkapital (*ek*) größer als 50000 DM. Für die Variable *stkap* wurde der Daten-

satz um die Kategorie 1 bereinigt, da in dieser keine Beobachtungen vorhanden waren. Somit dient die zweite Kategorie als Referenzkategorie. Für die Variable *neu* musste die Kategoriebezeichnung des ursprünglichen Datensatzes angepasst werden. Die Zielgröße wird mit der letzten Kategorie als Referenzkategorie in das Modell aufgenommen.

6.1.2 Auswertung

Unter Verwendung des sequentiellen Logit-Modells mit kategoriespezifischen Effekten, sind bei sechs Zielgrößenkategorien 150 Koeffizienten inklusive Intercepts zu schätzen. Basierend auf den Ergebnissen der Simulationsstudie, speziell der datenreichen Situation in Szenario 2 werden als Penalisierungsansätze das Group Lasso und das Sparse Group Lasso, jeweils mit adaptiver Gewichtung ausgewählt. Diese Entscheidung beruht auf deren verhältnismäßig großer Stabilität und Güte der geschätzten Koeffizienten und Wahrscheinlichkeiten, sowie der geringeren Fehlselektionsraten. Eine Refit-Modifikation schneidet zwar bezüglich der Selektionsgüte geringfügig besser ab, tendiert allerdings zu instabileren Schätzern.

Zur Bestimmung der Modellkoeffizienten für verschiedene, durch den Tuningparameter $\lambda \geq 0$ bestimmte, Penalisierungsintensitäten wurden aus dem Wertebereich $[\log(0.05), \log(158, 32)]$ 100 gleichabständige λ-Werte berechnet. Für jeden Wert des Penalisierungsparameters wurde ein Modell geschätzt. Der logarithmierte Wertebereich ermöglicht es, eine Vielzahl von Modellen für kleinere Penalisierungsparameter zu bestimmen, um eine etwaige Variablenselektion in kleineren Schritten nachvollziehen zu können. Der Wert 0.05 entspricht dabei einer vernachlässigbar geringen Penalisierung, sodass dieses Modell annähernd die unpenalisierten ML-Koeffizienten angibt. Aus diesen 100 Modellen verschiedener Penalisierungsstärke lässt sich, mittels Akaikes Informationskriterium (AIC), Bayesschen Informationskriterium (BIC) oder zehnfacher Kreuzvalidierung (CV) ein optimales Modell bestimmen. Da die Wahl eines kreuzvalidierten Modells von der zufälligen Wahl der Teildatensätze abhängt, das Bayessche Informationskriterium eine sehr restriktive Modellwahl durchführt, wird im Folgenden beispielhaft Akaikes Informationskriterum zur Wahl eines optimalen Modells verwendet.

Parameterauswertung

Tabelle 6.2 zeigt die, aus den penalisierten ML-Schätzungen, resultierenden Koeffizienten. Die Spalten kennzeichnen, auf welche der Zielgrößenkategorien sich ein Koeffizient der in den Zeilen angegebenen Variable bezieht. Die grau hinterlegten Zeilen markieren die, mittels adaptivem Group Lasso bestimmten Koeffizienten. Die jeweils unmittelbar darunter stehenden, nicht hinterlegten Werte geben die Koeffizienten der adaptiven Sparse Group Lasso-Penalisierung an. Diese Modellkoeffizienten entstammen den beiden AIC-optimalen Modellen für adaptives Group und adaptives Sparse Group Lasso.

Die Variablen Standort (*stdort2*), Neugründung (*neu2*), Zielmarkt (*zielm2*), Geschlecht des Gründers (*sex2*) und Berufserfahrung (*berf2,berf3*) wurden von beiden Penalisierungsvarianten aus dem AIC-optimalen Modell entfernt, besitzen somit keinen Einfluss auf die Dauer bis zu einer Insolvenz. Sofern für beide Ansätze Koeffizienten mit einem von null verschiedenen Wert vorliegen, besitzen diese in allen Fällen das gleiche Vorzeichen und einen ähnlichen Wert. Dies liegt daran,

dass Sparse Group Lasso Werte des Group Lasso nahe null direkt auf null schätzen würde. In Rückblick auf die Interpretation der Koeffizienten des sequentiellen Logit-Modells auf Seite 22 f. sei daran erinnert, dass ein negativer Koeffizient ausdrückt, dass die Chance (Risiko) in Zielgrößenkategorie r zu verbleiben, statt in eine höhere Kategorie aufzusteigen, sich um den Faktor exp(Koeffizient) verringert , wenn die entsprechende Einflussgröße eine andere Prädiktorkategorie, anstatt ihrer Referenzkategorie annimmt, unter sonst gleichen Einflussgrößen. Da der Verbleib in einer Zielgrößenkategorie, anstatt eines Aufstiegs in eine höhere Kategorie, eine frühere Insolvenz impliziert, wird im Folgenden vom Risiko des Verbleibs in einer Kategorie gesprochen. Dies bedeutet, dass ein negativer Koeffizient einen positiven Einfluss dieser Prädiktorkategorie auf die Überlebenszeit, also eine spätere Unternehmensinsolvenz, ausdrückt. Für einen positiven Koeffizienten gilt die umgekehrte Interpretation, sodass dieser eine frühere Insolvenz impliziert. Dies sei beispielhaft am Koeffizienten 1,412 der Prädiktorkategorie *recht4* (Rechtsform GbR, KG, OHG) für die fünfte Responsekategorie dargestellt: Hat eine Unternehmensgründung die Rechtsform GbR, KG oder OHG, dann erhöht sich das Risiko, eine Insolvenz in den Monaten 25 bis 30 nach Unternehmensgründung zu erfahren, anstatt nach dem 30. Monat, um das $\exp(1.412) = 4.104$-fache, gegenüber eines Kleingewerbes ohne Handelsregistereintrag, c.p.

Der Prädiktor Wirtschaftsbereich weist für Handel (*wirt2*) und Dienstleistung (*wirt3*) für alle Zielgrößenkategorien positive Werte auf, sodass in jeder Zielgrößenkategorie ein höheres Risiko auf Insolvenz besteht, als für den Wirtschaftsbereich Industrie/verarbeitendes Gewerbe. Der Effekt innerhalb der ersten sechs Monate ist am geringsten und wird duch das Sparse Group Lasso auf null geschätzt. Die Rechtsformen GmbH und GmbH & CoKG (*recht3*) haben in allen Zeitintervallen ein geringeres Insolvenzrisiko als ein Kleingewerbe ohne Handelsregistereintrag. Für *recht2* und *recht4* variieren die Effekte über die Zeit hinweg. SGL schätzt ebenfalls alle Koeffizienten als von null verschieden. Ist der Erwerbszweck Nebenerwerb, erhöht dies über die Zeitkategorien hinweg tendenziell das Risiko einer Insolvenz, gegenüber dem Vollerwerbszweck.

Für die Kapitalausstattung des Unternehmens ergibt sich ein sehr vielseitiges Bild. Eine starke Gesamtkapitalausstattung zur Unternehmensgründung verringert das Risiko einer frühzeitigen Insolvenz erheblich und besitzt tendenziell einen positiven Einfluss auf die Überlebenschancen der Neugründung über die Zeit hinweg. Eine Eigenkapitalausstattung von unter 20 000 DM (*ek2*) erhöht das Risiko einer Insolvenz innerhalb der ersten Monate. Je höher die Eigenkapitalausstattung (*ek3, ek4*), desto geringer ist das frühzeitige Insolvenzrisiko. Vorhandenes Fremdkapital (*fk2*) erhöht ebenfalls das Risiko einer frühzeitigen Insolvenz. Dies deckt sich mit der intuitven Erwartung, dass ein Unternehmen mit höherer Eigenkapitalausstattung von dieser die ersten Monate zehren kann. Unabhänig von der Finanzierungsart lässt der Effekt über die Zeit hinweg nach. Dies führt dazu, dass die Sparse Group Lasso-Penalisierung eine Vielzahl der Kategorieeffekte auf null schätzt. Für Zielgrößenkategorie 6 werden fast ausschließlich negative Koeffizienten geschätzt.

Eine kleine Zahl großer Kunden/ein großer Kunde (*kart2*) erhöht die Überlebenschancen ebenso, wie ein höherer Schulabschluss (*schab2, schab3*). Allerdings werden viele dieser Effekte, vor allem in den ersten Monaten nach Unternehmensgründung durch SGL auf null geschätzt. Große Kunden, sowie ein höherer

Kategorie	1	2	3	4	5	6
Intercept	-1.593	-1.653	-2.844	-2.321	-2.948	-2.624
	-1.753	-1.404	-3.275	-2.957	-3.367	-2.146
wirt2	0.014	0.524	0.84	0.689	0.767	0.29
	0	0.387	0.845	0.593	0.683	0.186
wirt3	0.076	0.311	0.838	0.448	0.824	0.533
	0	0.204	0.822	0.359	0.777	0.356
recht2	-0.309	-0.571	0.285	-0.56	0.744	-1.471
	-0.31	-0.484	0.214	-0.553	0.668	-1.325
recht3	-2.005	-0.887	-1.414	-1.667	-0.489	-1.322
	-2.214	-0.775	-1.378	-1.777	-0.311	-1.135
recht4	-0.377	-0.396	0.155	-0.247	1.412	-0.413
	-0.432	-0.353	0.093	-0.313	1.288	-0.393
stdort2	0	0	0	0	0	0
neu2	0	0	0	0	0	0
ezweck2	-0.616	0.201	0.014	0.582	-0.252	0.968
	-0.517	0	0	0.537	0	0.965
stkap2	-1.406	-0.688	0.434	-0.555	-0.724	0.341
	-1.508	-0.567	0.137	0	-0.672	0
stkap3	-2.95	-0.901	-0.377	-0.1	-1.613	-0.246
	-3.278	-0.776	-0.284	0	-1.261	0
ek2	0.684	0.162	0.151	-0.372	0.012	-0.345
	0.644	0	0	0	0	-0.363
ek3	0.29	0.271	-0.039	-0.379	-0.009	-0.074
	0.293	0	0	0	0	-0.017
ek4	0.244	0.135	0.077	-0.325	0.325	-0.092
	0.288	0	0	0	0	-0.186
fk2	0.673	0.339	-0.021	0.165	0.063	-0.51
	0.699	0.15	0	0	0	-0.541
zielm2	0	0	0	0	0	0
kart2	-0.221	-0.127	-0.352	-0.168	-0.083	-0.675
	0	0	-0.183	0	0	-0.748
schab2	-0.075	-0.113	-0.162	0.168	-0.5	0.103
	0	0	0	0	-0.499	0
schab3	-0.225	-0.359	0.103	0.038	-0.499	-0.221
	0	0	0	0	-0.525	0
sex2	0	0	0	0	0	0
berf2	0	0	0	0	0	0
berf3	0	0	0	0	0	0
be2	-0.311	0.278	-0.315	0.033	-0.6	-0.708
	0	0	-0.167	0	-0.391	-0.661
be3	-0.742	0.086	-0.574	-0.671	0.492	-2.279
	0	0	-0.254	0	0.271	-2.513
age	-0.02	-0.021	-0.012	-0.007	-0.01	0.009
	-0.02	-0.021	0	0	0	0

Tabelle 6.2: Gründerdatensatz: Modellkoeffizienten je Zielgrößenkategorie (Spalten) für adaptives Group Lasso (Zeilenhintergrund grau) und adaptives Sparse Group Lasso.

Schulabschluss spielen somit für das kurzfristige Insolvenzrisiko keine Rolle. Dies gilt ebenso für eine höhere Anzahl an Beschäftigten (*be2, be3*) bei der Unternehmensgründung. Die einzig metrisch aufgenommene Variable Alter (*age*) besitzt kaum einen Einfluss auf die Überlebenschancen. In den ersten Monaten scheint ein höheres Alter des Gründers das frühzeitige Risiko der Insolvenz zu verringern.

