Judith KLOAS

Ordinale Regressionsmodelle

MASTERARBEIT

zur Erlangung des akademischen Grades einer Diplom-Ingenieurin

Masterstudium Finanz- und Versicherungsmathematik

Technische Universität Graz

Betreuer:

A.o. Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig FRIEDL

Institut für Statistik

Graz, im Januar 2014

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....

(Unterschrift)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1. Ordinale kategoriale Daten

In der heutigen Zeit ist die Statistik Basis und Hilfsmittel in vielen Forschungsgebieten. Seien es technische Analysen, Wetterbeobachtungen, medizinische Untersuchungen oder Bevölkerungsbefragungen - sobald man versucht Zusammenhänge zwischen verschiedenen Größen herzustellen, kann man unterschiedliche statistische Verfahren verwenden um entscheidende Schlüsse zu ziehen. In vielen Fällen werden kategoriale Daten untersucht. Solche Daten können in Gruppen (Kategorien) eingeordnet werden. Sie umfassen nominale und ordinale Daten. Nominale Merkmale sind beispielsweise das Geschlecht, die Haarfarbe oder Hobbys einer Person. Hier ist wirklich nur die Gruppierung möglich. Es gibt keinerlei Ordnungsstruktur und die Unterschiede der Merkmalsausprägungen sind nicht messbar. Im Vergleich dazu kann man ordinale Daten ihrer Größe nach ordnen, es besteht also eine Rangordnung zwischen ihnen. Oft findet man ordinale Daten bei Personenbefragungen wieder, bei denen beispielsweise etwas mit gut/mittel/schlecht oder selten/häufig/oft bewertet werden soll. Schulnoten sind ebenso ein Beispiel für ein ordinales Skalenniveau, bei denen den Merkmalsausprägungen sehr gut/gut/befriedigend/ausreichend/nicht genügend Zahlen zugeordnet werden. Man weiß, dass die 1 besser ist als eine 2, die 2 besser ist als eine 3 und so weiter. Durch die Zahlenzuweisung behauptet man, dass eine 1 genauso weit entfernt von einer 2 ist wie die 2 von einer 3. In Wahrheit kennt man den Abstand zwischen den verschiedenen Ausprägungen allerdings nicht. Wer kann schon sagen wie weit „gut“ von „sehr gut“ entfernt ist? Somit ist es unsinnig zu denken, dass eine 1 doppelt so gut ist wie eine 2 oder ähnliches. Solch ein Vergleich funktioniert nur bei metrisch skalierten Daten. Dazu zählen zum Beispiel das Einkommen oder das Alter einer Person. Werden metrisch skalierte Daten gruppiert, so resultiert daraus schließlich wieder ein ordinaler Datensatz.

Skala	Erklärung	Beispiel
Nominalskala	Man unterscheidet verschiedene qualitative Kategorien, welche diskret und nicht miteinander vergleichbar sind.	Beruf, Wohnort, Geschlecht
Ordinalskala	Die Kategorien sind qualitativ, diskret und ihrer Größe nach sortierbar. Es gibt keine fixen Abstände zwischen den Klassen.	Schulbildung, Zufriedenheit, Schulnoten
Intervallskala	Hier liegen quantitative, stetige Daten vor, bei denen es sinnvoll ist Abstände zu betrachten. Es existiert kein absoluter Nullpunkt (verändert sich bei linearen Transformationen).	Jahreszahlen, Temperaturskala (Celsius)
Verhältnisskala	Daten dieser Skala sind ebenfalls quantitativ und stetig. Es existiert ein absoluter Nullpunkt.	Gewicht, Größe, Alter, Einkommen

Tabelle 1.1.: Skalenniveaus

So ist das Alter zwar eine metrische Größe, wird es jedoch in Bereichen beispielsweise „jünger als 20 Jahre“, „zwischen 20 und 40 Jahren“ und „älter als 40 Jahre“ angegeben, so wäre dieses Merkmal mit seinen drei Ausprägungen als ordinal anzusehen. Hier wird also eine einst stetige Variable durch eine Gruppierung in Intervalle kategorisiert.

Weist man ordinalen Daten Zahlen zu wie zum Beispiel bei Schulnoten, so lassen sich einige Größen wie der Mittelwert oder die Standardabweichung berechnen. Diese sind auf Grund der meist beliebigen Zahlenzuweisung nur beschränkt aussagekräftig. Sinnvolle Maßzahlen sind Quantile, insbesondere der Median. Die Betrachtung sogenannter Odds Ratios ist meist auch sehr aufschlussreich. Tabelle 1.1 zeigt eine Übersicht über verschiedene Datentypen. Die Nominal- und die Ordinalskala gehören zur topologischen Skala. Man spricht von kategorischen Daten. Die Intervall- und Verhältnisskala werden zur Kardinalskala zusammengefasst.

In dieser Arbeit wollen wir verschiedene Klassen von statistischen Regressionsmodel-

len für ordinal skalierte Daten vergleichen. Wir werden deren Umsetzung in R diskutieren. Insbesondere werden die kumulativen und die sequentiellen Ansätze betrachtet. Es sind aber auch log-lineare Modelle mit fest vorgegebenen Scores für derartige Situationen verwendbar. Wir werden einen Vergleich der Modellklassen anhand realer Daten durchführen.

Im Kapitel 2 werden wir zunächst einige Definitionen und die in dieser Arbeit verwendete Notation einführen, welche wir im Zusammenhang zur Analyse ordinaler Daten benötigen. Es geht unter anderem um Odds und Odds Ratios, deren Konfidenzintervalle, die logistische Funktion usw.

Anschließend wollen wir uns auf verschiedene Regressionsmodelle konzentrieren, mit denen man ordinale Daten analysieren und interpretieren kann. Die Modelle sind problemspezifisch einzusetzen. Alle beziehen die ordinale Struktur der Daten mit ein. Sie stellen einen Spezialfall der generalisierten linearen Modelle dar, welche im Anhang A.2 erläutert sind.

Im Kapitel 3 betrachten wir kumulative Modelle, welche wir über einen sogenannten Schwellenwertansatz und einer latenten Variable motivieren. Wie der Name sagt, werden in diesen Modellen zumeist kumulierte (aufaddierte) Wahrscheinlichkeiten betrachtet, anstelle von Punktwahrscheinlichkeiten. Geht es beispielsweise um den Geschmack einer Käsesorte, welchen man mit „sehr köstlich“, „angenehmer Geschmack“, „nicht mein Geschmack“ und „ungenießbar“ bewerten kann, so interessiert eher die Wahrscheinlichkeit mit „angenehmer Geschmack“ oder besser geantwortet zu haben, anstatt die Wahrscheinlichkeit, dass jemand mit „angenehmer Geschmack“ geantwortet hat. Außerdem spielen die Odds und Odds Ratios, welche wir im Kapitel 2 eingeführt haben, eine wichtige Rolle. Weiters werden einige der zur Analyse zur Verfügung stehenden R-Funktionen beschrieben und anhand von Beispielen miteinander verglichen. Schließlich wenden wir die Modellschätzung im Abschnitt 3.10 auf eine aktuelle Studie an.

Anschließend wollen wir uns den sequentiellen Modellen zuwenden und diese im Kapitel 4 erläutern. Auch hier wird der Schwellenwertansatz zur Motivation herangezogen. Schließlich sollen im Kapitel 5 Modelle mit fixen Scores genauer erläutert werden.

Die meisten Analysen ordinaler Daten lassen sich auf sinnvolle Art und Weise mit einem Modell aus diesen Klassen durchführen. Dennoch wollen wir weitere Ansätze im Kapitel 6 betrachten. Somit erhalten wir eine Übersicht über diese Modelle und ge-

Kumulative Modelle	Einfaches kumulatives Modell, Kumulatives Logit-Modell, Verallgemeinertes kumulatives Modell
Sequentielle Modelle	Allgemeines Modell, Sequentielles Logit-Modell
Log-lineare Poissonmodelle mit fixen Scores	Einfache log-lineare Modelle, Linear-by-linear Association-Modell, Row-Effect-Association-Modell
Weitere Modelle	Kumulative Modelle mit Skalierungsparameter, Hierarchisch strukturierte Modelle, Adjazent-Kategorie-Modelle, Stereotypische Modelle

Tabelle 1.2.: Übersicht über Regressionsmodelle für ordinale Daten

langen zu einem interessanten Vergleich der die Eigenschaften und Nützlichkeit dieser miteinbezieht. Tabelle 1.2 schafft einen Überblick der Modelle, welche in dieser Arbeit besprochen werden.

1.2. Motivation zur Nutzung ordinaler Methoden

Da ordinale Daten die bereits erwähnte Ordnungsstruktur besitzen, ist es sinnvoll bei der Analyse solcher Daten auf diese einzugehen und sie miteinzubeziehen. Beachtet man diese nicht, gehen wertvolle Informationen verloren. So ist das beispielsweise der Fall, wenn bei statistischen Methoden für kategoriale Daten angenommen wird, dass die Daten nominal wären wie zum Beispiel beim Pearson Chi-Quadrat-Test auf Unabhängigkeit. Wird die Ordnungsstruktur beachtet, so gestaltet sich die Interpretation viel leichter und Trends sind eher erkennbar. Die Ergebnisse sind viel klarer und stärker, als bei Analysen, die die ordinale Struktur missachten.

2. Definitionen und Notation

In dieser Arbeit werden Vektoren und Matrizen durch fett gedruckte Buchstaben gekennzeichnet. Weiters sind Schätzer von Parametern durch ein Dach („ ̂ “) markiert.

2.1. Odds und Odds Ratios

Wie bereits einleitend beschrieben ist es bei ordinalen Daten nicht möglich Vergleiche wie „doppelt so gut“ oder ähnliches durchzuführen. Daher muss man sich hier anderer Hilfsmittel bedienen. Odds und Odds Ratios sind zwei Größen, die sich besonders gut zum Vergleich ordinaler Daten eignen. Sie werden im Folgenden definiert.