Koeffizientenpfade

Bisher wurden die Koeffizienten zweier konkreter Modellschätzungen, deren Parameter penalisiert mit adaptivem Group Lasso bzw. adaptivem Sparse Group Lasso geschätzt wurden, interpretiert. Beide Modelle wurden unter Verwendung des AIC als Modellwahlkriterium aus den 100 Modellschätzungen, die mit unterschiedlichem Einfluss des Penalisierungsterms gerneriert wurden, ausgewählt. Eine Darstellung der Koeffizienten aller 100 Modelle eines spezifischen Penalisierungsansatzes gelingt mit Hilfe von Koeffizientenpfaden. Ein Koeffizientenpfad ist die grafische Darstellung eines kategoriespezifischen Regressionskoeffizienten in Abhängigkeit vom Penalisierungsparameter λ. Dieser Pfad zeigt an, wie mit zunehmender Stärke der Penalisierung der jeweilige Koeffizient gegen null geschrumpt wird. Die Darstellung dieser Pfade zeigt Abbildung 6.2. Innerhalb einer Grafik werden, für die in der Grafiküberschrift angegebene Kovariable (dummycodierte Prädiktorkategorie), die Pfade der Koeffizienten für alle sechs Zielgrößenkategorien angegeben. Die **Färbung der verschiedenen Pfade** je Zielgrößenkategorie ergibt sich wie folgt: schwarz (Kategorie 1), rot (Kategorie 2), grün (Kategorie 3), blau (Kategorie 4), türkis (Kategorie 5), pink (Kategorie 6). Der Wert des kategoriespezifischen Koeffizienten wird auf der Ordinate abgetragen, der Wert des Penalisierungsparameters, der für die Erzeugung dieses Koeffizienten ursächlich war, auf der Abszisse. Hierfür wird nicht direkt der Wert von λ verwendet, sondern $\log(1+\lambda)$. Durch diese Transformation erhält man für den Abszissenwert 0, die unpenalisierten ML-Koeffizienten am rechten Rand der Grafik. Für zunehmenden Grad der Penalisierung werden die Koeffizienten gegen null geschrumpft, wodurch die Pfade, für höhere Werte der Abszisse, gegen null laufen. Zusätzlich sind in jeder Grafik drei vertikale Linien für die Modellwahlkriterien (AIC, BIC, CV) eingezeichnet, deren Schnitt mit den Koeffizientenpfaden diejenigen Werte selektiert, deren Modell mit diesem Kriterium ausgewählt wurde. In der linken Spalte in Abbildung 6.2 sind die Koeffizientenpfade der kategoriespezifischen Intercepts, der Variable Startkapital (*stkap2*, *stkap3*) der adaptiven Group Lasso-Penalisierung eingezeichnet, in der rechten Spalte die Koeffizienten der adaptiven Sparse Group Lasso-Penalisierung. In Abbildung 6.3 sind die Koeffizientenpfade des vierkategorialen Merkmals Rechtsform (*recht2, recht3, recht4*) eingezeichnet.

Die kategoriespezifischen Intercepts besitzen für alle Kategorien negative Werte und werden für das SGL bereits mit geringerer Penalisierung auf null geschätzt. Die Intercepts werden ebenso wie die einzigen beiden Variablen Startkapital und Rechtsform für alle Auswahlkriterien mit fast allen Koeffizienten im Modell belassen. Es wird nochmals darauf hingewiesen, dass die Referenzkategorie der Variable Startkapital, ein Kapital zwischen 0 und 25000 DM angibt. Ein Startkapital zwischen 25 und 75 Tausend DM erhöht die Überlebenschancen in den ersten zwölf Monaten. Durch das Sparse Group Lasso werden die Koeffizienten der vierten, sechsten und dritten Kategorie auf null geschätzt. Ähnliche Effekte werden für ein Startkapital von mindestens 75 Tausend DM geschätzt.

Für die Variable Rechtsform werden als einzige alle Kategoriekoeffizienten von allen drei Kriterien für das Modell selektiert. Während für eine Einzelfirma oder einen Vollkaufmann (*recht2*) der Effekt nicht eindeutig über die Zeit ist, verringert sich für eine Neugründung in Form einer GmbH oder GmbH & CoKG (*recht3*) das Risiko einer Insolvenz für jeden Zeithorizont gegenüber einem Kleingewerbe. Der Effekt einer GbR, KG, OHG (*recht4*) gegenüber einem Kleingewerbe ist ebenfalls nicht eindeutig über die Zeit. Des Weiteren wurden einzelne Kategorieeffekte der Beschäftigtenzahl (*be3*) und des Erwerbszwecks (*ezweck2*), die durch Sparse Group Lasso einen von null verschiedenen Koeffizienten erhalten haben, mit allen drei Kriterien gewählt. Die Koeffizientenpfade aller übrigen Prädiktoren finden sich in Anhang B.1 auf Seite 76 ff.

Ein Vergleich der beiden durch adaptives Group Lasso und adaptives Sparse Group Lasso penalisierten Modelle, ergibt einen AIC-Wert des Group Lasso-Modells von 2155,09, den des Sparse Group Lasso-Modells mit 2174,45. Das AIC des unpenalisierten Modells kategoriespezifischer Effekte weist einen Wert von 2294,52 auf. Für die BIC-Werte ergeben sich 2307,5 für das GL-Modell, 2345,29 für das SGL-Modell und 3061,00 für das Modell unpenalisierter ML-Schätzung. Diesen beiden Kriterien folgend, würde die Entscheidung zugunsten des adaptive Group Lasso-Modells fallen.

(a) adaptives Group Lasso **(b)** adaptives Sparse Group Lasso

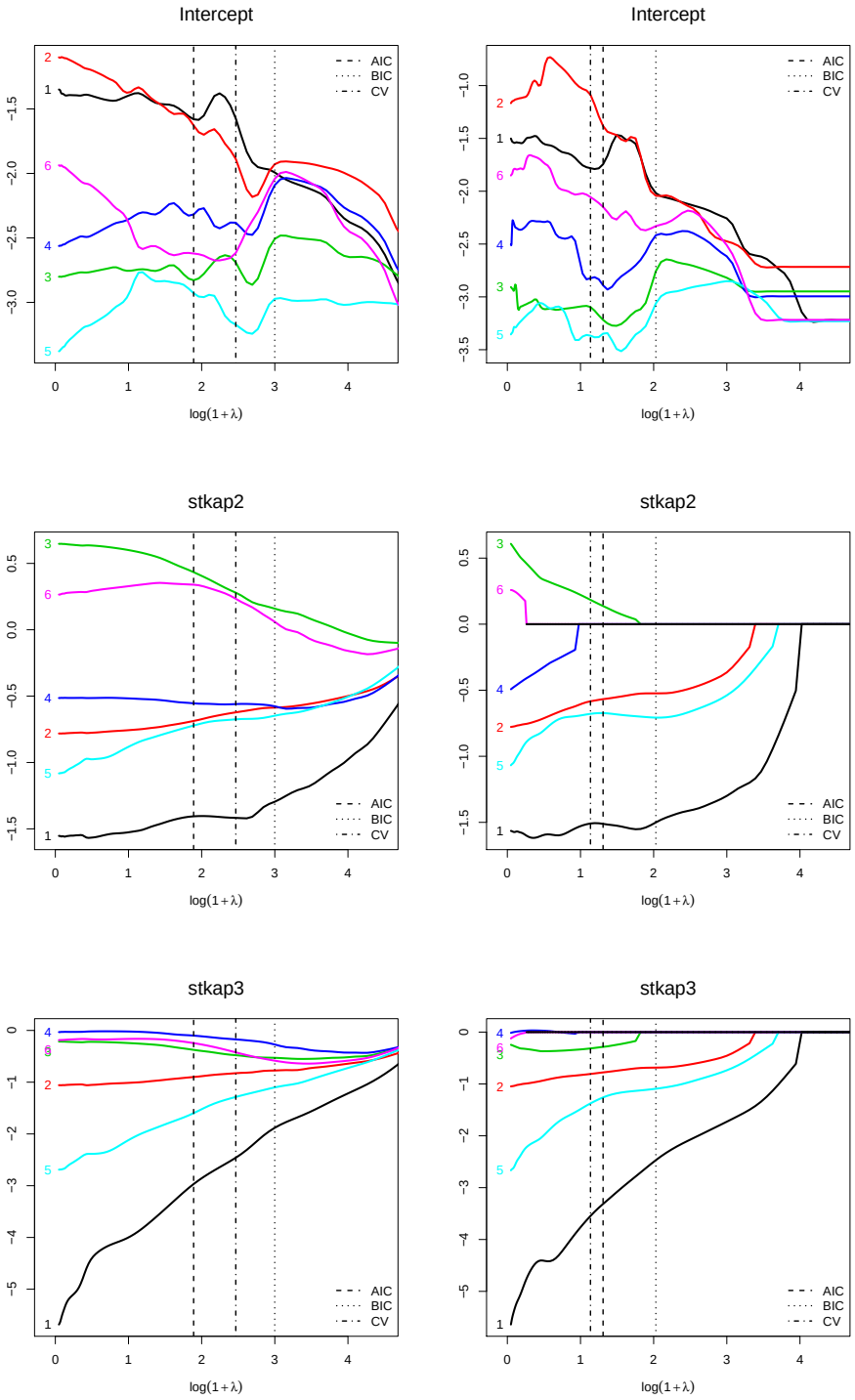

Abbildung 6.2: Gründerdatensatz: Koeffizientenpfade der kategoriespezifischen Intercepts, *stkap2* und *stkap3* für adaptives Group Lasso (linke Spalte) und adaptives Sparse Group Lasso (rechte Spalte).

(a) adaptives Group Lasso **(b)** adaptives Sparse Group Lasso

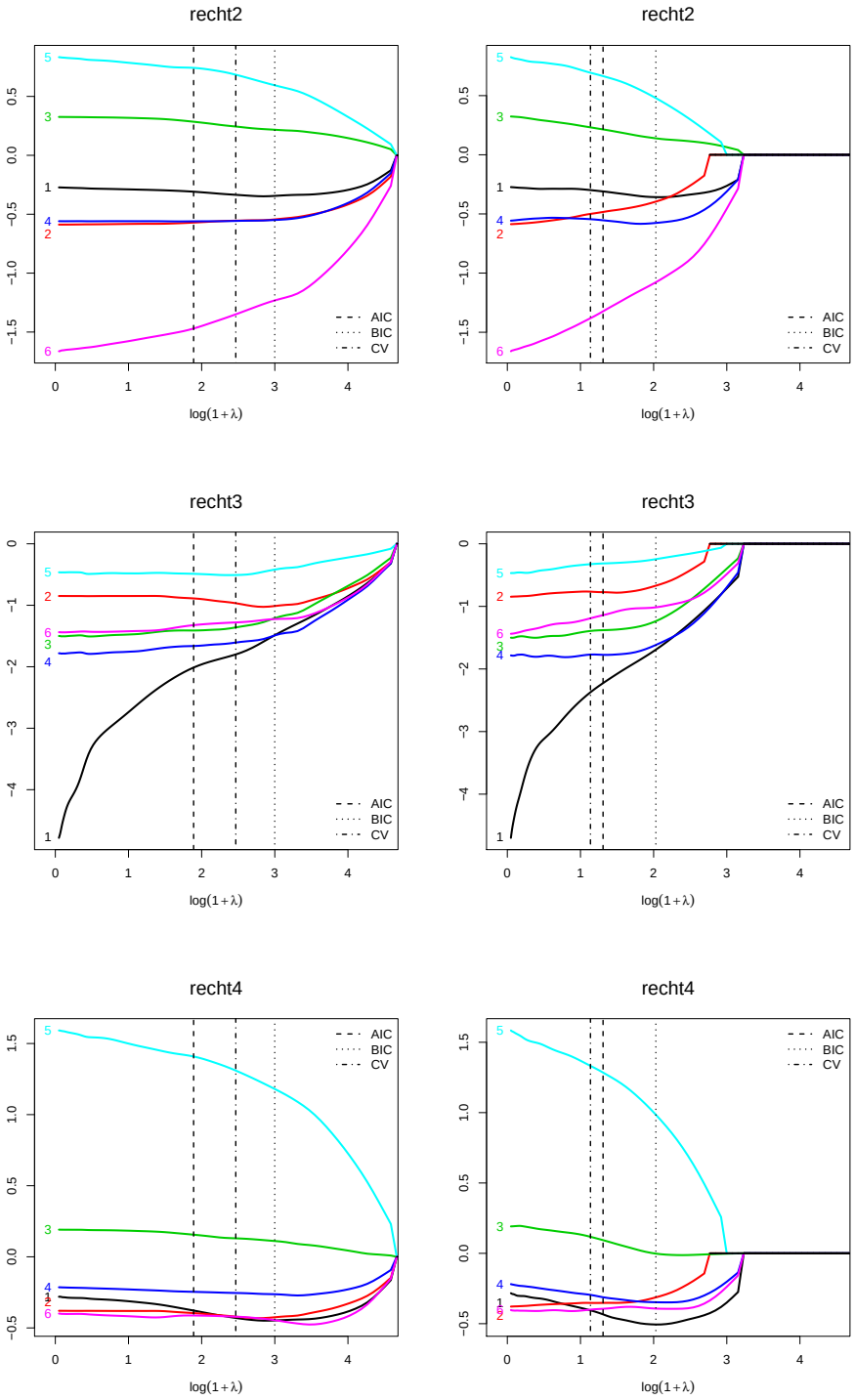

Abbildung 6.3: Gründerdatensatz: Koeffizientenpfade für *recht2*, *recht3* und *recht4* für adaptives Group Lasso (linke Spalte) und adaptives Sparse Group Lasso (rechte Spalte).