Definition 2.1.1 (Odds)**.** *Odds berechnen sich aus dem Quotienten der Wahrscheinlichkeit und der Gegenwahrscheinlichkeit eines Ereignisses. Hat man ein Ereignis A mit dessen Wahrscheinlichkeit $P(A)$, so gilt*

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)}.$$

Bemerkung 1. *Ist der Wert der Odds größer 1, so spricht das dafür, dass das Ereignis A eher eintritt. Ist es kleiner 1, tritt A eher nicht ein. Auf Grund der Eigenschaften von Wahrscheinlichkeiten, sind Odds stets größer oder gleich 0.*

Definition 2.1.2 (Odds Ratio)**.** *Hat man zwei Ereignisse A und B mit deren Wahrscheinlichkeiten $P(A)$ und $P(B)$, so ist das Odds Ratio (OR) definiert als*

$$\begin{aligned}\text{OR}(A, B) &= \frac{\text{odds}(A)}{\text{odds}(B)} \\ &= \frac{P(A)/(1 - P(A))}{P(B)/(1 - P(B))} \\ &= \frac{P(A) \cdot (1 - P(B))}{P(B) \cdot (1 - P(A))}.\end{aligned}$$

	Ereignis	Gegenereignis	Gesamtanzahl
Situation 1	n_{11} π_{11}	n_{12} π_{12}	$n_1 = n_{11} + n_{12}$
Situation 2	n_{21} π_{21}	n_{22} π_{22}	$n_2 = n_{21} + n_{22}$
Gesamtanzahl	$n_{11} + n_{21}$	$n_{12} + n_{22}$	$n = n_1 + n_2$

Tabelle 2.1.: Tabellarische Darstellung der beobachteten Anzahlen n_{ij} und Wahrscheinlichkeiten π_{ij} zweier Ereignisse

Bemerkung 2. *Odds Ratios können alle nichtnegativen Werte annehmen. Ist der Wert des Odds Ratios größer 1, so ist die Chance, dass eher A eintritt größer als die Chance, dass eher B eintritt. Ist der Wert kleiner 1 so gilt das Umgekehrte und bei einem Odds Ratio von 1 sind die Odds bei beiden Ereignissen gleich. Hat man eine Variable x und möchte man deren Einfluss auf die Ereignisse A und B untersuchen, so würde uns ein Odds Ratio nahe 1 signalisieren, dass das Merkmal x eher nicht erklärend ist.*

Hat man eine 2×2-Kontingenztafel mit beobachteten Anzahlen n_{ij} und zu schätzenden Zellwahrscheinlichkeiten $\pi_{ij}, i, j \in \{1, 2\}$ wie in Tabelle 2.1 dargestellt, so ergibt sich das Odds Ratio wie folgt

$$\begin{aligned} \mathrm{OR} &= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \\ &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \end{aligned}$$

Da man die Zellwahrscheinlichkeiten oft nicht kennt, wollen wir diese schätzen. Für $i, j = 1, 2$ ist der Maximum-Likelihood-Schätzer

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_i} \quad \text{mit}$$

$$n_i = \sum_{j=1}^{2} n_{ij}$$

und wir erhalten die Stichproben-Odds für die i-te Situation

$$\begin{aligned} \widehat{\mathrm{odds}}_i &= \frac{n_{i1}/n_i}{n_{i2}/n_i} \\ &= \frac{n_{i1}}{n_{i2}} \quad i = 1, 2. \end{aligned}$$

	Anzahl sonniger Tage	Anzahl nicht sonniger Tage
Februar	8	20
Juli	20	11
August	24	7

Tabelle 2.2.: Anzahl sonniger Tage

Das Stichproben-Odds Ratio ergibt sich aus

$$\widehat{\mathrm{OR}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Bemerkung 3. *Hat man größere Tabellen, so können die Odds und Odds Ratios auch von 2×2-Untertabellen gebildet werden. Die Herleitung der Schätzer $\hat{\pi}_{ij}$ ist im Anhang A.1 zu finden.*

Beispiel 1. *Wir betrachten die Anzahl sonniger Tage in den Monaten Februar, Juli und August eines fiktiven Jahres, welche in Tabelle 2.2 dargestellt sind. Sei S das Ereignis, dass ein Tag sonnig ist, so gilt für die Stichproben-Odds in den einzelnen Monaten*

$$\widehat{\mathrm{odds}}(S_F) = \frac{8/28}{1 - 8/28} = \frac{8}{20} = 0.4,$$

$$\widehat{\mathrm{odds}}(S_J) = \frac{20/31}{1 - 20/31} = \frac{20}{11} = 1.8,$$

$$\widehat{\mathrm{odds}}(S_A) = \frac{24/31}{1 - 24/31} = \frac{24}{7} = 3.4.$$

Man kann diesen Werten entnehmen, dass die Chance eher sonnige statt nicht sonnige Tage zu haben im Februar bei 0.4, im Juli bei 1.8 und im August bei 3.4 liegt. Das heißt im Februar haben wir eher selten sonniges Wetter. Im Juli und August haben wir eher sonnige Tage anstatt Tage, an denen die Sonne nicht scheint. Um die Chancen zu vergleichen betrachten wir die dazugehörigen Odds Ratios. Es ergibt sich

$$\widehat{\mathrm{OR}}(S_F, S_J) = \frac{0.4}{1.8} = 0.2,$$

$$\widehat{\mathrm{OR}}(S_F, S_A) = \frac{0.4}{3.4} = 0.1,$$

$$\widehat{\mathrm{OR}}(S_A, S_J) = \frac{3.4}{1.8} = 1.9.$$

Die Chance eher im Februar sonnige Tage zu haben statt nicht sonnige ist nur 0.2 mal so groß wie im Juli und nur 0.1 mal so groß wie im August. Und die Chance eher im August einen sonnigen Tag zu haben als einen nicht sonnigen ist fast doppelt so groß wie im Juli.

Mit den Odds und Odds Ratios haben wir zwei Größen, die im Bereich der nichtnegativen Zahlen liegen. Insbesondere testen wir auf Werte nahe bei 1 um den Einfluss eines Merkmals x zu untersuchen. Es wäre aber schöner, wenn wir beim Wert 0 sagen könnten, dass x nicht einflussreich ist. Dies erreichen wir, indem wir den Logarithmus der Odds bzw. des Odds Ratios bilden. Zudem erhalten wir damit eine Größe, welche sich über die gesamten reellen Zahlen erstreckt. Für eine einfache Darstellung bedienen wir uns der Logit-Funktion.

Definition 2.1.3 (Logit-Funktion und logistische Funktion)**.** *Sei $p \in (0,1)$, so ist der Logit von p (siehe Abbildung 2.1) definiert als*

$$\text{logit}(p) = \log \frac{p}{1-p}.$$

Falls p eine Wahrscheinlichkeit ist, so kann man sagen, dass der Logit von p dem Logarithmus der Odds von p entspricht. Man nennt diese dann auch Log-Odds. Die Umkehrfunktion ist die logistische Funktion, welche sich wie folgt definiert

$$\text{logit}^{-1}(p) = \frac{1}{1+e^{-p}} = \frac{e^p}{1+e^p}.$$

Dies entspricht der logistischen Verteilungsfunktion.

$$F_X(x) = P(X \leq x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}.$$

Die Abbildung 2.2 zeigt diese gemeinsam mit der zugehörigen Dichtefunktion, welche sich ergibt zu

$$f_X(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^x}{(1+e^x)^2}.$$

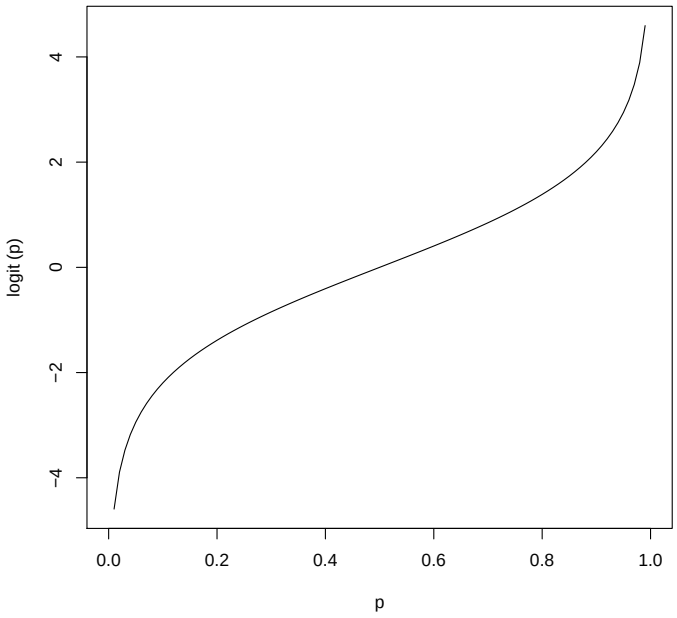

Abbildung 2.1.: Logitfunktion

Damit können wir den Logarithmus des Odds Ratios mit Hilfe der Logit-Funktion darstellen

$$\begin{aligned}\log(\mathrm{OR}(A,B)) &= \log \frac{P(A)/(1-P(A))}{P(B)/(1-P(B))} \\ &= \log \frac{P(A)}{1-P(A)} - \log \frac{P(B)}{1-P(B)} \\ &= \mathrm{logit}(P(A)) - \mathrm{logit}(P(B)).\end{aligned}$$

Dies wird als Log-Odds Ratio oder die Logit-Differenz bezeichnet. Ein Odds Ratio kleiner 1 entspricht somit einem negativen Log-Odds Ratio, ein Odds Ratio größer 1 einem positiven.

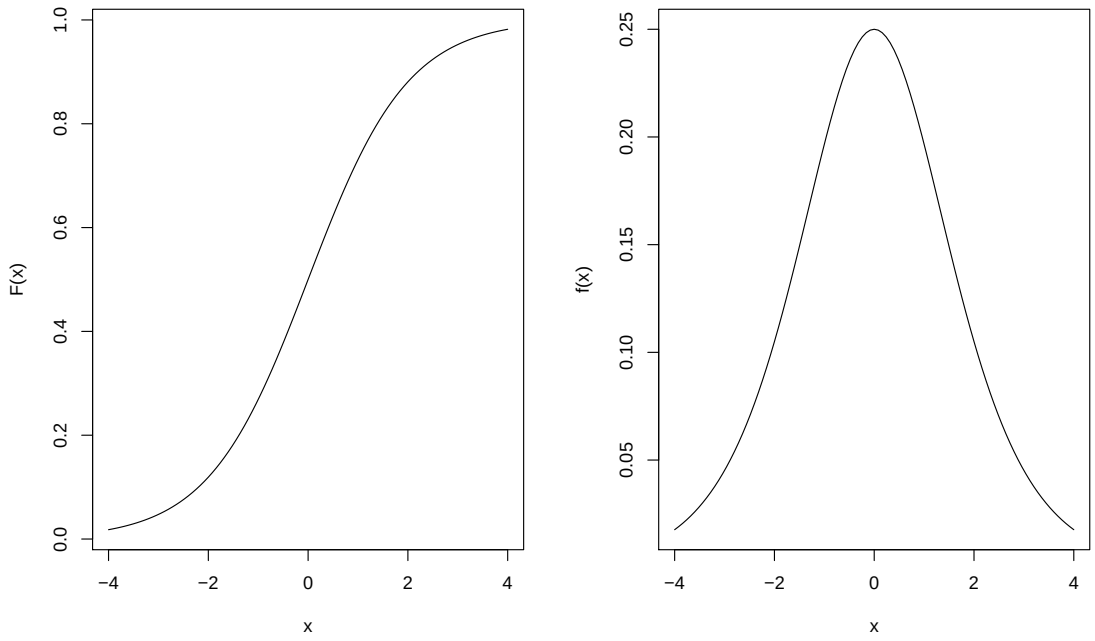

Abbildung 2.2.: Verteilungs- und Dichtefunktion der logistischen Verteilung

2.2. Konfidenzintervalle

Als nächstes wollen wir Konfidenzintervalle der Odds Ratios angeben (vergleiche Agresti, 2002). Dazu ermitteln wir die Konfidenzintervalle der Log-Odds Ratios und transformieren diese schließlich, indem wir sie exponieren. Für die Berechnung benötigen wir den beobachteten Wert und den Standardfehler SE der Log-Odds Ratios.