6.2 Datensatz: Gleason-Score

6.2.1 Beschreibung

Der Gleason-Score dient als ein Maß zur Bestimmung der Agressivität eines Prostatakarzinoms (Prostatkrebs). Zur Bestimmung des Scores, der Werte zwischen 2 und 10 annehmen kann, wird der männlichen Prostata Drüsengewebe entnommen. Dieses Gewebe kann durch eine Gewebeentnahme (Prostatabiopsie) oder teilweise/vollständige Prostataentfernung (Prostatektomie) gewonnen werden. Zur Beurteilung des Prostatagewebes, anhand einer Einstufung durch den Gleason-Score, wird die Struktur der am häufigsten und am zweithäufigsten vorkommenden Zellen analysiert. Beide Zellarten erhalten einen Wert von 1 bis 5. Je entdifferenzierter die Zellstruktur ist, desto höher der resultierende Wert. Stark entdifferenziertes Gewebe ähnelt dem Wachstumsmuster eines normalen Gewebes sehr wenig. Der Entdifferenzierungsgrad für beide Zellarten wird addiert und ergibt den Gleason-Score. In der Praxis wird ein Gleason-Score von 2-4 als wohl differenziert, von 5-7 als mäßig differenziert und von 8-10 als schlecht differenziert eingestuft. (Vgl. Humphrey (2004))

Als Faktoren für die Entstehung von Prostatakrebs werden neben dem Alter, Ernährungsgewohnheiten, Lebensstil und Umweltfaktoren genannt. (Vgl. Robert Koch-Institut (2010), S. 72 ff.) Geographische Muster in der Verbreitung von Prostatakrebs werden von Baade et al. (2009) festgestellt. Dies bestätigt mitunter die Vermutung, genetische Risikofaktoren hinsichtlich eines Erkrankungsrisikos in Betracht zu ziehen. Hierzu untersuchen Singh et al. (2002) einen Zusammenhang zwischen Genexpressionen und dem Gleason-Score. Unter der Expression eines Gens versteht man den Ausdruck der genetischen Information, die im menschichen Erbgut veranlagt ist (Genotyp), zu einem konkreten Phänotyp, auf dem fundamentalsten Level der Genetik. Singh et al. (2002) verwenden zur Identifikation derjenigen Gene, die die stärkste Korrelation mit dem durch den Gleason-Score ermittelten Grad der Tumordifferenzierung aufweisen, einen Datensatz von 235 Patienten, bei denen in den Jahren 1995 bis 1997 eine vollständige Prostataentfernung druchgeführt wurde. Hochqualitative Daten sind für ca. 12600 Gene von 52 Patienten, in diesem öffentlich zugänglichen Datensatz, verfügbar. Die Identifikation von Genen, die einen Zusammenhang zum Gleason-Score aufweisen, erlaubt es, anhand dieser den klinischen Verlauf der Krankheit zu antizipieren und entsprechende Behandlungen anzuwenden.

Der im folgenden verwendete Datensatz enthält die Expressionen von 244 Genenen dieser 52 Patienten. Deren Gleason-Score verteilt sich entsprechend Abbildung 6.4 auf die Werte 6 bis 10, somit einer mäßigen bis schlechten Differenzierung der Zellen des Prostatagewebes. Für eine Kategorisierung der Gleason-Scores bilden die 26 Patienten mit Gleason-Score 6 die niedrigste Kategorie, die 20 Patienten mit Gleason-Score 7 eine mittlere Kategorie und die 6 Patienten mit Werten von 8, 9, 10 werden zur höchsten Kategorie zusammengefasst - entsprechend den vertikalen Trennlinien der Abbildung.
Chu et al. (2005) argumentieren, dass sich der Gleason-Score als ordinale Variable auffassen lässt, da die Rangwerte zwar geordnet sind, aber deren Abstände sich nicht metrisch interpretieren lassen. Für die Verwendung des sequentiellen Logit-Modells erscheint die Annahme erfüllt, dass die Grade des Gleason-Score

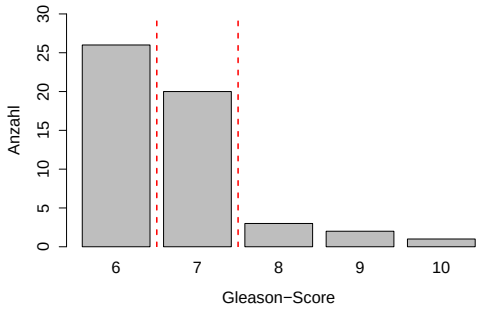

Abbildung 6.4: Werte des Gleason-Score der 52 Patienten

nur sukzessive erreicht werden können. Dies lässt sich durch das zunehmend entdifferenzierende Wachstum der Krebszellen, deren ursprüngliches Gewebe einen vollkommenen differenzierten Zustand der Zellen aufwies, begründen. Mittels oligonucleotider Microarraymessungen werden die Expressionslevel der 244 Gene bestimmt. Diese lassen sich, den Grad an Unter- bzw. Überexpression ausdrückend, als metrische Kovariablen in das Modell aufnehmen. (Vgl. Balakrishnan & Rao (2004), S. 675 f.)

6.2.2 Auswertung

Für das sequentielle Logit-Modell mit kategoriespezifischen Effekten sind in dieser Datenkonstellation, mit drei Zielgrößenkategorien und 244 metrischen Einflussgrößen, 490 Parameter zu schätzen - eingeschlossen der beiden Interceptparameter. Hierfür stehen die Informationen von 52 Beobachtungseinheiten zur Verfügung, sodass eine extrem datenarme Schätzsituation vorliegt. Für die penalisierte Schätzung wurden, ebenfalls wie für den Gründerdatensatz, Group Lasso und Sparse Group Lasso mit adaptiver Gewichtung angewendet.

Parameterauswertung

Identisch zum vorangegangenen Datenbeispiel, wurde für den Penalisierungsparameter λ ein Gitter von 100 Werten bestimmt, das zu 100 geschätzten Modellen führt. Die Koeffizientenschätzungen sind in Tabelle 6.3 dargestellt. Dabei wurden die Koeffizienten, der jeweils zwei Responsekategorien, jeweils für die drei Modellwahlkriterien AIC, BIC und Kreuzvalidierung (zehnfach) in die Spalten der Tabelle eingetragen. Die grau hinterlegten Zeilen geben die Koeffizientenschätzungen, der mit V1 bis V244 codierten Gene, für die Modelle mit adaptiver Group Lasso-Penalisierung an, die unmittelbar darunter liegenden Zeilen, jeweils die Koeffizientenschätzungen für die Modelle der adaptiven Sparse Group Lasso-Penalisierung. Insgesamt wurden 25 Gene durch das unpenalisierte Modell, mit einer Korrektur, um in dieser überparametrisierten Schätzsituation Schätzer zu erhalten, mit von null verschiedenen Werten bestimmt, sodass 219 Gene unmittelbar auf null geschätzt wurden. Von diesen 25 Genen wurden für adaptive GL-Penalisierung 20, im durch Kreuzvalidierung bestimmten Modell, zwölf in dem durch AIC bestimmten und drei in dem durch BIC bestimmten Modell behalten.

Speziell in dieser Datensituation ist auf die Abhängigkeit des Kreuzvalidierungskriteriums von den gewählten Teildatensätzen hinzuweisen, sodass dieses Kriterium zu hoher Variabilität in der Wahl seines optimalen Modells führen kann. In den unter adaptivem SGL geschätzten Modellen, wurde durch Kreuzvalidierung und das AIC ein Modell mit je 16 Einflussgrößen, durch das BIC ein optimales Modell mit einer Einflussgröße gewählt. Das unter adaptiver GL-Penalisierung geschätzte Modell hat einen AIC-Wert von 58.912, das unter adaptiver SGL-Penalisierung geschätzte, einen Wert von 55.428. Die Werte des BIC betragen jeweils 89.261 und 83.421. Hinsichtlich beider Kriterien würde das entsprechende SGL-Modell gewählt werden. Für die Interpretation werden die durch Kreuzvalidierung gewählten Koeffizienten nicht weiter betrachtet, da die Koeffizienten des GL-Modells weniger restringierte Koeffizenten aufweisen, als das AIC-optimale Modell und die Koeffizienten des SGL-Modells ähnlich dem AIC-opimalen Modell sind.

Ausgehend von der Interpretation des sequentiellen logistischen Modells, führt ein positiver Regressionskoeffizient für ein Individuum, mit einem um einen Messpunkt höheren Expressionslevel, zu einer um exp(Koeffizientenwert)-fachen Chance, in gegebener Kategorie zu verbleiben, als in eine höhere Gleason-Score-Kategorie zu fallen. Ein negativer Regressionskoeffizient erhöht somit das Risiko, in eine höhere Gleason-Score-Kategorie überzugehen.

Zunächst werden die Koeffizienten der AIC-optimalen Modelle betrachtet: die Variablen V9, V30, V109, V136 und V212 des GL-penalisierten Modells erhalten für beide Responsekategorien positive Effekte, erhöhen somit die Chance in der gegbenen Gleason-Score-Kategorie zu verbleiben. Demhingegen erhalten die Variablen V43, V62 und V145 für beide Responsekategorien negative Effekte, erhöhen somit das Risiko eines Übergangs hin zu einem höheren Gleason-Score. Die Variablen V1, V57 und V85 besitzen wechselnde Effekte mit einem negativen Koeffizienten in der zweiten Responsekategorie. Unter adaptiver SGL-Penalisierung erhielt für jede Variable nur einer der beiden Koeffizienten einen von null verschiedenen Wert. Abgesehen von den Variablen V43, V81 und V204, resultierte für alle nicht auf null geschätzten Koeffizienten der ersten Responsekategorie ein positiver Wert (9 Variablen). Auffällig sind die Variablen V62, V119 und V145, für die unter SGL-Penalisierung ausschließlich in der zweiten Responsekategorie ein Wert geschätzt wurde mit einem negativen Vorzeichen. Für diese drei Variablen erhöht sich somit das Risiko zu einem höheren Gleason-Score.

Das die Komplexität eines Modells stärker bestrafende BIC, resultiert für diese Modellschätzung in extrem parametersparsammen Modellen, im Vergleich zu den AIC-optimalen Modellen. Unter adaptiver GL-Penalisierung ergeben sich für das BIC-optimale Modell die Variablen V9, V30 und V109 mit fast ausschließlich positiven Koeffizienten. Unter adaptiver SGL-Penalisierung resultiert ein einziger von null verschiedener Koeffizient der Variable V109 im BIC-optimalen Modell.

Generell ist für die Schätzergebnisse zu berücksichtigen, dass in der Zielgrößenkategorie 3 der Gleason-Scores 8,9 und 10 lediglich sechs Beobachtungen vorhanden sind.

Kriterium	**AIC**		**CV**		**BIC**	
Kategorie	**1**	**2**	**1**	**2**	**1**	**2**
Intercept	-2.8627	5.7466	-9.3888	11.9428	-0.9619	1.3134
	-5.8217	9.6475	-5.2623	8.5035	-0.4883	1.2799
V1	0.0205	-0.001	0.0819	0	0	0
	0.05	0	0.043	0	0	0
V9	0.5277	0.0808	1.2914	0.1564	0.1325	0.0026
	0.8514	0	0.7327	0	0	0
V30	0.0605	0.169	0.1031	0.461	0	0
	0	0.2433	0	0.214	0	0
V38	0.1321	0.003	0.2737	0.0303	0.0698	-0.0147
	0.1836	0	0.1761	0	0	0
V43	-0.0953	-0.0563	-0.2014	-0.1241	0	0
	-0.1425	0	-0.1263	0	0	0
V57	0.0635	-0.003	0.124	-0.004	0	0
	0.0808	0	0.0698	0	0	0
V62	-0.0017	-0.0126	-0.0031	-0.0313	0	0
	0	-0.0195	0	-0.0181	0	0
V81	0	0	-0.1516	-0.0037	0	0
	-0.0971	0	-0.0773	0	0	0
V85	0.1022	-0.0354	0.2143	-0.0798	0	0
	0.1044	0	0.0825	0	0	0
V94	0	0	0.0514	0.0092	0	0
	0	0	0	0	0	0
V95	0	0	-0.0064	-0.0189	0	0
	0	0	0	0	0	0
V109	0.2747	0.1241	0.7759	0.3395	0.0873	0.0211
	0.399	0	0.3451	0	0.0805	0
V119	0	0	0	0	0	0
	0	-0.0413	0	-0.038	0	0
V127	0	0	-0.0176	-0.0049	0	0
	0	0	0	0	0	0
V136	0.0342	0.003	0.07	0.0109	0	0
	0.0676	0	0.062	0	0	0
V140	0	0	0.0413	0.0126	0	0
	0	0	0	0	0	0
V145	-0.0039	-0.0248	-0.0156	-0.0607	0	0
	0	-0.0375	0	-0.0324	0	0
V174	0	0	0.095	0.0561	0	0
	0.068	0	0.0628	0	0	0
V188	0	0	0.043	0.0383	0	0
	0	0	0	0	0	0
V204	0	0	-0.1386	0.0442	0	0
	-0.0813	0	-0.0683	0	0	0
V212	0.0475	0.0244	0.1349	0.079	0	0
	0.0495	0	0.0428	0	0	0

Tabelle 6.3: Gleason-Score-Datensatz: Modellkoeffizienten der Zielgrößenkategorien (jeweils zwei Spalten pro Modellwahlkriterium) für adaptives Group Lasso (grau hinterlegte Zeilen) und adaptives Sparse Group Lasso.

Koeffizientenpfade

Die Koeffizientenpfade der beiden Penalisierungsvarianten sind für die ausgewählten Variablen V9, V85 und V109 in Abbildung 6.5 dargestellt. Die Koeffizientenpfade aller Variablen, die einen von null verschiedenen Koeffizienten bei minimaler Penalisierung erhalten haben, sind zudem im Anhang, in den Abbildungen B.5, B.6, B.7 und B.8 auf Seite 82 ff., zu finden.

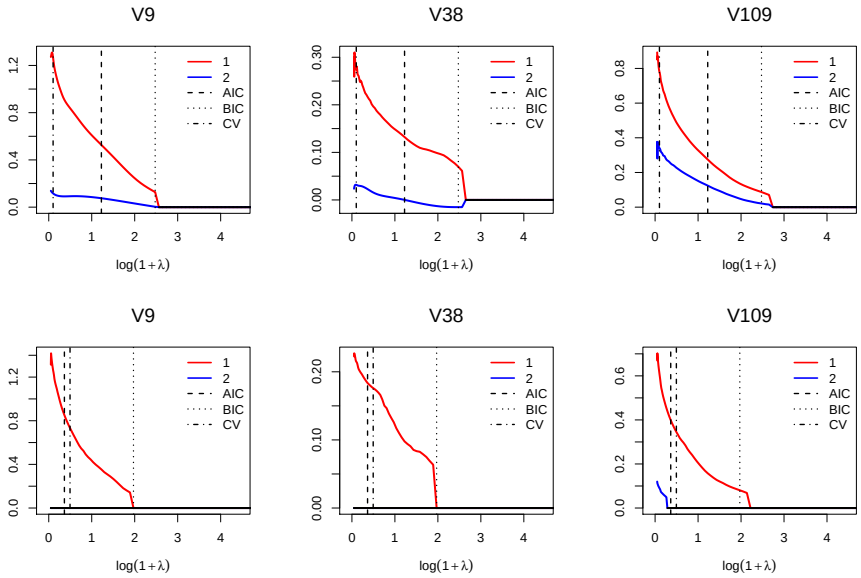

Abbildung 6.5: Gleason-Score-Datensatz: Koeffizientenpfade, der unter adaptivem Group Lasso (obere Zeile) und adaptivem Sparse Group Lasso (untere Zeile) bestimmten Modellkoeffizienten.

Die Koeffizientenpfade der zu Responsekategorie 1 gehörigen Koffizienten sind rot gefärbt, die zu Responsekategorie 2 gehörigen blau. Die Schnittpunkte der Pfade mit den vertikalen, gestrichelten Linien entsprechen den Koeffizientenwerten, der mittels AIC, BIC oder zehnfacher Kreuzvalidierung (CV) gewählten Modellen. Der minimale Penalisierungsgrad wird an den linksseitigen Enden der Pfade erzeugt.

Dargestellt sind die drei Genvariablen, die selbst unter dem restriktiven BIC-Kriterium in das Modell aufgenommen werden. In der oberen Zeile von Abbildung 6.5 befinden sich die unter adaptiver GL-Penalisierung bestimmten Koeffizienten, in der unteren Zeile diejenigen, die unter adaptiver SGL-Penalisierung geschätzt wurden. Es zeigt sich in diesen sechs Grafiken ein generell schwächerer Effekt für die zweite Responsekategorie, dessen Koeffizienteninterpretation in engem Zusammenhang zu hohen Gleason-Scores stehen würde. Diese Effekte werden durch SGL-Penalisierung bereits bei relativ geringer Penalisierungsstärke auf null geschätzt, sodass diese unberücksichtigt bleiben. Aus der ersten Zeile (GL-Penalisierung) der Koeffizientenpfade wird ersichtlich, dass das BIC-optimale Modell für denjenigen Penalisierungsgrad resultiert, der diese drei Variablen gerade noch als von null verschieden schätzt. Das BIC-optimale Modell der SGL-

Penalisierung liegt dort, wo zwar gerade noch Variable V109 einen positiven Koeffizienten erhält, allerdings werden für diesen Penalisierungsgrad gerade keine von null verschiedenen Koeffizienten für die anderen beiden Variablen erzielt. Alle drei Variablen erhalten positive Koeffizienten und erhöhen somit die Chance in der betrachteten Gleason-Score-Kategorie zu verbleiben, wenn das Genexpressionslevel steigt. Aus den im Anhang dargestellten Koeffizientenpfaden wird deutlich, dass eine Vielzahl der Variablen sowohl für das AIC-optimale, als auch für das CV-optimale Modell von null verschiedene Koeffizienten erhält, die selbst bei steigendem Penalisierungsgrad von null verschiedene Koeffizienten behalten, allerdings nicht mehr im BIC-optimalen Modell berücksichtigt werden.

Anhand der in Kapitel 5 vorgestellten Simulationsergebnisse aus Modell 2.1 in Szenario 2 und weiterer durchgeführter Simulationen, deren Auswertungen nicht in diese Arbeit eingebunden wurden, ist anzumerken, dass datenarme Schätzsituationen zu extrem hohen Falsch-Negativ-Raten führen, somit die Gefahr besteht, eine große Anzahl von Genen, fälschlicherweise als irrelevant (Null-Koeffizient) einzustufen.

6.3 Zusammenfassung

In diesem Kapitel wurden das Group und das Sparse Group Lasso mit adaptiver Gewichtung, in einer penalisierten ML-Schätzung sequentieller Logit-Modelle, auf den Gründer- und den Gleason-Score-Datensatz angewandt. Für jeden der beiden Datensätze wurden, über ein Gitter von 100 verschiedenen Penalisierungsparametern, Modelle mit adaptiver GL- oder SGL-Penalisierung geschätzt. Die Auswahl der jeweils optimalen Modelle fand mit Hilfe von Modellwahlkriterien (AIC, BIC, CV) statt.

Mit 1224 Beobachtungseinheiten und 150 zu schätzenden Parametern, bei sieben Responsekategorien und 15 (kategorialen) Prädiktoren, lag für das Modell des Gründer-Datensatzes eine sehr datenreiche Schätzsituation vor. Für die beiden AIC-optimalen Modelle, eines unter GL-Penalisierung, das andere unter SGL-Penalisierung, wurden 5 der 15 Prädiktoren ohne Einfluss geschätzt. Diese Prädiktoren waren Standort, Neugründung, Zielmarkt, Geschlecht und Berufserfahrung des Gründers. Für die BIC-optimalen Modelle erhielten die Prädiktoren Rechtsform und Startkapital für fast alle kategoriespezifischen Koeffizienten von null verschiedene Einflüsse. Dadurch, dass die sieben Responsekategorien aufeinanderfolgende Zeitintervalle beschreiben, fällt die Analyse von Unternehmensinsolvenzen gleichermaßen in das Gebiet der Survivaldaten.

Für das Modell des Gleason-Score-Datensatzes lag, mit 52 Beobachtungen und 490 zu schätzenden Parametern, bei drei Responsekategorien und 244 metrischen Einflussgrößen, eine datenarme Schätzsituation vor. Unter adaptiver Group Lasso-Penalisierung (Sparse Group Lasso-Penalisierung) wurden für das AIC-optimale Modell 20 (16) der 244 Genexpressionslevel selektiert, für das BIC-optimale Modell noch 3 (1) Prädiktor(-en).

Eine Veränderung der geschätzten kategoriespezifischen Effekte, in Abhängigkeit der Penalisierungsstärke, wurde mit Hilfe von Koeffizientenpfaden für jede Zielgrößenkategorie grafisch veranschaulicht.

Kapitel 7

Zusammenfassung

In diesem abschließenden Kapitel werden zunächst die zentralen Aspekte der theoretischen Grundlagen, die zu einer Verwendung likelihoodbasierter, koeffizientengruppierender Penalisierungsansätze im allgemeinen sequentiellen Logit-Modell geführt haben, zusammengefasst. Weiterhin wird ein Überblick über die Ergebnisse der verschiedenen Simulationsszenarien und Datenauswertungen gegeben. Parallel hierzu werden Modifikations- und Erweiterungsmöglichkeiten angedeutet, die im Verlauf der Analyse aufgefallen sind, aber nicht weiter verfolgt werden konnten.

In Kapitel 2 wurde die allgemeine Struktur generalisierter linearer Modelle, sowie deren Erweiterung auf multivariate Responsevariablen erläutert. Diese Erweiterung erlaubt es, unter Verwendung der Multinomialverteilung als stochastische Komponente und einer vektorwertigen Funktion zur Verknüpfung der Zielgrößenvariable mit dem linearen Prädiktor, die beiden ordinalen Regressionsmodelle - das kumulative und das sequentielle Modell - in das allgemeine GLM-Rahmenwerk einzubinden. Dies ermöglicht eine Maximum-Likelihood Schätzung der Regressionskoeffizienten mittels iterativer Verfahren und bildet den Anknüpfungspunkt zu likelihoodbasierten Penalisierungsansätzen. Ehe diese Penalisierungsansätze betrachtet wurden, deren Koeffizientenschätzung und Selektionsfähigkeit von den Charakteristika eines gegebenen Modelltyps abhängen, wurden in Kapitel 3 die beiden ordinalen Regressionsmodelle formuliert. Während das sequentielle Modell in seiner Anwendung auf ordinale Zielgrößen, deren Kategorien nur sukzessive erreichbar sind, beschränkt ist, ist diese Einschränkung für das kumulative Modell nicht erforderlich. Dennoch wurden die Penalisierungsansätze auf das sequentielle Modell angewandt, da an dieses Modell weniger Parameterrestriktionen gebunden sind und Schätzer leichter gewonnen werden können. Ein weiterer Vorteil ist, dass durch das sequentielle Modell Fragestellungen der Survival-Analyse berücksichtigt werden können, sofern die abhängige Zeitvariable diskrete Werte annimmt. Wirft eine konkrete Fragestellung ein Modell auf, das eine große Anzahl von Responsekategorien oder (kategoriespezifische) Prädiktoren besitzt, besteht die Gefahr, dass bei unzureichender Datenlage keine ML-Schätzer mehr generiert werden können oder diese instabil sind. Neben der Erzeugung von stabilisierten ML-Schätzern durch eine Penalisierung der log-Likelihood, gelingt mit den in Kapitel 4 dargestellten Ansätzen, eine Selektion von Prädiktoren. Penalisierungsansätze, wie das

Group oder das Sparse Group Lasso sind in der Lage, die Zugehörigkeit mehrere Koeffizienten zu einem Prädiktor, durch eine gruppierte Penalisierung, zu berücksichtigen und somit implizit eine Variablenselektion zu erzielen, indem alle Koeffizienten, einer zu einem Prädiktor gehörigen Gruppe, auf null geschätzt werden.