Ein 95%-Konfidenzintervall für einen Parameter β entspricht einem Hypothesentest $H_0 : \beta = \beta_0$ mit einem p-Wert von 0.05. Das meist genutzte Konfidenzintervall ist das Wald-Konfidenzintervall, da es leicht zu konstruieren ist. Man benötigt allein den Schätzer $\hat{\beta}$ und den Standardfehler $SE(\hat{\beta})$, welcher von 0 verschieden sein sollte. Unter H_0 kann man für die Teststatistik $(\hat{\beta} - \beta_0)/SE(\hat{\beta})$ approximativ die Normalverteilung annehmen, wenn der Schätzer $\hat{\beta}$ nahezu normalverteilt ist. Das Wald-Konfidenzintervall

betrachtet jene Menge der β_0, für die

$$\frac{1}{SE(\hat{\beta})} \mid \hat{\beta} - \beta_0 \mid < z_{1-\alpha/2}$$

gilt, wobei z_q dem $q \times 100\%$-Quantil einer Standardnormalverteilung entspricht. Dies liefert die Grenzen des Konfidenzintervalls für einen Parameter β

$$\hat{\beta} \pm z_{1-\alpha/2} \cdot SE(\hat{\beta}).$$

Ein weiterer Zugang zur Bestimmung eines Konfidenzintervalls ist der Likelihood-Quotienten-Test. Die darauf basierenden Konfidenzintervalle sind für einen kleineren Stichprobenumfang besser geeignet als das Wald-Konfidenzintervall. Der Likelihood-Quotienten-Test ist ein klassisches Testverfahren der schließenden Statistik.

Um ein Konfidenzintervall anzugeben betrachten wir zunächst eine Stichprobe, die wie in Tabelle 2.1 dargestellt werden kann und gehen davon aus, dass die Anzahlen einer Multinomialverteilung genügen. Damit wir ein Konfidenzintervall nach dem Wald-Test für das Odds Ratio angeben können, müssen wir zeigen, dass der Schätzer $\widehat{\mathrm{OR}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ bzw. $\log \widehat{\mathrm{OR}} = \log \frac{n_{11}n_{22}}{n_{12}n_{21}}$ asymptotisch normalverteilt ist. Dafür betrachten wir zunächst allgemein den Zusammenhang zwischen den Parametern in einer multinomialverteilten Stichprobe zu jenen in einer Normalverteilung.

Asymptotische Normalverteilung von Funktionen von multinomialverteilten Daten

Es wird angenommen, dass es bei einer Datenauswertung zu N verschiedenen Ausprägungen kommen kann, wobei jede Ausprägung A_j mit der Anzahl $n_j, j = 1, \ldots, N$, belegt ist (siehe Tabelle 2.3). Diese Anzahlen seien multinomialverteilt und weiters enthält $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ die Wahrscheinlichkeiten dafür, dass eine einzelne Ausprägung j eintritt für alle $j = 1, \ldots, N$. Es sei $n = n_1 + \cdots + n_N$ die Anzahl aller Beobachtungen. Der Vektor $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_N)^T$ beschreibt die relativen Häufigkeiten $\hat{\pi}_j = n_j/n$. Dies ist der Maximum-Likelihood-Schätzer für $\boldsymbol{\pi}$. Außerdem sei $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN})$ mit $Y_{ij} = 1$, falls die i-te Beobachtung Ausprägung j aufweist und sonst 0 für $i = 1, \ldots, n$. Daher folgt, dass $Y_{ij}Y_{ik} = 0$ für $j \neq k$ und $\sum_j Y_{ij} = 1$ ist, da jede Beobachtung in nur eine

A_1	A_2	$\dots$	A_N	$\sum$
n_1	n_2	$\dots$	n_N	n
π_1	π_2	$\dots$	π_N	1

Tabelle 2.3.: Tabellarische Darstellung der Anzahlen und Wahrscheinlichkeiten von N Ereignissen

Zelle fällt. Des Weiteren ergibt sich für alle $j = 1, \dots, N$

$$\hat{\pi}_j = \frac{n_j}{n} = \frac{\sum_{i=1}^{n} Y_{ij}}{n}$$

und

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_j,$$
$$E(Y_{ij}^2) = \pi_j,$$
$$E(Y_{ij}Y_{ik}) = 0, \quad j \neq k.$$

Sei nun $\boldsymbol{\Sigma}(\boldsymbol{Y}_i) = (\sigma_{jk})$ die Varianz-Kovarianz-Matrix von $\boldsymbol{Y}_i$ mit

$$\sigma_{jj} = Var(Y_{ij}) = E(Y_{ij}^2) - E(Y_{ij})^2 = \pi_j - \pi_j^2 = \pi_j(1 - \pi_j)$$
$$\sigma_{jk} = Cov(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) = -\pi_j\pi_k, \quad j \neq k,$$

dann erhält man insgesamt

$$E(\boldsymbol{Y}_i) = \boldsymbol{\pi}$$
$$\boldsymbol{\Sigma}(\boldsymbol{Y}_i) = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T,$$

wobei $diag(\boldsymbol{\pi})$ die Werte π_j in den Diagonalelementen enthält und sonst 0 ist. Der Vektor $\hat{\boldsymbol{\pi}}$ kann als Mittelwert von n unabhängigen Beobachtungen gesehen werden, denn

$$\hat{\boldsymbol{\pi}} = \frac{\sum_{i=1}^{n} \boldsymbol{Y}_i}{n},$$
$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{n}.$$

Daher folgt

$$Var(\hat{\pi}_j) = \frac{1}{n^2} \cdot n \cdot Var(Y_{ij}) = \frac{Var(Y_{ij})}{n}$$
$$Cov(\hat{\pi}_j, \hat{\pi}_k) = \frac{Cov(Y_{ij}, Y_{ik})}{n}.$$

Somit ergeben sich Erwartungswert und Varianz-Kovarianz-Matrix von $\hat{\boldsymbol{\pi}}$ zu

$$E(\hat{\boldsymbol{\pi}}) = \frac{\sum_{i=1}^{n} E(\boldsymbol{Y}_i)}{n} = \frac{n\boldsymbol{\pi}}{n} = \boldsymbol{\pi}$$
$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\pi}}) = \frac{1}{n}(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T).$$

Der multivariate zentrale Grenzwertsatz liefert uns folgendes Ergebnis

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T).$$

Wendet man hier die in A.4.1 beschriebene Delta-Methode an, so erklärt sich, dass auch Funktionen von $\hat{\boldsymbol{\pi}}$, deren Differential bei $\boldsymbol{\pi}$ verschieden von $\mathbf{0}$ ist, ebenfalls asymptotisch normalverteilt sind. Sei dazu $g(t_1, \ldots, t_N)$ eine differenzierbare Funktion und sei

$$\phi_j = \partial g(\boldsymbol{t})/\partial t_j \Big|_{\boldsymbol{t}=\boldsymbol{\pi}}, \quad j = 1, \ldots, N$$

die Ableitung $\partial g/\partial t_i$ an der Stelle $\boldsymbol{t} = \boldsymbol{\pi}$ ausgewertet. Mit der Delta-Methode für Vektoren folgt

$$\sqrt{n}(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\phi}^T(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)\boldsymbol{\phi}),$$

wobei sich die Varianz wie folgt darstellen lässt

$$\begin{aligned}\boldsymbol{\phi}^T(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)\boldsymbol{\phi} &= \boldsymbol{\phi}^T diag(\boldsymbol{\pi})\boldsymbol{\phi} - (\boldsymbol{\phi}^T\boldsymbol{\pi})^2 \\ &= \sum_{j=1}^{N} \pi_j \phi_j^2 - \left(\sum_{j=1}^{N} \pi_j \phi_j\right)^2.\end{aligned}$$

Es folgt daher, dass sich $\hat{\boldsymbol{\pi}}$ asymptotisch normalverteilt verhält. Da das Odds Ratio und das Log-Odds Ratio Funktionen von $\boldsymbol{\pi}$ bzw. deren Schätzer Funktionen vom $\hat{\boldsymbol{\pi}}$ sind, folgt auch für diese die asymptotische Normalverteilung. Das Log-Odds Ratio liefert uns eine additive Struktur. Zudem konvergiert es schneller gegen die Normalverteilung als

das Odds Ratio. Aus diesem Grund betrachten wir das Wald-Konfidenzintervall des Log-Odds-Ratios, welches sich mittels

$$\log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} SE(\log \widehat{\mathrm{OR}})$$

berechnen lässt.

Berechnung des Standardfehlers

Um den Standardfehler des Log-Odds Ratios zu bestimmen zu können, verwenden wir die Delta-Methode. Es sei $\hat{\boldsymbol{\pi}} = (n_1, \ldots, n_N)/n$. Wie zuvor gezeigt gilt

$$\begin{aligned} E(\hat{\pi}_j) &= \pi_j, \\ Var(\hat{\pi}_j) &= \frac{\pi_j - \pi_j^2}{n}, \\ Cov(\hat{\pi}_j, \hat{\pi}_k) &= -\frac{\pi_j \pi_k}{n}. \end{aligned}$$

Die Schätzer $(\hat{\pi}_1, \ldots, \hat{\pi}_{N-1})$ verhalten sich für große Stichprobenumfänge multivariat normalverteilt. Für eine differenzierbare Funktion $g(\boldsymbol{\pi})$ auf $\mathbb{R}$ folgt mittels der Delta-Methode, dass für $n \to \infty$

$$\frac{\sqrt{n}(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi}))}{\sigma} \xrightarrow{d} N(0,1),$$

wobei

$$\sigma^2 = \sum_{j=1}^{N} \pi_j \phi_j^2 - \left(\sum_{j=1}^{N} \pi_j \phi_j \right)^2.$$