Derartige Selektionseigenschaften wurden neben Schätz- und Prädiktionsgüte, in den Simulationsszenarien des Kapitel 5, für verschiedene Penalisierungsansätze miteinander verglichen. In Szenario 1 wurde festgestellt, dass eine Fehlspezifikation der wahren Koeffizientenstruktur weniger die prädiktive Devianz beeinflusst, umso mehr aber die Qualität der geschätzten Koeffizienten und Wahrscheinlichkeiten, sowie die Selektion relevanter Variablen. Unter der Annahme, dass sich die wahre Koeffientenstruktur aus einer Mischung globaler und kategoriespezifischer Effekte zusammensetzt, lässt sich überlegen, ob es für Schätzung, Prädiktion und Selektion von Vorteil sein könnte, mit Hilfe von Vorwissen für einzelne Prädiktoren die Schätzung globaler Effekte zu forcieren. Eine Simulation, mit dem in Szenario 2 gegebenen Modell gemischter Koeffizienteneffekte, für das die globalen und kategoriespezifischen relevanten Variablen im Penalisierungsansatz korrekt spezifiziert wurden, hat keine wesentlichen Unterschiede zu einer Schätzung mit vollständig kategoriespezifischen Effekten gezeigt. Wird eine Schätzung kategoriespezifischer Effekte durchgeführt, erhält jede Kovariable kategoriespezifische Effekte, unabhängig davon, ob die wahre Struktur der Effekte global ist. Ein Ansatz, der diesem Problem begegnen könnte, bestünde in der Konstruktion eines Penalisierungsterms, der gleichartige Effekte zueinanderhinschrumpft, somit implizit globale Effekte für Prädiktoren schätzen kann. Hiervon ausgehend könnten für einzelne kategoriespezifische Effekte Abweichungen von dem globalen Effektniveau eines Prädiktors bestimmt werden, sodass ein kategoriespezifischer Effekt, für alle Koeffizienten einer Gruppe, in eine kategoriespezifische und eine für diese Gruppe globale Komponente zerlegt werden könnte.

Im Szenario 3 und Modell 2.1 des Szenarios 2, die eine solide Datengrundlage zur Verfügung hatten, wurde die Überlegenheit von Penalisierungsvarianten, die gruppiert kategoriespezifische Effekte penalisieren, gegenüber einer unpenalisierten ML-Schätzung, der Schätzung (un-) penalisierter globaler Effekte, sowie der klassischen Lasso-Penalisierung deutlich. Eine adaptive Group Lasso-Penalisierung schneidet dabei geringfügig besser ab, als eine adaptive Sparse Group Lasso-Penalisierung. Dabei sind adaptive Gewichte, aufgrund stabilerer Schätzer und geringerer Falsch-Negativ-Raten, einer Penalisierung ohne Gewichte oder einem Refit vorzuziehen. Problematisch ist allerdings, dass sofern die Datensituation nicht deutlich über der Anzahl zu schätzender Koeffizienten liegt, wie in Szenario 3, Falsch-Positiv-Raten sehr hoch sind. In allen Szenarien liefert die Schätzung (un-) penalisierter globaler Effekte zwar generell eine gute prädiktive Devianz, aber hohe FNR und ein erhöhtes MSE-Niveau.

In datenarmen Schätzsituationen (Modell 2.2 in Szenario 2, Szenario 4) besteht neben instabileren Schätzern ein hohes Risiko relevante Variablen fälschlicherweise aus dem Modell zu entfernen. Je ungünstiger die Datenlage wird, desto höher werden auch die Falsch-Negativ-Raten. Es gelingt keiner der Penalisierungsvarianten in (deutlich) überparametrisierten Modellsituationen geringe Fehlselektionsraten aufzuweisen, womit sich diese Ansätze in datenarmen Situationen hinsichtlich einer Zweckmäßigkeit der Variablenselektion in Frage stellen lassen.

In den bisher durchgeführten Simulationen wurden maximal zehn Responsekategorien berücksichtigt. Vor allem in Bezug auf Survival-Datensätze, kann die Anzahl diskretisierter Zeitkategorien deutlich höher liegen, sodass diesbezüglich Untersuchungsbedarf besteht. In einem Szenario mit vielen Responsekategorien und einer geringen Anzahl Prädiktoren, kann unter Umständen eine einfache (adaptive) Lasso-Penalisierung genügen, um relevante kategoriespezifische Effekte zu selektieren.

Des Weiteren wurde für Prädiktoren mit kategoriespezifischen Effekten eine verhältnismäßig geringe Anzahl wahrer Null-Koeffizienten angenommen. In Simulationen, in die unter die wahren kategoriespezifischen Effekte ein Anteil von durchschnittlich 20-25 % Null-Koeffizienten gemischt wurden, die dieser Auswertung nicht hinzugefügt worden sind, konnte kein positiver Effekt auf Schätz- und Selektionsgüte durch eine Sparse Group Lasso-Penalisierung verglichen mit einer Group Lasso Penalisierung festgestellt werden. In weiteren Untersuchungen könnte die Null-Koeffizientenquote erhöht und gleichzeitig eine Struktur der Null-Koeffizienten eingebunden werden, um wahre Nulleffekte für aufeinanderfolgende Responsekategorien zu simulieren.

In Kapitel 6 wurde eine Penalisierung im sequentiellen Logit-Modell, unter adaptivem Group und adaptivem Sparse Group Lasso, auf die beiden Datensätze Gründer und Gleason-Score angewandt. In der datenreichen Schätzsituation des Modells des Günder-Datensatzes, wurden für die beiden AIC-optimalen Modelle 10 der 15 Prädiktoren selektiert, für die BIC-optimalen Modelle erhielten die Prädiktoren Rechtsform und Startkapital für fast alle kategoriespezifischen Koeffizienten, von null verschiedene Einflüsse. Für das Modell des Gleason-Score-Datensatzes lag hingegen eine datenarme Schätzsituation vor. Unter adaptiver Group Lasso-Penalisierung (Sparse Group Lasso-Penalisierung) wurden für das AIC-optimale Modell 20 (16) der 244 Genexpressionslevel selektiert, für das BIC-optimale Modell noch 3 (1) Prädiktor(-en). Die Auswirkungen verschiedener Penalisierungsstärken auf die Werte der geschätzten Koeffizienten, wurden mit Hilfe von Koeffizientenpfaden für jede Zielgrößenkategorie grafisch veranschaulicht.

In dieser Ausarbeitung wurde verdeutlicht, dass ein, für ein multivariates Regressionsmodell geeignetes, likelihoodbasiertes Penalisierungskriterium, hinsichtlich einer Selektion von Variablen, erkennen muss, dass ein Prädiktor durch eine Gruppe von kategoriespezifischen Koeffizienten auftritt. Erst wenn alle Koeffizienten dieser Gruppe auf null geschätzt werden, kann diese Variable aus dem Modell entfernt werden. Solange dies bei der Konstruktion berücksichtigt wird, ist ein adäquates Penalisierungsfunktional nicht auf das Group oder das Sparse Group Lasso beschränkt. Ebenso ist eine Variablenselektion nicht auf Ansätze mit einer penalisierten Likelihoodfunktion beschränkt.

Ideen, die in hochparametrisierten Modellen ebenfalls Variablenselektion erzielen, stammen aus der Informatik und dem Maschinellen Lernen. Zu einflussreichen Ansätzen zählen *Boosting*, *Support Vector Machines* oder *Random Forests*. (Vgl. Hofner et al. (2009)) Boosting (Schapire, 1990; Freund & Schapire, 1996), dessen Grundidee es ist, eine Basisprozedur durch schrittweise Gewichtung der Zwischenergebnisse zu verbessern, lässt sich bspw. für generalisierte lineare, generalisierte additive Modelle und Regressionsmodelle für Survivalanalysen adaptieren. (Vgl. Bühlmann & Hothorn (2007))

Anhang A

Theoretische Grundlagen

A.1 Die Multinomialverteilung

Die Zielgröße des multivariaten Regressionsmodells sei für Beobachtungseinheit i die Realisation einer von k Kategorien, mit $Y_i \in \{1, \ldots, k\}$. Dabei trete Kategorie r mit der kategoriespezifischen Wahrscheinlichkeit $P(Y_i = r) = \pi_{ir}$ ein, für alle $r = 1, \ldots, k$. Die k kategoriespezifischen Wahrscheinlichkeiten summieren sich hierbei zu 1. In Abschnitt 2.3 wird eine redundanzfreie vekorwertige Darstellung für Beobachtungseinheit i mit Hilfe des dummycodierten q-dimensionalen Zielgrößenvektors $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})'$ eingeführt, dessen r-ter Eintrag eine 1 annimmt, wenn die Zielgröße in diese Kategorie fällt, für alle anderen Werte eine 0, mit $r = 1, \ldots, q$, wobei $q = k - 1$. Fällt die Zielgröße in die Referenzkategorie k, ist der Zielgrößenvektor ein Nullvektor. Die q Wahrscheinlichkeiten lassen sich ebenfalls in einem q-dimensionalen Vektor $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ darstellen, mit $\pi_{ik} = 1 - \sum_{r=1}^{q} \pi_{ir}$.

In ihrer allgemeinen Form beschreibt die Multinomialverteilung, für eine Stichprobe des Umfangs m, die Anzahl der Beobachtungen, die in Kategorie r fallen. Somit geben die Einträge des k-dimensionalen Vektors $\mathbf{y} = (y_1, \ldots, y_k)'$, die Anzahl der in Kategorie r gezählten Einheiten an. (Vgl. Tutz (2012), S. 209.) Dieser Vektor besitzt die Dichtefunktion:

$$f(\mathbf{y}) = \frac{m!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdot \ldots \cdot \pi_k^{y_k} \quad \text{mit} \quad y_r \in \{1, \ldots, m\}, \quad \sum_{r=1}^{k} y_r = m \tag{A.1}$$

$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ bezeichne den Vektor der kategoriespezifischen Auftretenswahrscheinlichkeiten. Ebenso, wie der oben genannte Responsevektor, ist auch diese Darstellung nicht redundanzfrei, da sich beispielsweise aus $q = k-1$ Kategorien errechnen lässt, wieviele der m Beobachtungen in der verbleibenden Kategorie sind. Eine redundanzfreie Dichtefunktion mit $\mathbf{y} = (y_1, \ldots, y_q)'$ und $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q)'$ ist durch

$$f(\mathbf{y}) = \frac{m!}{y_1! \cdot \ldots \cdot y_q!(m - y_1 - \ldots - y_q)!} \pi_1^{y_1} \cdot \ldots \cdot \pi_q^{y_q} \cdot (1 - \pi_1 - \ldots - \pi_q)^{(m - y_1 - \ldots - y_q)} \tag{A.2}$$

gegeben. Der Vektor $\mathbf{y}$ mit dieser Dichtefunktion folgt dann einer Multinomialverteilung mit den Parametern m und $\boldsymbol{\pi}$: $\mathbf{y} \sim M(m, \boldsymbol{\pi})$. Der Erwartungswert der r-ten Komponente berechnet sich als $\mathbb{E}(y_r) = m\pi_r$, die Varianz als $\mathbb{V}(y_r) = m\pi_r(1 - \pi_r)$. Für die Kovarianz zweier Kategorien r und s, mit $r \neq s$ gilt: $cov(y_r, y_s) = -m\pi_r\pi_s$. In Matrixdarstellung ergibt sich für den Erwartungswertvektor und die Varianz-Kovarianzmatrix:

$$\mathbb{E}(\mathbf{y}) = \begin{pmatrix} m\pi_1 \\ \vdots \\ m\pi_q \end{pmatrix}; \quad Cov(\mathbf{y}) = \begin{pmatrix} m\pi_1(1-\pi_1) & \cdots & -m\pi_1\pi_q \\ \vdots & \ddots & \vdots \\ -m\pi_q\pi_1 & \cdots & m\pi_q(1-\pi_q) \end{pmatrix} \tag{A.3}$$

Für den Responsevektor $\mathbf{y}_i$ des multivariaten Modells für Beobachtung i, resultiert der Spezialfall der Multinomialverteilung mit m=1, sodass $\mathbf{y}_i \sim M(1, \boldsymbol{\pi}_i)$ mit der in Abschnitt 2.3 angegebenen Dichtefunktion, Erwartungswertvektor und Varianz-Kovarianzmatrix.

Anhang B

Anwendungsbeispiele

B.1 Gründerdatensatz

Auf den folgenden Seiten sind die Koeffizientenpfade aller im sequentiellen Logit-Modell des Gründerdatensatzes berücksichtigten Kovariablen, dargestellt. In den Abbildungen B.1 und B.2 die Koeffizientenpfade des Modells mit adaptiver Group Lasso-Penalisierung, in den Abbildungen B.3 und B.4 die des Modells mit adaptiver Sparse Group Lasso-Penalisierung. Die Färbung der verschiedenen Koeffizientenpfade je Zielgrößenkategorie ergibt sich wie folgt: schwarz (Kategorie 1), rot (Kategorie 2), grün (Kategorie 3), blau (Kategorie 4), türkis (Kategorie 5), pink (Kategorie 6). Die Schnittpunkte der Pfade mit den vertikalen, gestrichelten Linien entsprechen den Koeffizientenwerten, der mittels AIC, BIC oder zehnfacher Kreuzvalidierung (CV) gewählten Modellen.

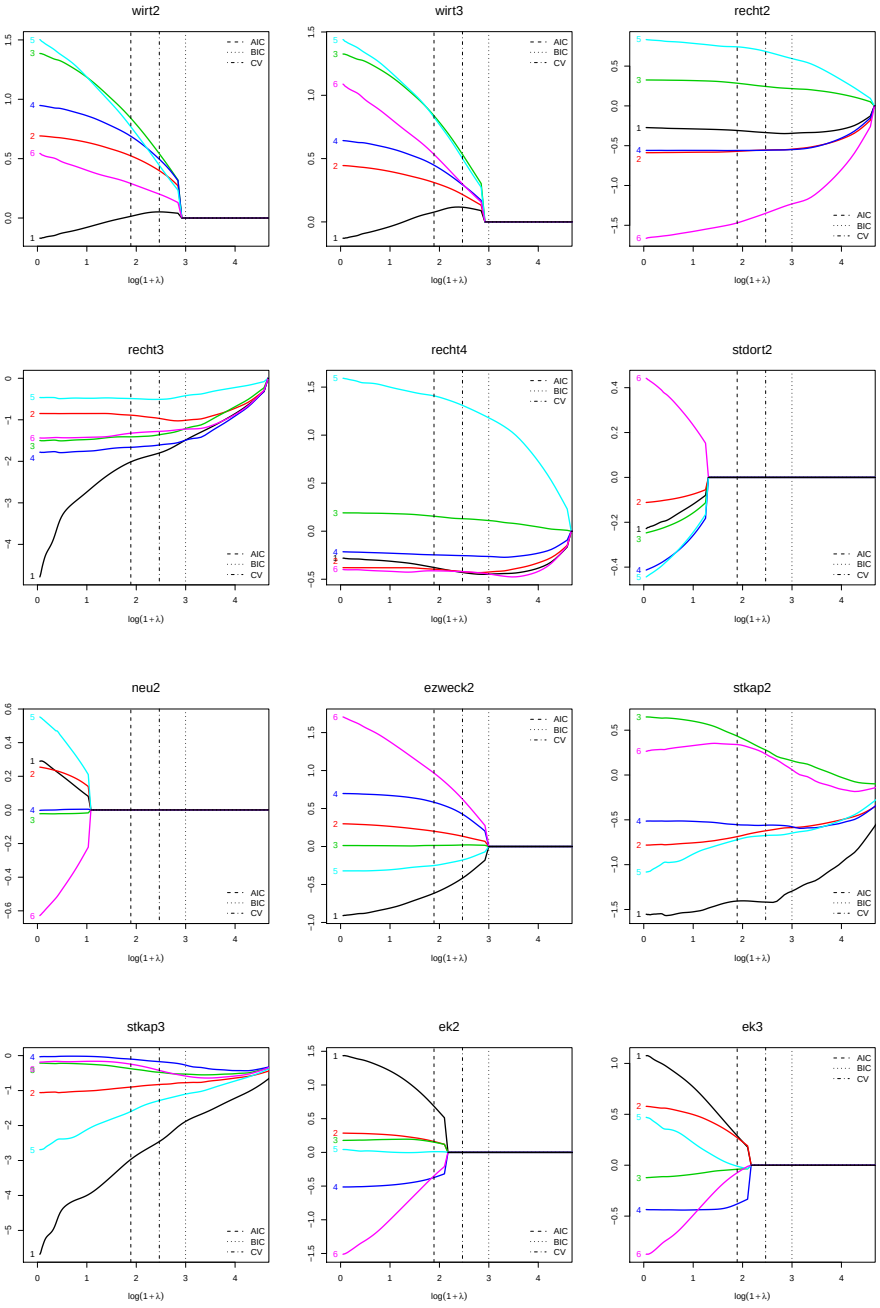

Abbildung B.1: Gründerdatensatz: Pfade, der unter adaptiver Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

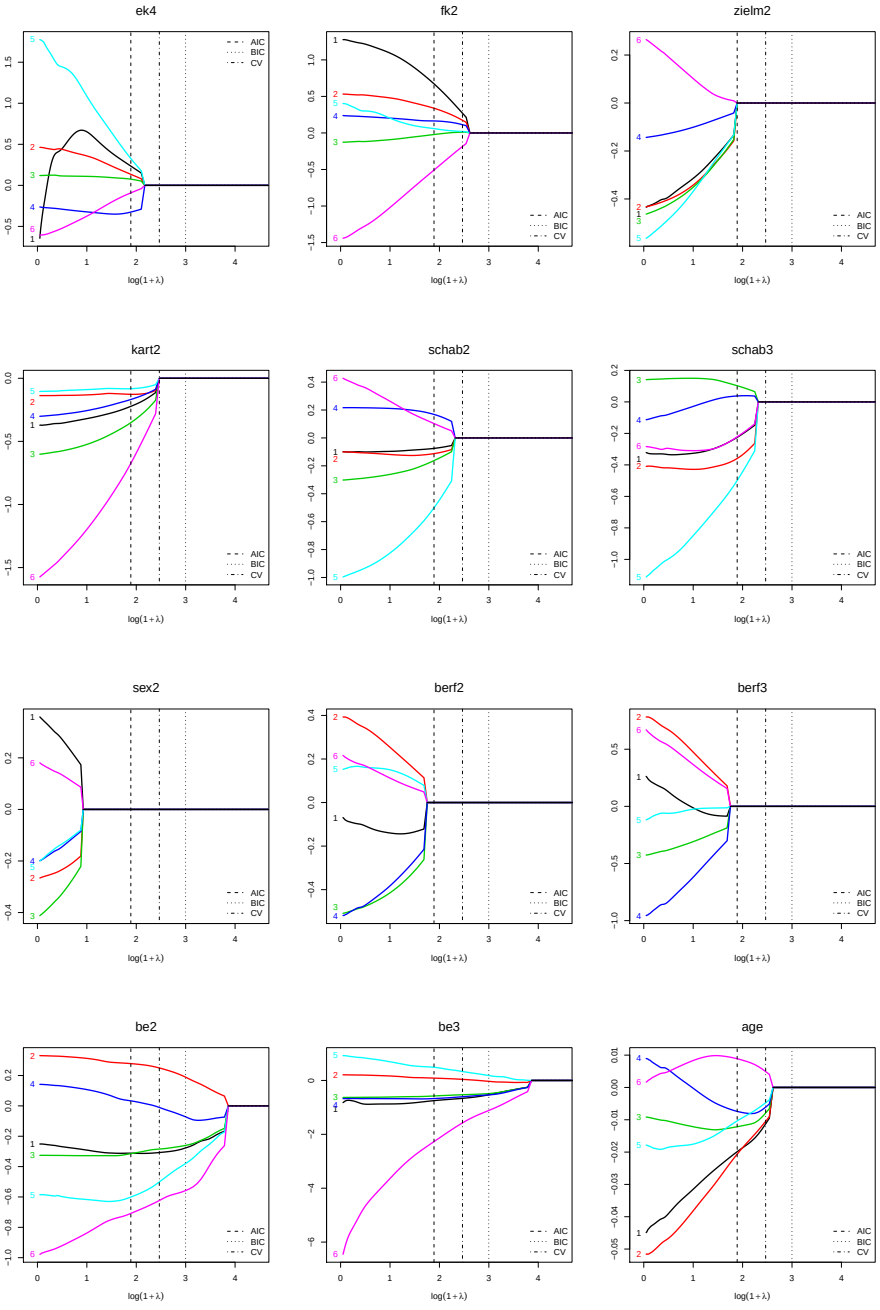

Abbildung B.2: Gründerdatensatz: Pfade, der unter adaptiver Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

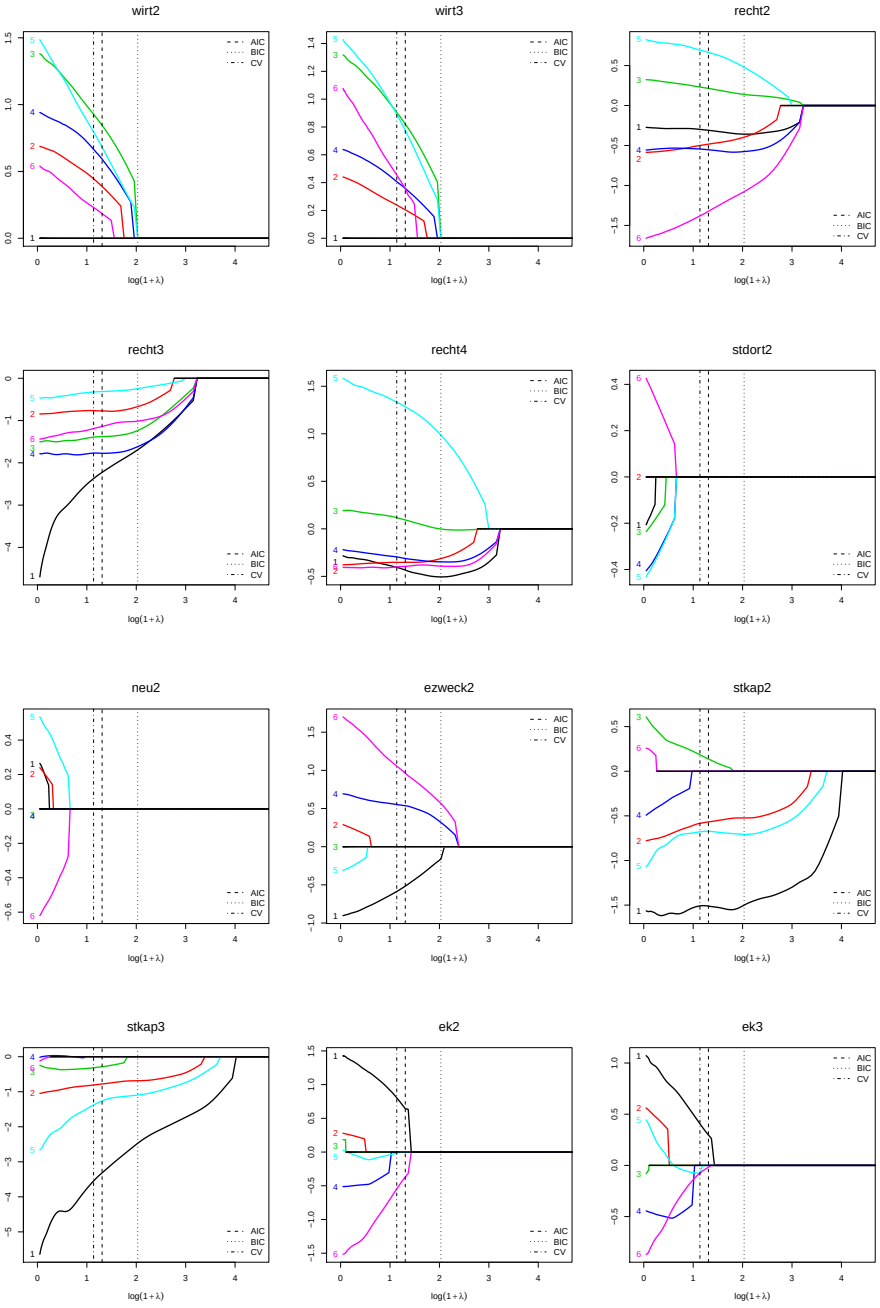

Abbildung B.3: Gründerdatensatz: Pfade, der unter adaptiver Sparse Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