Anders ausgedrückt bedeutet das, dass $g(\hat{\boldsymbol{\pi}})$ gegen eine Normalverteilung mit Erwartungswert $g(\boldsymbol{\pi})$ und Standardabweichung $\sigma/\sqrt{n}$ strebt. Da die π_j unbekannt sind, wollen wir sie durch ihre Schätzer ersetzen und erhalten somit auch Schätzer für σ. Für $g(\hat{\boldsymbol{\pi}})$ ist $\hat{\sigma}/n$ der Standardfehler. Die Grenzverteilung bleibt die Normalverteilung, wenn wir σ durch $\hat{\sigma}$ ersetzen. Dieser Zusammenhang lässt sich wie folgt erklären. Die Schätzer $\hat{\pi}_j = n_j/n$ stellen das arithmetische Mittel von n unabhängig und identisch verteilten Variablen dar und konvergieren daher nach dem Schwachen Gesetz der großen Zahlen

für wachsendes n in Wahrscheinlichkeit gegen π_j. Da $\hat{\sigma}$ bezüglich $\hat{\boldsymbol{\pi}}$ stetig ist, konvergiert dieses in Wahrscheinlichkeit gegen σ und daher $\sigma/\hat{\sigma}$ gegen 1. Mit der Umformung

$$\frac{\sqrt{n}(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi}))}{\hat{\sigma}} = \frac{\sqrt{n}(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi}))}{\sigma}\frac{\sigma}{\hat{\sigma}}$$

kann man erkennen, dass der Ausdruck auf der linken Seite nach dem Satz von Slutsky gegen eine Standardnormalverteilung konvergiert, da $\sqrt{n}(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi}))/\sigma$ in Verteilung gegen die Standardnormalverteilung und $\sigma/\hat{\sigma}$ in Wahrscheinlichkeit gegen 1 strebt. Somit erhält man die Grenzen des Wald-Konfidenzintervalls für $g(\boldsymbol{\pi})$

$$g(\hat{\boldsymbol{\pi}}) \pm z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}.$$

Schlussendlich wollen wir diese allgemeinen Herleitungen auf das Log-Odds Ratio anwenden. Hierbei ist

$$\begin{aligned} g(\boldsymbol{\pi}) &= \log(\mathrm{OR}) \\ &= \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21}. \end{aligned}$$

Wir erhalten für $\phi_{ij} = \partial(\log(\mathrm{OR})/\partial\pi_{ij}, i, j \in \{1, 2\}$

$$\begin{aligned} \phi_{11} &= \frac{1}{\pi_{11}}, \\ \phi_{12} &= -\frac{1}{\pi_{12}}, \\ \phi_{21} &= -\frac{1}{\pi_{21}}, \\ \phi_{22} &= \frac{1}{\pi_{22}} \end{aligned}$$

und somit folgt

$$\sum_{i=1}^{2}\sum_{j=1}^{2} \pi_{ij}\phi_{ij} = 0,$$

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2} \pi_{ij}\phi_{ij}^2 \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{1}{\pi_{ij}}. \end{aligned}$$

Der asymptotische Standardfehler vom geschätzten Log-Odds Ratio, $\log \widehat{\mathrm{OR}}$, für eine multinomialverteilte Stichprobe ist

$$(Var(\log \widehat{\mathrm{OR}}))^{\frac{1}{2}} = \sigma/\sqrt{n} = \left(\sum_i \sum_j \frac{1}{n\pi_{ij}} \right)^{\frac{1}{2}}.$$

Ersetzen wir π_{ij} durch seinen Schätzer und wenden an, dass $n\hat{\pi}_{ij} = n_{ij}$, so ergibt sich

$$SE(\log \widehat{\mathrm{OR}}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Nun können wir die Grenzen berechnen und das zweiseitige $(1-\alpha)$-Konfidenzintervall angeben

$$\begin{aligned} CI(\log \mathrm{OR}) &= \log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} SE(\log \widehat{\mathrm{OR}}) \\ &= \log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \end{aligned}$$

Das entsprechende Konfidenzintervall für die Odds Ratios erhalten wir, indem wir beide Grenzen exponieren

$$CI(\mathrm{OR}) = \exp\left(\log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} SE(\log \widehat{\mathrm{OR}}) \right). \tag{2.1}$$

Durch die Berechnung des Konfidenzintervalls wissen wir, in welchem Bereich der wahre Wert des Odds Ratios mit $(1-\alpha)\cdot 100\%$-iger Sicherheit liegt. Die entscheidende Frage dabei ist, ob das Konfidenzintervall (2.1) die 1 überdeckt oder eben nicht. Überdeckt es die 1, so können wir nicht ausschließen, dass die Chancen für die zwei betrachtete Ereignisse gleich sind.

Beispiel 2. *Betrachten wir noch einmal die Monate Juli und August aus Beispiel 1 und berechnen das Konfidenzintervall des dazugehörigen Odds Ratios. Wir erhalten für den geschätzten Wert des Log-Odds Ratios*

$$\log(\widehat{\mathrm{OR}}(A,J)) = \log \frac{24 \cdot 11}{7 \cdot 20} = 0.63$$

und als Standardfehler ergibt sich

$$SE(\log \widehat{\mathrm{OR}}(A,J)) = \sqrt{\frac{1}{24} + \frac{1}{7} + \frac{1}{20} + \frac{1}{11}} = 0.57.$$

Wählen wir $\alpha = 0.1$ *so benötigen wir für das 90%-Konfidenzintervall das Quantil* $z_{0.95} = 1.65$ *und somit ist*

$$\begin{aligned} CI(\log \mathrm{OR}(A, J)) &= [0.63 - 1.65 \cdot 0.57, 0.63 + 1.65 \cdot 0.57] \\ &= [-0.3040, 1.5726] \,. \end{aligned}$$

Dieses überdeckt die 0. Durch Exponieren sehen wir schließlich, dass das Konfidenzintervall für das Odds Ratio äquivalent die 1 überdeckt

$$\begin{aligned} CI(\mathrm{OR}(A, J)) &= [\exp(-0.31), \exp(1.57)] \\ &= [0.7378, 4.8194] \,. \end{aligned}$$

Es kann daher nicht ausgeschlossen werden, dass die Chancen eher sonnige Tage zu haben in den Monaten Juli und August gleich sind. Vergleichen wir hingegen die Monate Februar und Juli, so lässt das Konfidenzintervall wie folgt berechnen

$$\begin{aligned} CI(\mathrm{OR}(F, J)) &= \exp(\log(8 * 11/(20 * 20)) \pm z_{0.95} * (1/20 + 1/8 + 1/20 + 1/11)^{1/2}) \\ &= [0.0873, 0.5545] \end{aligned}$$

mit 90%-iger Sicherheit ausschließen, dass in diesen Monaten die Chancen eher sonnige Tage zu haben identisch sind.

3. Kumulative Modelle

In diesem und in den folgenden Kapiteln werden wir uns der Regressionsanalyse ordinaler Daten widmen. Zunächst wollen wir ordinale Daten mit Hilfe einer ordinalen Responsevariable Y, die c verschiedene Werte annehmen kann, beschreiben. Das Verhalten dieser ist dabei abhängig vom Spaltenvektor $\boldsymbol{x}$, welcher alle erklärenden Variablen (Merkmale, Faktoren) enthält. Weiters sei $\boldsymbol{\beta}$ der Vektor der Parameter. Er beschreibt den Einfluss der erklärenden Variablen auf unsere Responsevariable. Beispielsweise könnte x_{i1} das Alter und x_{i2} das Geschlecht der i-ten Person sein und wir sind interessiert daran, wie oft die Person in ihrer Freizeit wandern geht. Die Responsevariable Y besäße in dem Fall die Ausprägungen „sehr oft“, „oft“, „selten“und „gar nicht“. Ziel ist es nun Werte des Parameters $\boldsymbol{\beta}$ zu erhalten, mit denen wir den Zusammenhang zwischen $\boldsymbol{x}$ und Y möglichst gut beschreiben können. Dieser Zusammenhang ist jedoch nicht direkt linear, da kumulative Modelle zu den generalisierten linearen Modellen gehören. Im Anhang A.2 sind zu generalisierten linearen Modellen allgemeine Erläuterungen zu finden. Im nächsten Abschnitt leiten wir uns einen geeigneten Zusammenhang her, mittels dem wir das Problem anhand des Schwellenwertansatzes motivieren wollen.

3.1. Das einfache kumulative Modell und der Schwellenwertansatz

Dieser Ansatz baut wie in Tutz (2012) beschrieben auf der Annahme auf, dass man ein Regressionsmodell einer stetigen Zufallsvariable Z betrachtet, wobei der Bereich, den Z annehmen kann, in c Kategorien geteilt wird. Z ist hier also eine latente Variable. Weiters gibt es Schwellenwerte a_j mit der Eigenschaft

$$-\infty = a_0 < a_1 < \cdots < a_c = \infty.$$

Mittels der ordinalen Responsevariable Y beschreiben wir in welchen Bereich Z realisiert

$$a_{j-1} \leq Z \leq a_j \Leftrightarrow Y = j, \quad j = 1, \ldots, c.$$

Für Z wird die Darstellung

$$Z = \alpha^* + \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon$$

angenommen, wobei α^* der Intercept und ϵ eine Zufallsvariable mit Verteilungsfunktion F ist. Mit $\alpha_j := a_j - \alpha^*$ ergibt sich folgender Zusammenhang

$$\begin{aligned} P(Y \leq j \mid \boldsymbol{x}) &= P(Z \leq a_j \mid \boldsymbol{x}) \\ &= P(\alpha^* + \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon \leq a_j) \\ &= P(\epsilon \leq a_j - \alpha^* - \boldsymbol{\beta}^T \boldsymbol{x}) \\ &= F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}). \end{aligned}$$

Beispiel 3. *Nehmen wir an Z beschreibt die Zufriedenheit einer Person mit einem gekauften Produkt (beispielsweise ein neues Fahrrad) auf einer Skala von 1 bis 10. Verschiedene Personen würden nun für sich die Zahlen mit unterschiedlichen Interpretationen belegen. Für den einen entspräche nur die 1 einem „sehr zufrieden“, für einen anderen würden die Zahlen 2 und 3 auch noch in diesen Bereich fallen. Die sich ergebenen Zahlenwerte sind nicht eindeutig interpretierbar. Daher ist es in solchen Fällen sinnvoll, die Skala in verschiedene Bereiche 1-3, 4-7, 8-10 zu teilen und diese Bereiche zum Beispiel mit „sehr zufrieden“, „zufrieden“ und „unzufrieden“ zu deklarieren. Dies ist eine ordinale Skalierung und diese wird durch die Responsevariable Y beschrieben.*

In Abbildung 3.1 ist die latente Variable Z im Zusammenhang zur ordinalen Variable Y für verschiedene $\boldsymbol{x}$ dargestellt. Sei $f(\cdot)$ die zugehörige Dichtefunktion zu F. Die Wahrscheinlichkeit $P(Y = j \mid \boldsymbol{x})$ ergibt sich durch Integration der um $-\alpha^* - \boldsymbol{\beta}^T \boldsymbol{x}$ verschobenen Dichte f zwischen a_{j-1} und a_j. Folgende Rechnung lässt dies genauer