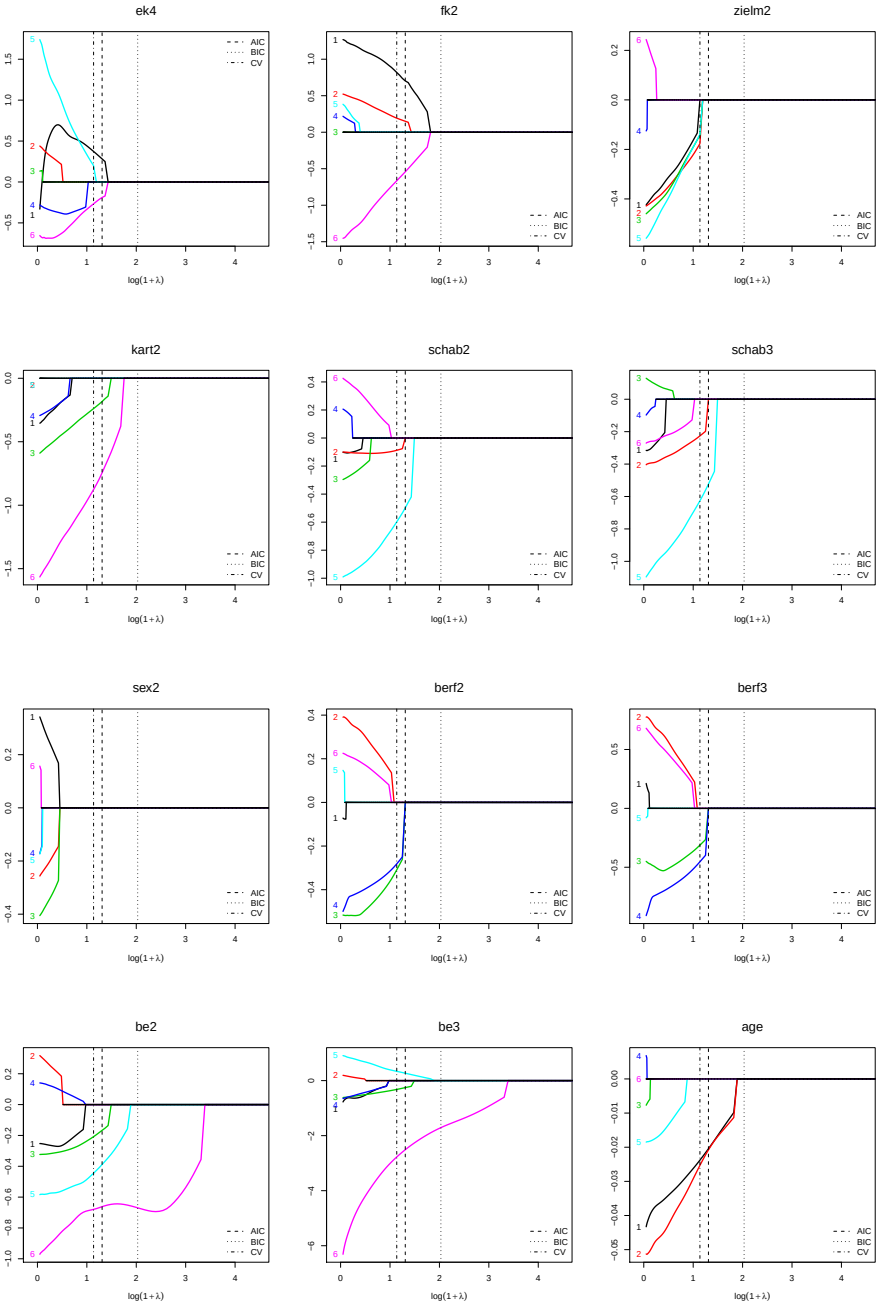

Abbildung B.4: Gründerdatensatz: Pfade, der unter adaptiver Sparse Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

B.2 Gleason-Score-Datensatz

Auf den folgenden Seiten sind die Koeffizientenpfade aller im sequentiellen Logit-Modell des Gleason-Score-Datensatzes berücksichtigten Kovariablen, mit von null verschiedenen Koeffizientenwerten, inklusive Intercepts dargestellt. In den Abbildungen B.5 und B.6 die Koeffizientenpfade des Modells mit adaptiver Group Lasso-Penalisierung, in den Abbildungen B.7 und B.8 die, des Modells mit adaptiver Sparse Group Lasso-Penalisierung. Die Färbung der verschiedenen Koeffizientenpfade je Zielgrößenkategorie ergibt sich wie folgt: rot (Kategorie 1), blau (Kategorie 2). Die Schnittpunkte der Pfade mit den vertikalen, gestrichelten Linien entsprechen den Koeffizientenwerten, der mittels AIC, BIC oder zehnfacher Kreuzvalidierung (CV) gewählten Modellen.

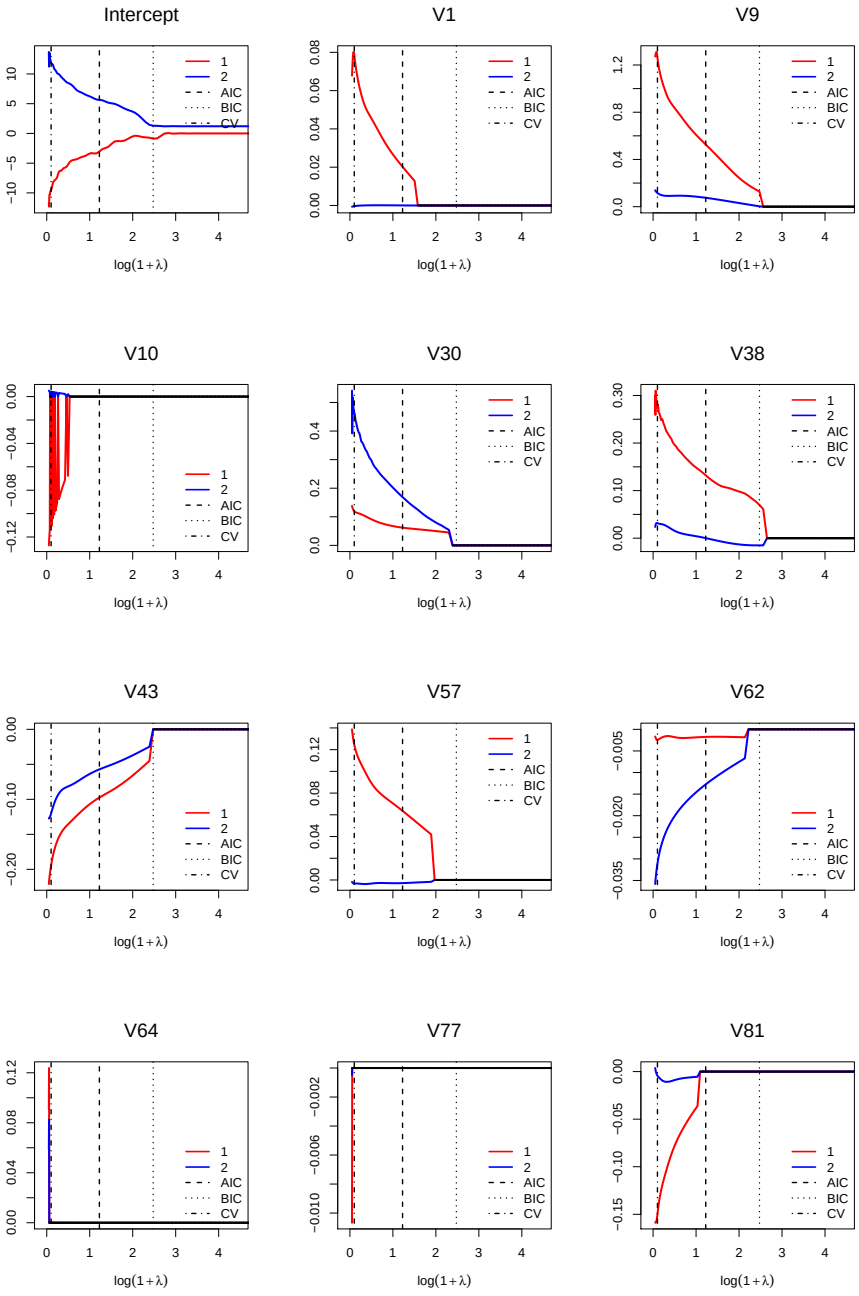

Abbildung B.5: Gleason-Score-Datensatz: Pfade, der unter adaptiver Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

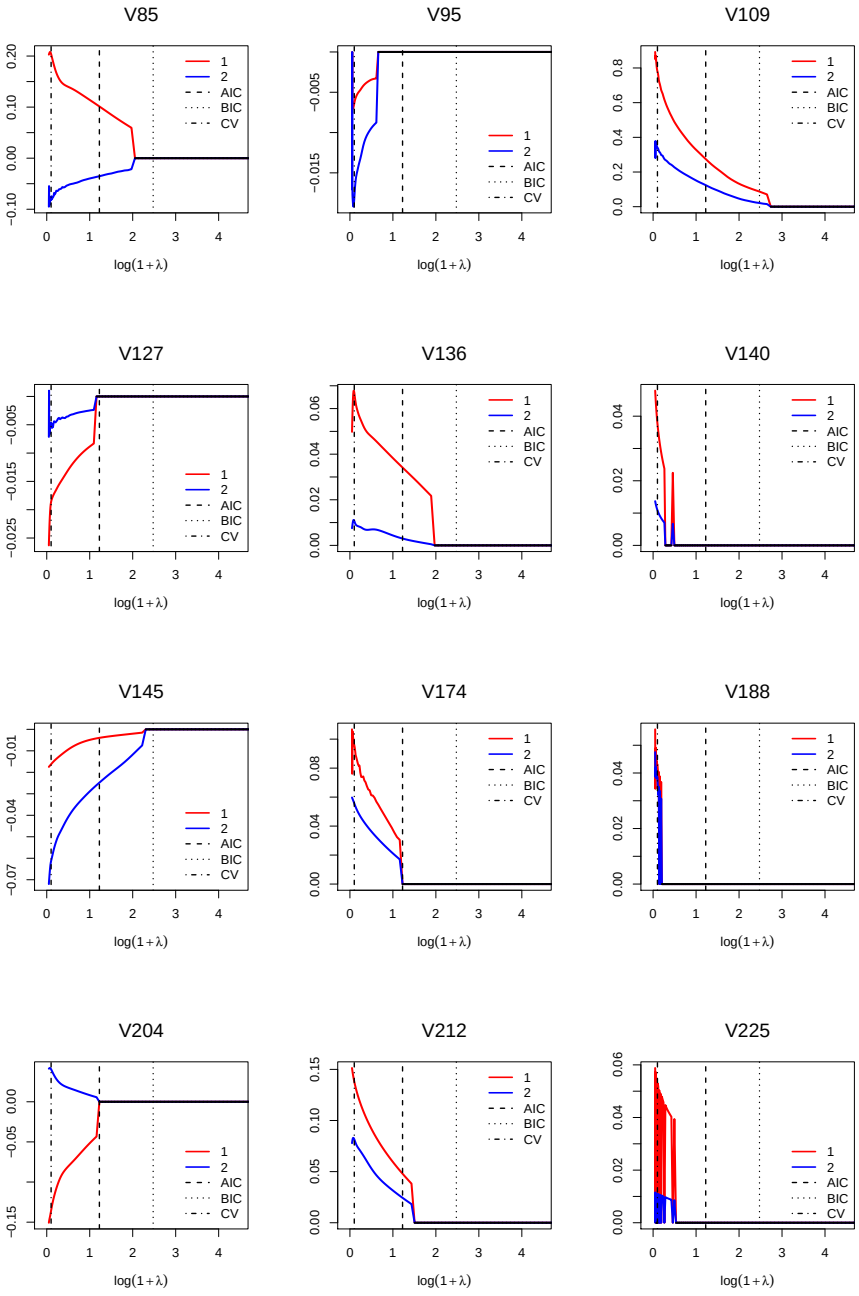

Abbildung B.6: Gleason-Score-Datensatz: Pfade, der unter adaptiver Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