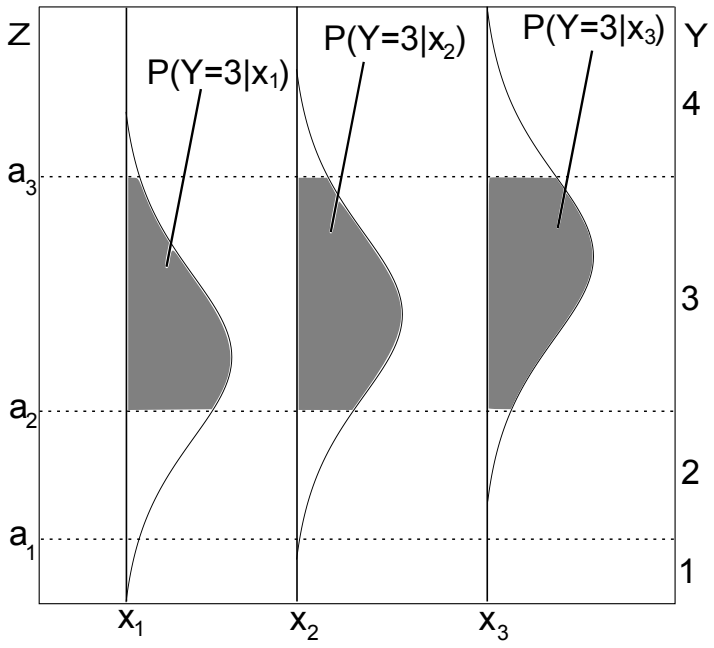

Abbildung 3.1.: Zusammenhang zwischen Z und Y

nachvollziehen

$$\begin{aligned}
P(Y = j \mid \boldsymbol{x}) &= P(Z \leq a_j \mid \boldsymbol{x}) - P(Z \leq a_{j-1} \mid \boldsymbol{x}) \\
&= P(\alpha^* + \boldsymbol{\beta}^T\boldsymbol{x} + \epsilon \leq a_j) - P(\alpha^* + \boldsymbol{\beta}^T\boldsymbol{x} + \epsilon \leq a_{j-1}) \\
&= P(\epsilon \leq a_j - \alpha^* - \boldsymbol{\beta}^T\boldsymbol{x}) - P(\epsilon \leq a_{j-1} - \alpha^* - \boldsymbol{\beta}^T\boldsymbol{x}) \\
&= F(a_j - \alpha^* - \boldsymbol{\beta}^T\boldsymbol{x}) - F(a_{j-1} - \alpha^* - \boldsymbol{\beta}^T\boldsymbol{x}) \\
&= \int_{a_{j-1}-\alpha^*-\boldsymbol{\beta}^T\boldsymbol{x}}^{a_j-\alpha^*-\boldsymbol{\beta}^T\boldsymbol{x}} f(t)\mathrm{d}t \\
&= \int_{a_{j-1}}^{a_j} f(t - \alpha^* - \boldsymbol{\beta}^T\boldsymbol{x})\mathrm{d}t.
\end{aligned}$$

Besteht der Parametervektor $\boldsymbol{\beta}$ aus negativen Komponenten, so verschiebt sich die Dichte mit wachsendem $\boldsymbol{x}$ nach links (bzw. oben).

Einfaches kumulatives Modell

Im kumulativen Modell gilt

$$P(Y \leq j \mid \boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}), \tag{3.1}$$

wobei

$$-\infty = \alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_c = \infty.$$

In Tutz (2012) findet man dafür auch die Bezeichnung „Schwellenwertmodell (Threshold Modell)“. Außerdem erkennt er, dass man dieses Modell auch als binäres Regressionsmodell betrachten kann. Man teilt den Wertebereich der Responsevariable in zwei Bereiche $\{1, \ldots, j\}$ und $\{j+1, \ldots, c\}$ und definiert

$$W_j = 1 \Leftrightarrow Y \in \{1, \ldots, j\},$$

$$W_j = 0 \Leftrightarrow Y \in \{j+1, \ldots, c\}.$$

Somit erreicht man, dass

$$P(Y \leq j \mid \boldsymbol{x}) = P(W_j = 1 \mid \boldsymbol{x}).$$

Folglich ist das obige Modell ein binäres Regressionsmodell für $W_j \in \{0, 1\}$. Es gibt auch weitere Ansätze um ordinale Response-Modelle aus binären Response-Modellen zu konstruieren. Diese unterscheiden sich in der Transformation der Kategorien zum obigem Ansatz.

Nimmt man an, dass F die logistische Verteilungsfunktion ist, die wir in Definition 2.1.3 spezifiziert haben, so entspricht dies dem **kumulativen Logit-Modell**, welches wir im nächsten Abschnitt besprechen wollen, und es gilt

$$P(Y \leq j \mid \boldsymbol{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})} \quad \text{bzw.}$$

$$\operatorname{logit}(P(Y \leq j \mid \boldsymbol{x})) = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}. \tag{3.2}$$

Dabei ist der lineare Prädiktor $\boldsymbol{\beta}^T \boldsymbol{x}$ eine lineare Funktion der erklärenden Variablen.

Es ist auch möglich andere Verteilungen für F zu verwenden. Die Gompertz-, Gumbel- oder Normalverteilung werden wir als alternative Linkfunktionen im Abschnitt 3.9 betrachten. Die logistische Verteilung ist dennoch die meist genutzte, da sich mittels dieser ein Zusammenhang zu den Odds Ratios bzw. Log-Odds Ratios ergibt und somit die Interpretationen der Parameter leicht möglich sind.

3.2. Das kumulative Logit-Modell

In diesem Abschnitt wollen wir das meist verwendete kumulative Modell besprechen - das kumulative Logit-Modell. Dieses besitzt nützliche Eigenschaften, durch welche sich einige Vorteile gegenüber anderen Modellen ergeben. Insbesondere die Interpretation der Parameter stellt sich als relativ einfach heraus. Man findet es auch unter dem Namen „Proportional Odds Model“ beispielsweise bei Tutz (2012).

Definition 3.2.1 (Kumulativer Logit)**.** *Sei Y eine ordinale Responsevariable, die c verschiedene Kategorien mit Wahrscheinlichkeiten $\pi_1, \ldots, \pi_c$ annehmen kann. Dann sind die kumulativen Logits für $j = 1, \ldots, c-1$ über*

$$\begin{aligned}\operatorname{logit}\left(P\left(Y \leq j\right)\right) &= \log\left(\frac{P\left(Y \leq j\right)}{P\left(Y > j\right)}\right) \\ &= \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right)\end{aligned}$$

definiert.

Das kumulative Logit Modell bezieht die erklärenden Variablen $\boldsymbol{x}$ mit ein und beschreibt die obige Wahrscheinlichkeit wie folgt

$$\begin{aligned}\operatorname{logit}\left(P\left(Y_i \leq j|\boldsymbol{x}_i\right)\right) &= \alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_i \\ &= \alpha_j - \beta_1 x_{i1} - \cdots - \beta_k x_{ir}, \quad j = 1, \ldots, c-1.\end{aligned}$$

Mit der Umkehrfunktion gelangt man zum äquivalenten Modell

$$P\left(Y_i \leq j|\boldsymbol{x}_i\right) = \frac{\exp\left(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_i\right)}{1 + \exp\left(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_i\right)}$$

für $j = 1, \ldots, c-1$. Man bemerke, dass dieser Ausdruck für $j = c$ gerade 1 ergeben muss (dies ist ein sogenanntes sicheres Ereignis). Ist man unter diesem Modell an der Wahrscheinlichkeit interessiert, dass die Responsevariable genau den Wert j annimmt,

so können die Punktwahrscheinlichkeiten berechnet werden durch

$$P\left(Y=j|\boldsymbol{x}\right)=\begin{cases}\dfrac{\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)} & j=1,\\[2ex] \dfrac{\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)}-\dfrac{\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)} & 1<j<c,\\[2ex] 1-\dfrac{\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)} & j=c.\end{cases}$$

Mit der Festlegung, dass $\alpha_0=-\infty$ und $\alpha_c=\infty$ lässt sich der Ausdruck zusammenfassen zu

$$P\left(Y=j|\boldsymbol{x}\right)=\frac{\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}\right)}-\frac{\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)}{1+\exp\left(\alpha_{j-1}-\boldsymbol{\beta}^T\boldsymbol{x}\right)},\quad j=1,\ldots,c.$$

3.2.1. Folgerungen und Eigenschaften

Monoton wachsende Intercepts

Im einfachen kumulativen Modell geht man davon aus, dass es nur einen Parametervektor $\boldsymbol{\beta}$ für alle c Kategorien gibt. Kategorieabhängig sind allein die Intercepts α_j. Damit keine negativen Wahrscheinlichkeiten entstehen, müssen diese monoton wachsend sein, das heißt

$$\alpha_1\leq\alpha_2\leq\cdots\leq\alpha_c.$$

Im Abschnitt 3.1 sind die Intercepts über $\alpha_j=a_j-\alpha^*$ definiert, wobei α^* der Intercept des Regressionsmodells des Fehlerterms ϵ ist. Die a_j sind die Schwellenwerte. Da diese monoton wachsend sind, ergibt sich die monotone Steigung der α_j für wachsendes j. Diese Eigenschaft gilt für alle einfachen kumulativen Modelle. Abbildung 3.2 zeigt die kumulativen Wahrscheinlichkeiten für verschiedene α_j und konstantem β in Abhängigkeit von x. Nimmt man von diesen die Logitfunktion, so gelangt man zur unteren Abbildung. Alle Kurven haben dieselbe Form. In

$$P(Y\leq j\mid\boldsymbol{x})=\frac{\exp(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x})}{1+\exp(\alpha_j-\boldsymbol{\beta}^T\boldsymbol{x})},\quad j=1,\ldots,c-1$$

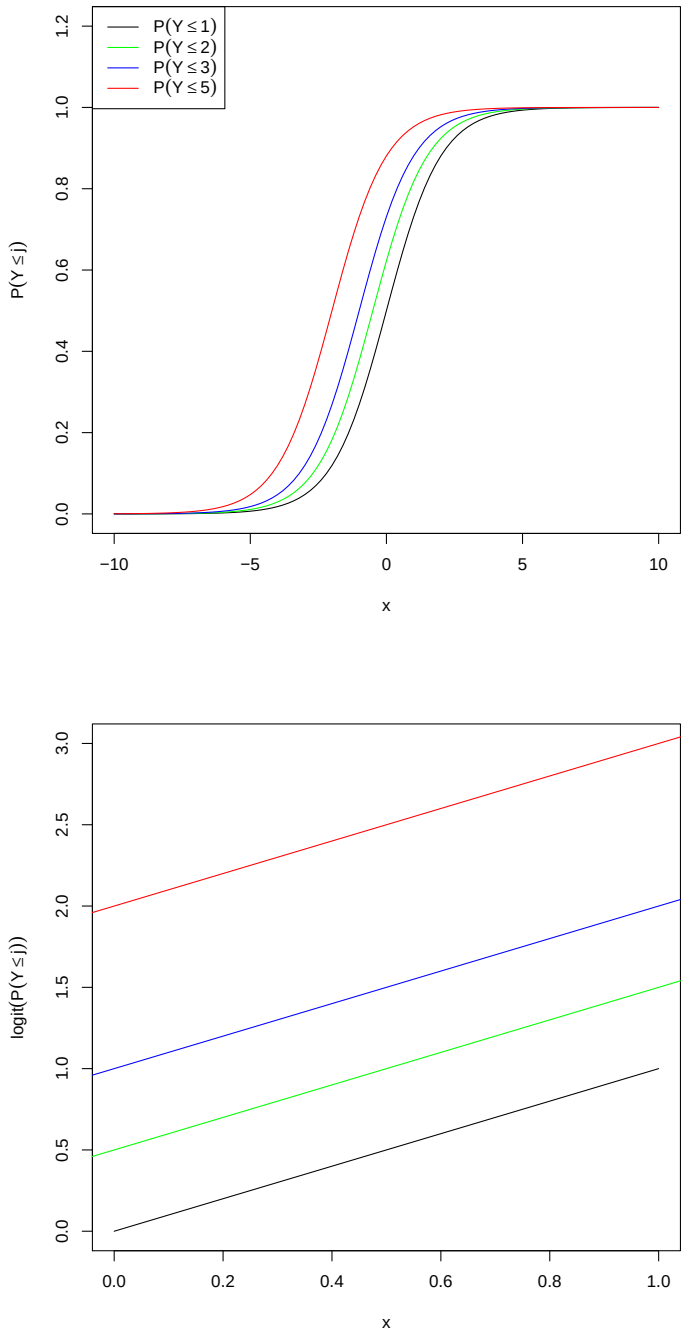

Abbildung 3.2.: Kumulative Wahrscheinlichkeiten $P(Y \leq j)$ für verschiedene α_j und konstantem $\boldsymbol{\beta}$ (oben) und deren Logit-Transformierten (unten).