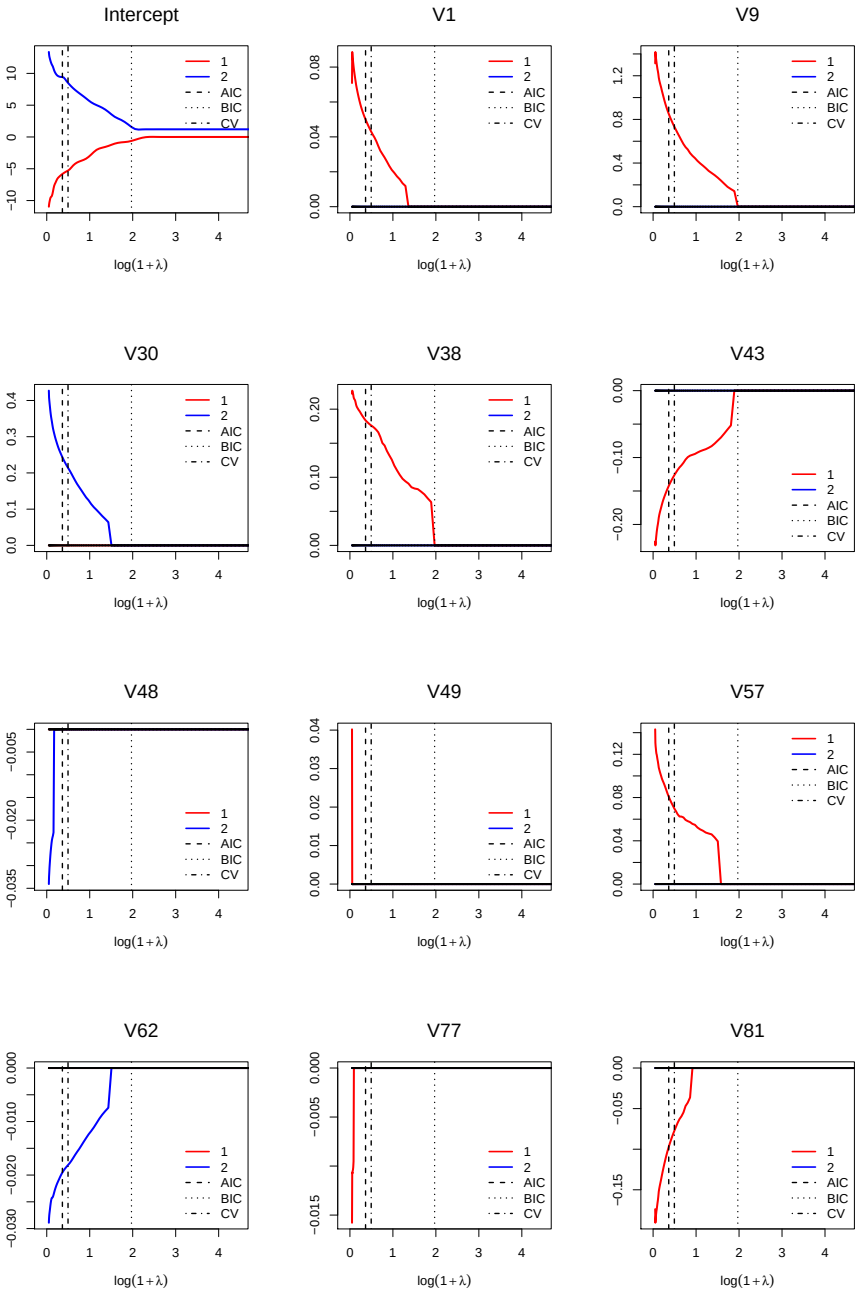

Abbildung B.7: Gleason-Score-Datensatz: Pfade, der unter adaptiver Sparse Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

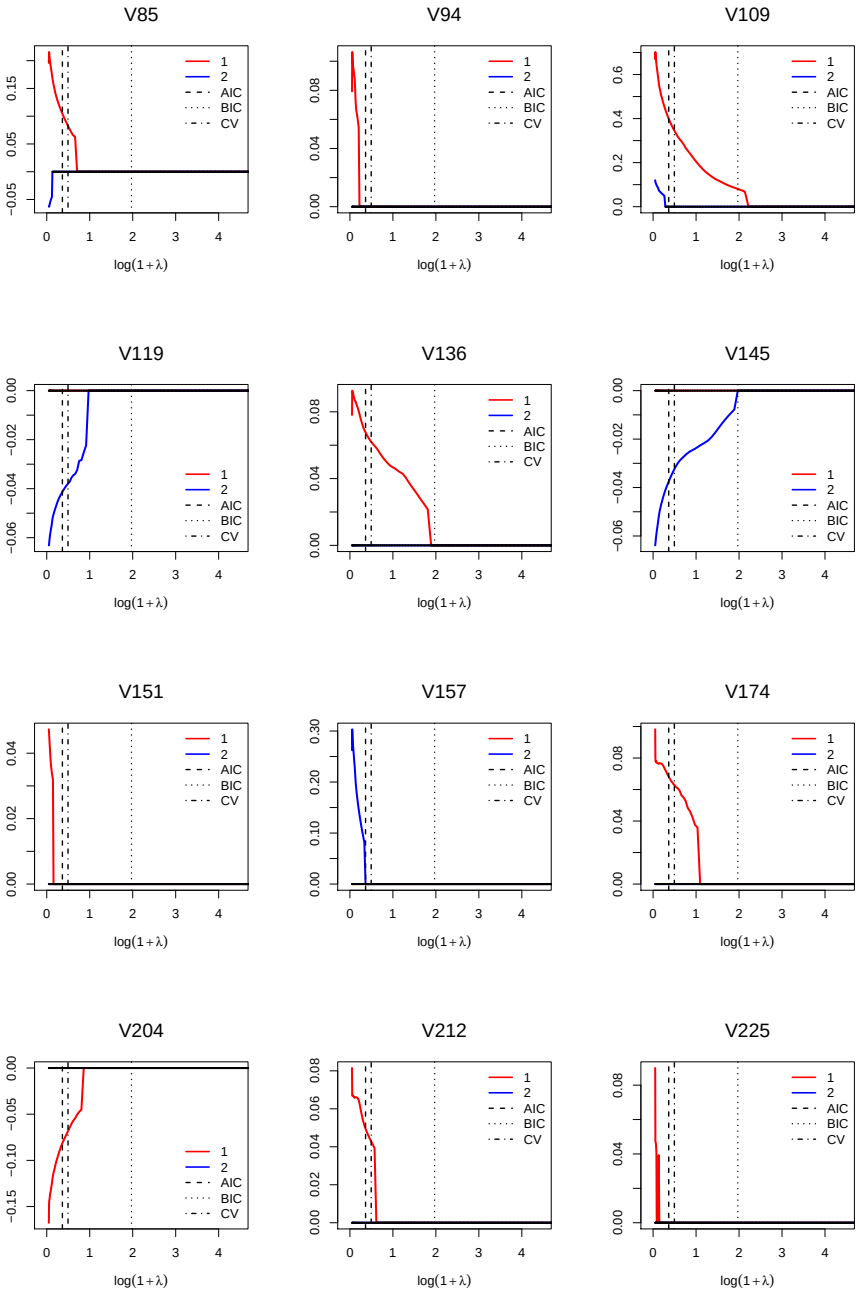

Abbildung B.8: Gleason-Score-Datensatz: Pfade, der unter adaptiver Sparse Group Lasso-Penalisierung bestimmten Modellkoeffizienten.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Agresti, A. (2007), 'An Introduction to Categorical Data Analysis', 2. Auflage, John Wiley & Sons, Hoboken, New Jersey.

Anderson, J.A. (1984), 'Regression and Ordered Categorical Variables', in: Journal of the Royal Statistical Society, Series B (Vol. 46, No. 1), 1-30.

Baade, P.D, Youlden, D.R & Krnjacki, L.J. (2009), 'International epidemiology of prostate cancer: Geographical distribution and secular trends', in: Molecular Nutrition & Food Research (Vol. 53, No. 2), 171-184.

Balakrishnan, N. (Hrsg.) & Rao, C.R. (Hrsg.) (2004), 'Handbook of Statistics 23 - Advances in Survival Analysis', 1. Auflage, Elsevier B.V., Amsterdam.

Beck, A. & Teboulle, M. (2009), 'A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems', in: SIAM Journal on Imaging Sciences (Vol. 2, No. 1), 183-202.

Brüderl, J., Preisendörfer, P. & Ziegler, R. (1992), 'Survival Chances of Newly Founded Business Organizations', in: American Sociological Review (Vol. 57, No. 2), 227-242.

Bühlmann, P. & Hothorn, T. (20007), 'Boosting Algorithms: Regularization, Prediction and Model Fitting', in: Statistical Science (Vol. 22, No. 4), 477-505.

Candes, E. & Tao, T. (2007), 'The Dantzig selector: statistical estimation when p is much larger than n', in: The Annals of Statistics (Vol. 35, No. 6), 2313-2351.

Chu, W. et al. (2005), 'Biomarker discovery in microarry gene expression data with Gaussian processes', in: Bioinformatics (Vol. 21, No. 16), 3385-3393.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least Angle Regression', in: The Annals of Statistics (Vol. 32, No. 2), 407-499.

Fahrmeir, L. et al. (2007), 'Statistik - Der Weg zur Datenanalyse', 5. Auflage, Springer-Verlag, Berlin.

Fahrmeir, L., Kneib, T. & Lang, S. (2009), 'Regression - Modelle, Methoden und Anwendungen', 2. Auflage, Springer-Verlag, Berlin.

Fahrmeir, L. & Tutz, G. (2001), 'Multivariate statistical modelling based on generalized linear models', 2. Auflage, Springer-Verlag, New York.

Frank, I.E. & Friedman, J.H. (1993), 'A Statistical View of Some Chemometrics Regression Tools', in: Technometrics (Vol. 35, No. 2), 109-135.

Freund, Y. & Schapire, R.E. (1996), 'Experiments with a New Boosting Algorithm', in: Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'A note on the group lasso and a sparse-group lasso', http://arxiv.org/abs/1001.0736, 1-8.

Hastie, T., Tibshirani, R. & Friedman, J. (2011), 'The Elements of Statistical Learning', Springer Series in Statistics, Kalifornien.

Hoerl, A.E. & Kennard, R.W. (1970), 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', in: Technometrics (Vol. 12, No. 1), 55-67.

Hofner, B. et al. (2009), 'A Framework for Unbiased Model Selection Based on Boosting', Technical Report 072, Department of Statistics, University of Munich.

Humphrey, P.A. (2004), 'Gleason grading and prognostic factors in carcinoma of the prostate', in: Modern Pathology (No. 17), 292-306.

Klein, J.P. & Moeschberger, M.L. (2003), 'Survival Analysis - Techniques for Censored and Truncated Data', 2.Auflage, Springer Verlag, New York.

McCullagh, P. (1980), 'Regression Models for Ordinal Data', in: Journal of the Royal Statistical Society, Series B (Vol. 42, No. 2), 109-142.

Nelder, J.A. & Wedderburn, R.W.M (1972), 'Generalized Linear Models', in: Journal of the Royal Statistical Society, Series A (Vol. 135, No. 3), 370-384.

R Development Core Team (2012), 'R: A Language and Environment for Statistical Computing', R Foundation for Statistical Computing, Wien, URL: http://www.R-project.org.

Robert Koch-Institut (2010), 'Krebs in Deutschland 2005/2006. Häufigkeiten und Trends', 7. Ausgabe, Robert Koch-Institut (Hrsg.) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (Hrsg.), Berlin.

Schapire, R.E. (1990), 'The Strength of Weak Learnability', in: Machine Learning (Vol. 5, No. 2), 197-227.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2012), 'A sparse-group lasso', in: Journal of Computational and Graphical Statistics, in press.

Singh, D. et al. (2002), 'Gene expression correlates of clinical prostate cancer behavior', in: Cancer Cell (Vol. 1), 203-209.

Stevens, S.S. (1946), 'On the Theory of Scales of Measurement', in: Science (Vol. 103, Nr. 2684), 677-680.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', in: Journal of the Royal Statistical Society (Vol. 58, Part 1), 267-288.

Tutz, G. (2000), 'Die Analyse kategorialer Daten: anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression', Oldenbourg Verlag, Wien.

Tutz, G. (2012), 'Regression für Categorical Data', Cambridge University Press, Cambridge.

Tutz, G., Pößnecker, W. & Uhlmann, L. (2012), 'Variable Selection in General Multinomial Logit Models', Technical Report 126, Department of Statistics, University of Munich.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', in: Journal of the Royal Statistical Society (Vol. 68, Part 1), 49-67.

Zou, H. (2006), 'The Adaptive Lasso and its Oracle Properties', in: Journal of the American Statistical Association (Vol. 101, No. 476), 1418-1429.

Zou, H., & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', in: Journal of the Royal Statistical Society (Vol. 67, Part 2), 301-320.