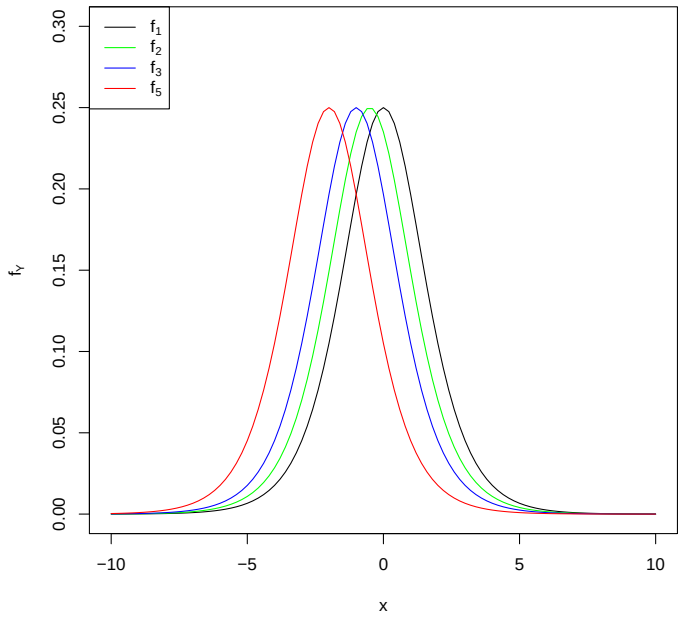

Abbildung 3.3.: Dichtefunktionen f_Y zur logistischen Verteilung für verschiedene α_j.

wird α_j mit wachsendem j größer und die zugehörigen Plots verschieben sich nach links. Dies bestätigt die Tatsache, dass für ein fixes $\boldsymbol{x}$ die Wahrscheinlichkeit $P(Y \leq j \mid \boldsymbol{x})$ mit größer werdendem j steigt. Weiters ist die Linearität für

$$\text{logit}(P(Y \leq j \mid \boldsymbol{x})) = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}$$

in Abbildung 3.2 zu erkennen. Für verschiedene α_j sind die Geraden parallel. In der Abbildung 3.3 sehen wir die Dichtefunktion der logistischen Verteilung für verschiedene α.

Strikte stochastische Ordnung

Vergleicht man zwei Responses von unterschiedlichen Objekten mit erklärenden Variablen $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ miteinander und betrachtet unter dem kumulativen Logit-Modell das

kumulative Log-Odds Ratio

$$\begin{aligned}\text{logit}\left(P\left(Y \leq j|\boldsymbol{x}_1\right)\right) - \text{logit}\left(P\left(Y \leq j|\boldsymbol{x}_2\right)\right) &= \left(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_1\right) - \left(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_2\right) \\ &= -\boldsymbol{\beta}^T\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right),\end{aligned}$$

so kann man erkennen, dass dieses unabhängig von den Parametern α_j ist. Somit ist das kumulative Log-Odds Ratio proportional zur Distanz zwischen $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ für alle $j = 1, \ldots, c-1$. Aus diesem Grund wurde dieses Model in McCullagh (1980) das „Proportional Odds Model“ genannt. Möchte man dieses Modell für die Analyse von Daten verwenden, so sollte man zunächst untersuchen, ob die Odds wirklich proportional sind. Andererseits kann es zu verfälschten Ergebnissen kommen. Weitere Erläuterungen über die Annahme proportionaler Odds sind in Abschnitt 3.4 zu finden. Die kumulativen Odds Ratios sind auch unabhängig von α, denn es gilt

$$\begin{aligned}\frac{P\left(Y \leq j|\boldsymbol{x}_1\right)/P\left(Y > j|\boldsymbol{x}_1\right)}{P\left(Y \leq j|\boldsymbol{x}_2\right)/P\left(Y > j|\boldsymbol{x}_2\right)} &= \frac{\exp(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_1)}{\exp(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}_2)} \\ &= \exp\left(-\boldsymbol{\beta}^T\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right).\end{aligned}$$

Die Chance $Y \leq j$ bei $\boldsymbol{x} = \boldsymbol{x}_1$ zu erhalten, ist das $\exp\left(-\boldsymbol{\beta}^T\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right)$-fache der Chance bei $\boldsymbol{x} = \boldsymbol{x}_2$ für alle $j = 1, \ldots, c-1$. Durch die Unabhängigkeit vom Intercept sind die Odds Ratios für alle Kategorien gleich und nur von $\boldsymbol{x}$ abhängig. Dies macht die Interpretation relativ einfach. Diese Eigenschaft der strikten stochastischen Ordnung hält auch für alle anderen einfachen kumulativen Modelle, der Zusammenhang zu den Odds bzw. Odds Ratios ist jedoch nicht herstellbar. Bei der Interpretation der Parameter, d.h. des Prädiktors $\boldsymbol{\beta}^T\boldsymbol{x}$, muss beachtet werden, dass ein größerer Wert des Prädiktors bedeutet, dass die Wahrscheinlichkeit einer höheren Stufe steigt.

Flexibilität (Collapsibility)

Gruppiert man Daten, in dem man einzelne Ausprägungen zusammenfasst und untersucht, was für ein Modell resultiert, so könnte man erwarten, dass sich die Schlussfolgerungen im Vergleich zum ungruppierten Ausgangsmodell verändern. Ist dies nicht der Fall, so hat man ein zerlegbares (flexibles) Modell. In Tutz (2012) spricht man von *Collapsibility*, was man als Flexibilität eines Modells verstehen kann.

Definition 3.2.2 (Flexibilität, Collapsibility)**.** *Sei Y eine Responsevariable, die in c verschiedene Kategorien fallen kann. Gruppiert man die c Ausprägungen in $G_r = \{g_{r-1}+1, \ldots, g_r\}, r = 1, \ldots, k$ mit $g_0 = 0$ und $g_k = c$ und definiert die Variable $\tilde{Y} = j$ falls $Y \in G_j$, so spricht man von einem flexiblen oder zerlegbaren Modell bezüglich einem Parameter, falls dieser Parameter sowohl für das Modell bezüglich Y als auch für das Model bezüglich $\tilde{Y}$ derselbe ist, das heißt durch die Transformation mittels Gruppierung bleibt der Parameter des Modells unverändert.*

Für das kumulative Modell gilt

$$P(Y \leq j \mid \boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}) \tag{3.3}$$

und mit den entsprechenden Gruppierungen

$$P(\tilde{Y} \leq j \mid \boldsymbol{x}) = F(\alpha_{g_j} - \boldsymbol{\beta}^T \boldsymbol{x}).$$

Wir können daraus schließen, dass es flexibel bezüglich des Parameters $\boldsymbol{\beta}$ ist. Durch das Zusammenfassen der Kategorien kann man zudem eine Flexibilität bezüglich des Intercepts α_j erhalten.

Weitere Darstellungen

Oftmals findet man in (3.3) vor dem Parametervektor $\boldsymbol{\beta}$ auch ein positives Vorzeichen. Je nach Auswahl der Statistik-Software, sollte vorher darauf geachtet werden, welches Vorzeichen verwendet wird, damit es nicht zu falschen Interpretationen kommt. `SPSS` und die Funktionen `polr`, `lrm` und `clm` in `R` verwenden das Modell mit negativem, `SAS` und die `R`-Funktion `vglm` jenes mit positivem Vorzeichen. Obige Erklärung motiviert jedoch eher die Wahl des negativen Vorzeichens.

3.3. Verallgemeinertes kumulatives Modell

Bisher haben wir angenommen, dass unser zu schätzender Parametervektor $\boldsymbol{\beta}$ für alle Kategorien gleich ist. Es gibt aber auch Modelle, bei denen das nicht so ist, das heißt, dass sich bei diesen für jede Stufe j ein anderer Vektor $\boldsymbol{\beta}_j$ erlaubt ist. Es ist auch möglich, dass es für einige Stufen ein globales $\boldsymbol{\beta}$ gibt und sich für die restlichen Stufen

verschiedene $\boldsymbol{\beta}_j$ ergeben. Bringen wir das ganze wieder in den Zusammenhang zu einem binären Modell und teilen unsere Kategorien in Gruppen $\{1, \ldots, j\}$ und $\{j+1, \ldots, c\}$, so kann das verallgemeinerte kumulative Modell als eine Kombination binärer Response-Modelle angesehen werden. Dabei sind die Parameter für jede Gruppierung verschieden. Dieses Modell wird auch als heteroskedastisches Modell bezeichnet.

Das verallgemeinerte kumulative Modell

Es gilt

$$P(Y \leq j \mid \boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x}), \quad j = 1, \ldots, c-1.$$

Dabei muss

$$P(Y \leq j \mid \boldsymbol{x}) \leq P(Y \leq j+1 \mid \boldsymbol{x})$$

erfüllt sein. Dies impliziert für alle $\boldsymbol{x}$ und alle c Kategorien folgendes

$$\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x} \leq \alpha_{j+1} - \boldsymbol{\beta}_{j+1}^T \boldsymbol{x}.$$

Der Unterschied zwischen dem einfachen und dem verallgemeinerten kumulativen Modell besteht also nur im linearen Prädiktor

$$\eta_j = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x} \quad \text{(einfaches kumulatives Modell)},$$

$$\eta_j = \alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x} \quad \text{(verallgemeinertes kumulatives Modell)}.$$

Hier wird auch das kumulative Logit-Modell am häufigsten verwendet, obwohl die simple Interpretation wie beim einfachen kumulativen Modell nicht mehr möglich ist. Der Einfluss des Parameters auf $\boldsymbol{x}$ muss für jede Unterteilung extra interpretiert werden.

Das verallgemeinerte kumulative Logit-Modell

Das Modell ist beschrieben durch

$$P(Y \leq j \mid \boldsymbol{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x})}, \quad \text{bzw.}$$

$$\log \frac{P(Y \leq j \mid \boldsymbol{x})}{P(Y > j \mid \boldsymbol{x})} = \alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x}.$$

Wie zu Beginn des Abschnitts schon angemerkt ist es auch möglich, dass es einen Teil der erklärenden Variablen gibt, für den der Einfluss vom Parameter $\boldsymbol{\beta}$ für alle Kategorien gleich ist und einen Teil, für den es stufenabhängige $\boldsymbol{\beta}_j$ gibt. Deswegen kann es beim letzteren Prädiktor auch sinnvoll sein, $\boldsymbol{\beta}_j$ in einen von den Kategorien unabhängigen und einen davon abhängigen Teil zu zerlegen, sodass

$$\eta_j = \alpha_j - \bar{\boldsymbol{\beta}}_j^T \bar{\boldsymbol{x}} - \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{x}}.$$

Teilen wir den Vektor der erklärenden Variablen in diese zwei Komponenten $\boldsymbol{x}^T = (\bar{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$, wobei für $\bar{\boldsymbol{x}}$ die Parameter kategorie-abhängig sind (kategorie-spezifischer Effekt) und für $\tilde{\boldsymbol{x}}$ die Parameter alle gleich sind (globaler Effekt), so können wir das sogenannte **partielle kumulative Logit-Modell** schreiben als

$$\log \frac{P(Y \leq j \mid \boldsymbol{x})}{P(Y > j \mid \boldsymbol{x})} = \alpha_j - \bar{\boldsymbol{\beta}}_j^T \bar{\boldsymbol{x}} - \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{x}}.$$

Proportionale Odds erhalten wir hierbei nur für den Teil in $\tilde{\boldsymbol{x}}$.

Sind also die erklärenden Variablen für zwei Objekte nur im Teil $\bar{\boldsymbol{x}}$ verschieden und im Teil $\tilde{\boldsymbol{x}}$ ident, das heißt wenn sie sich beschreiben lassen durch $(\bar{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}), (\bar{\boldsymbol{x}}_2, \tilde{\boldsymbol{x}})$, dann erhalten wir Odds Ratios, die unabhängig von der Kategorie sind

$$\begin{aligned}\frac{P\left(Y_1 \leq j|(\bar{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}})\right) / P\left(Y_1 > j|(\bar{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}})\right)}{P\left(Y_2 \leq j|(\bar{\boldsymbol{x}}_2, \tilde{\boldsymbol{x}}\right) / P\left(Y_2 > j|(\bar{\boldsymbol{x}}_2, \tilde{\boldsymbol{x}})\right)} &= \frac{\exp(\alpha_j - \bar{\boldsymbol{\beta}}^T \bar{\boldsymbol{x}}_1 - \tilde{\boldsymbol{\beta}}_j^T \tilde{\boldsymbol{x}})}{\exp(\alpha_j - \bar{\boldsymbol{\beta}}^T \bar{\boldsymbol{x}}_2 - \tilde{\boldsymbol{\beta}}_j^T \tilde{\boldsymbol{x}})} \\ &= \exp\left(-\bar{\boldsymbol{\beta}}^T \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)\right).\end{aligned}$$

3.4. Testen auf proportionale Odds

Beim kumulativen Logit-Modell haben wir gesehen, dass die Odds Ratios sich zu

$$\frac{P\left(Y_1 \leq j|\boldsymbol{x}_1\right) / P\left(Y_1 > j|\boldsymbol{x}_1\right)}{P\left(Y_2 \leq j|\boldsymbol{x}_2\right) / P\left(Y_2 > j|\boldsymbol{x}_2\right)} = \exp\left(-\boldsymbol{\beta}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right) \quad j = 1, \ldots, c-1.$$

ergeben. Sie sind unabhängig vom Intercept α und daher gleich für alle Kategorien. Außerdem sind sie proportional zum Abstand der erklärenden Variablen $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$. Soll dieses Modell verwendet werden, so ist zu überprüfen ob die Odds, so wie hier dargestellt, proportional sind. Dies erweist sich als relativ leicht, wenn man davon ausgeht,

dass das kumulative Logit-Modell ein Untermodell des verallgemeinerten kumulativen Modells mit Parametervektoren $\boldsymbol{\beta}_j$ ist. Will man auf proportionale Odds testen, so ist dies äquivalent dazu, die Nullhypothese

$$H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{c-1}$$

zu überprüfen. Es ist auch möglich diese Hypothese für jede erklärende Variable x_j extra zu betrachten und zu untersuchen ob man die Nullhypothese

$$H_0 : \beta_{1j} = \cdots = \beta_{c-1,j} = \beta_j$$

verwerfen kann oder nicht. Ist H_0 wahr, so bedeutet dies, dass für zwei erklärende Variablen $\boldsymbol{x}_1^T = (x_1, \ldots, x_j, \ldots, x_p), \boldsymbol{x}_2^T = (x_1, \ldots, \tilde{x}_j, \ldots, x_p)$, welche sich nur in der j-ten Komponente unterscheiden, das kumulative Odds Ratio

$$\frac{P(Y \leq j \mid \boldsymbol{x}_1)/P(Y > j \mid \boldsymbol{x}_1)}{P(Y \leq j \mid \boldsymbol{x}_2)/P(Y > j \mid \boldsymbol{x}_2)} = \exp(\beta_j(x_j - \bar{x_j})$$

unabhängig von der Kategorie ist. Um diese Hypothese zu testen kann man beispielsweise den Likelihood-Quotienten-Test oder den Wald-Test (siehe Abschnitt 2.2) verwenden.

3.5. Schätzen der Parameter mittels der Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode ist eine der am häufigsten verwendeten Methoden zur Konstruktion plausibler Schätzer, mit denen die Modellverteilung die beobachteten Daten möglichst genau beschreibt. Der Vorteil ist, dass die Schätzer asymptotisch erwartungstreu und konsistent sind, das heißt je größer der Stichprobenumfang, desto besser die Schätzung. Außerdem konvergieren die Schätzer in Verteilung gegen eine normalverteilte Zufallsvariable. Die Varianz dieser ist die Inverse der Fisher-Information, das heißt sie erreicht die Cramér-Rao-Schranke und ist daher in asymptotischer Betrachtung optimal. Man spricht dabei von asymptotischer Effizienz.

Mit Hilfe dieser Methode (vergleiche Agresti, 2010) können wir Schätzer für die Parameter α und $\boldsymbol{\beta}$ herleiten. Dafür gehen wir von n unabhängigen multinomialverteilten Zufallsvariablen $Y_i, i = 1, \ldots, n$, aus. Die zugehörigen erklärenden Variablen $\boldsymbol{x}_i$ enthalten alle r Informationen, auf die die zu berechnenden Wahrscheinlichkeiten bedingt sind.

Jede der n Beobachtungen Y_i fällt in eine von c Kategorien mit einer Wahrscheinlichkeit von

$$\pi_j(\boldsymbol{x}_i) = P(Y_i = j \mid \boldsymbol{x}_i), \quad j = 1, \ldots, c,$$

welche sich auch über die Differenz der kumulativen Wahrscheinlichkeiten darstellen lässt

$$\begin{aligned} P(Y_i = j \mid \boldsymbol{x}_i) &= P(Y_i \leq j \mid \boldsymbol{x}_i) - P(Y_i \leq j-1 \mid \boldsymbol{X} = \boldsymbol{x}_i) \\ &= F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i). \end{aligned}$$

Mit Hilfe von Indikatorvariablen

$$y_{ij} = \begin{cases} 1 & \text{wenn } Y_i = j \\ \\ 0 & \text{sonst} \end{cases}$$

erhält man folgende Produktdarstellung

$$P(Y_i = j \mid \boldsymbol{x}_i) = \prod_{j=1}^{c} P(Y_i = j \mid \boldsymbol{x}_i)^{y_{ij}}.$$

Um die Likelihood-Funktion aufzustellen, müssen die Wahrscheinlichkeiten aller unabhängigen Beobachtungen aufmultipliziert werden, das heißt mit $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_c)^T$ ist

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} P(Y_i = j \mid \boldsymbol{x}_i)^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} \left[P(Y_i \leq j \mid \boldsymbol{x}_i) - P(Y_i \leq j-1 \mid \boldsymbol{x}_i) \right]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} \left[F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) \right]^{y_{ij}} \right\}. \end{aligned}$$

Nimmt man wie im kumulativen Logit-Modell für F die logistische Verteilung, so erhalten wir die folgende Likelihood-Funktion

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} \left[\frac{\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right]^{y_{ij}} \right\}.$$

Durch Nullsetzen der partiellen Ableitungen nach den zu schätzenden Parametern, erhält man deren Schätzer. Oft ist es jedoch leichter die partiellen Ableitungen des Logarithmus der Likelihood-Funktion zu berechnen und diese Null zu setzen, was zum selben Ergebnis führt. Wir erhalten

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y}) = \log L(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{c} y_{ij} \log \left[F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) \right] \right\}.$$

Die partiellen Ableitungen nach α_j für $j = 1, \ldots, c$ und β_k für $k = 1, \ldots, r$ lassen sich nun leicht berechnen. Sei f die zugehörige Dichtefunktion zu F ($F' = f$) und x_{ik} die k-te Komponente von $\boldsymbol{x}_i$, so ergibt sich

$$\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \alpha_j} = \sum_{i=1}^{n} \left\{ y_{ij} \frac{f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)} - y_{i,j+1} \frac{f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{F(\alpha_{j+1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right\}$$

und

$$\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_k} = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{c} y_{ij} \frac{x_{ik}(f(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) - f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i))}{F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right\}.$$

Um die Schätzer $\hat{\alpha}_j$ und $\hat{\beta}_k$ zu berechnen, müssen die Gleichungen

$$\sum_{i=1}^{n} \left\{ \frac{y_{ij} f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)} - \frac{y_{i,j+1} f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)}{F(\alpha_{j+1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right\} = 0$$

$$\sum_{i=1}^{n} \left\{ \sum_{j=1}^{c} y_{ij} \frac{x_{ik}(f(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) - f(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i))}{F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right\} = 0$$

gelöst werden. Da diese nicht linear in den Parametern sind, muss man sich iterativen Verfahren bedienen. Es erweist sich beispielsweise der Fisher-Scoring-Algorithmus als sehr hilfreich, welcher auch von McCullagh (1980) genutzt wurde. Er ist angelehnt an das Newton-Raphson-Verfahren, welches numerische Lösungen von nichtlinearen Gleichungen und Gleichungssystemen generiert. Sie unterscheiden sich bezüglich der Wahl der Ableitungs-Matrix. Während Newton-Raphson auf der beobachteten negativen Hessematrix aufbaut, verwendet der Fisher-Scoring-Algorithmus deren Erwartungswert, die sogenannte Informationsmatrix. Für kanonische Linkfunktionen stimmen diese beiden

überein, da die Hessematrix unabhängig von den Beobachtungen Y ist. Beide Algorithmen liefern dieselben Schätzer für $\hat{\boldsymbol{\beta}}$. Eine genauere Beschreibung dazu ist im Abschnitt A.2 zu finden.

3.5.1. Schätzen der Standardabweichungen

Wie zu Beginn vom vorhergehenden Abschnitt besprochen, sind Maximum-Likelihood-Schätzer asymptotisch effizient. Daher kann die Varianz der Schätzer mittels der Berechnung der Inversen der Fisher-Informationsmatrix bestimmt werden. Die Fisher-Informationsmatrix ist die Varianz des Scorevektors. Wir stellen sie hier als Matrix dar, welche aus vier Teilmatrizen besteht, das heißt

$$\boldsymbol{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{A}^1 & \boldsymbol{A}^2 \\ \boldsymbol{A}^3 & \boldsymbol{A}^4 \end{pmatrix},$$

wobei $\boldsymbol{A}^1$ eine $c \times c$ - , $\boldsymbol{A}^2$ eine $c \times r$ - , $\boldsymbol{A}^3$ eine $r \times c$ - und $\boldsymbol{A}^4$ eine $r \times r$ - Matrix ist. Da der Erwartungswert des Scorevektors $\mathbf{0}$ ist, ergibt sich folgende Darstellung für die vier Teilmatrizen in Vektornotation

$$\begin{aligned}
\boldsymbol{A}^1 &= E\left[\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\alpha}^T}\right)\right] \quad j, l = 1, \ldots, c \\
\boldsymbol{A}^2 &= E\left[\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\alpha}_j}\right)\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_k}\right)\right] \quad j = 1, \ldots, c, k = 1, \ldots, r \\
\boldsymbol{A}^3 &= E\left[\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_k}\right)\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\alpha}_j}\right)\right] \quad k = 1, \ldots, r, j = 1, \ldots, c \\
\boldsymbol{A}^4 &= E\left[\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_k}\right)\left(\frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_m}\right)\right] \quad k, m = 1, \ldots, r.
\end{aligned}$$

Dabei sind $\boldsymbol{\alpha}$ und $\boldsymbol{\beta}$ die zu schätzenden Parameter und ℓ ist die Log-Likelihood-Funktion der Stichprobe. Man kann hier die Symmetrie der Fisher-Informationsmatrix erkennen.

Unter gewissen Regularitätsbedingungen können wir die Fisher-Informationsmatrix auch mittels der zweiten partiellen Ableitungen der Log-Likelihood-Funktion berechnen

$$\boldsymbol{A}^1 = -E\left[\frac{\partial^2 \ell(\alpha, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\right].$$

Analoges gilt für $\boldsymbol{A}^2$, $\boldsymbol{A}^3$ und $\boldsymbol{A}^4$. Da die Hessematrix zur Log-Likelihood-Funktion $\boldsymbol{H}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ genau die zweiten partiellen Ableitungen enthält, ergibt sich im Bezug zur

Fisher-Informationsmatrix $\boldsymbol{I}$ folgenden Zusammenhang

$$\boldsymbol{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -E[\boldsymbol{H}(\boldsymbol{\alpha}, \boldsymbol{\beta})].$$

Es genügt also die Hessematrix zu bestimmen, um die Fisher-Informationsmatrix und damit die Varianz-Kovarianz-Matrix der Schätzer zu erhalten. Kennt man wie in unserem Fall die wahren Parameter nicht, so ersetzt man diese durch ihre Schätzer und gelangt auf diese Weise zur geschätzten asymptotischen Varianz-Kovarianz-Matrix. Zieht man die Wurzel aus den Diagonalelementen erhält man die geschätzten Standardabweichung der Schätzer. Die negative Hessematrix ausgewertet in den geschätzten Parametern $-H(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ wird auch als beobachtete Fisher-Information bezeichnet.

Schlussendlich wollen wir die zweiten partiellen Ableitungen der Log-Likelihoodfunktion für kumulative Modelle angeben, welche die Hessematrix definieren. Dabei soll die Notation $t_{ij} = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i$ verwendet werden. Für die zweiten partiellen Ableitungen nach $\boldsymbol{\alpha}$ unterscheiden wir drei Fälle.

$$\frac{\partial^2 \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \alpha_j \partial \alpha_l} = \begin{cases} \displaystyle\sum_{i=1}^{n} y_{ij} \left\{ \frac{f(t_{ij})}{F(t_{ij}) - F(t_{i,j-1})} \right\} & l = j-1, \\ \displaystyle\sum_{i=1}^{n} y_{ij} \left\{ \frac{f'(t_{ij})}{F(t_{ij}) - F(t_{i,j-1})} - \frac{f(t_{ij})(f(t_{ij}) - f(t_{i,j-1}))}{(F(t_{ij}) - F(t_{i,j-1}))^2} \right\} & \\ \quad - y_{i,j+1} \left\{ \frac{f'(t_{ij})}{F(t_{i,j+1}) - F(t_{ij})} - \frac{f(t_{ij})(f(t_{i,j+1}) - f(t_{ij}))}{(F(t_{i,j+1}) - F(t_{ij}))^2} \right\} & l = j, \\ \displaystyle\sum_{i=1}^{n} y_{i,j+1} \left\{ \frac{f(t_{ij})}{F(t_{i,j+1}) - F(t_{i,j})} \right\} & l = j+1. \end{cases}$$

Die zweiten partiellen Ableitungen nach $\boldsymbol{\alpha}$ sind in allen anderen Fällen gleich 0. Für die gemischten Ableitungen ergibt sich

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \alpha_j \partial \beta_k} = & \sum_{i=1}^{n} y_{i,j+1} x_{ik} \left\{ \frac{f'(t_{ij})}{F(t_{i,j+1}) - F(t_{ij})} - \frac{f(t_{ij})(f(t_{i,j+1}) - f(t_{ij}))}{(F(t_{i,j+1}) - F(t_{ij}))^2} \right\} \\ & - y_{ij} x_{ik} \left\{ \frac{f'(t_{ij})}{F(t_{ij}) - F(t_{i,j-1})} - \frac{f(t_{ij})(f(t_{ij}) - f(t_{i,j-1}))}{(F(t_{ij}) - F(t_{i,j-1}))^2} \right\} \end{aligned}$$

und die zweiten partiellen Ableitungen bezüglich $\boldsymbol{\beta}$ sind

$$\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y}) \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \beta_k \partial \beta_m} = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} x_{ik} x_{im} \left\{ \frac{(f(t_{i,j-1}) - f(t_{ij}))(f(t_{ij}) - f(t_{i,j-1}))}{(F(t_{ij}) - F(t_{i,j-1}))^2} \right.$$
$$\left. - \frac{f'(t_{i,j-1}) - f'(t_{ij})}{F(t_{ij}) - F(t_{i,j-1})} \right\}.$$

In der R-Funktion `polr` werden die Parameter wie hier beschrieben geschätzt. Im nächsten Abschnitt gehen wir darauf noch genauer ein.

3.6. Die Funktion `polr`

Um die Parameter $\boldsymbol{\alpha}$ und $\boldsymbol{\beta}$ in kumulativen Modellen zu schätzen, steht uns die R-Funktion `polr` zur Verfügung. Außerdem werden damit die Standardfehler der Schätzer berechnet. Wie zuvor beschrieben werden in `polr` die zweiten partiellen Ableitungen (Hessematrix) bestimmt. Somit gelangt man zur Fisher-Informationsmatrix und erhält durch Wurzelziehen der Diagonalelemente die geschätzten Standardabweichungen.

Die Funktion `polr` steht für *proportional odds logistic regression* und erlaubt unter anderem folgende Argumente

```
polr(formula, data, weights, start, Hess = FALSE,
     method = c("logistic", "probit", "cloglog", "cauchit"),...).
```

In `formula` gibt man sein Modell in der Form `Response` $\sim$ `Prädiktor` an, wobei die Response ein geordneter Faktor, also eine ordinal skalierte Variable sein sollte. Der Faktor wird als ordinale Response interpretiert, wobei von einer lexikografischen Sortierung ausgegangen wird. Bei einer Zahlencodierung wird also davon ausgegangen, dass $0 < 1 < 2 < \ldots$. Man kann hier auch mit der Funktion `ordered` arbeiten, welche die Sortierung der Variablen (auch in lexikografischer Reihenfolge) annimmt. Der Unterschied zwischen einem Faktor und einem mit `ordered` geordneten Faktor liegt also nur darin, das zweiterer Level-Attribute hat. Mit `as.ordered` erreicht man dasselbe. Das nachfolgende Beispiel verdeutlicht dies.

```
> Status
[1] "Niedrig" "Hoch"    "Mittel"  "Hoch"    "Mittel"
> ordered(Status)
[1] Niedrig Hoch    Mittel  Hoch    Mittel
```

```
Levels: Hoch < Mittel < Niedrig
> as.ordered(Status)
[1] Niedrig Hoch    Mittel  Hoch    Mittel
Levels: Hoch < Mittel < Niedrig
> Status.sortiert <- factor(Status, levels=c("Niedrig",
                            "Mittel", "Hoch"), ordered=TRUE)
> Status.sortiert
[1] Niedrig Hoch    Mittel  Hoch    Mittel
Levels: Niedrig < Mittel < Hoch
```

Zuletzt ist auch angeführt, wie man einen Faktor nach eigenem belieben sortieren kann.

Die zugrundeliegenden Daten bezieht man über `data` mit ein. Mittels `weight` können (müssen aber nicht) Gewichte zugeordnet werden. Ebenso kann man optional über `start` Startwerte für die Parameter $\boldsymbol{\alpha}, \boldsymbol{\beta}$ vorgeben. Möchte man sich das Summary des entstehenden `polr`-Objekts ansehen, so sollte `Hess=TRUE` gesetzt werden. Dies liefert die Informationsmatrix, welche man zur Berechnung der Standardfehler benötigt. Mittels `method` kann die Verteilung der latenten Variable über die Linkfunktion festgelegt werden, wobei hier die logistische, die Normal-, die Gumbel- und die Cauchy-Verteilung zur Verfügung stehen. Da man mittels der Transformation der Gumbel- zur Gombertz-Verteilung gelangt, ist auch diese nicht ausgeschlossen. Wird nichts angegeben, so wird default-mäßig die logistische Verteilung angenommen, die der Funktion ihren Namen gegeben hat.

Man erhält die Schätzer der Parameter α_j und $\boldsymbol{\beta}$, welche unter `coefficients` im `summary` des Objekts zu finden sind. Im Summary findet man neben den Schätzern die Standardfehler und weiters die Residual Deviance sowie den Wert des Akaike Information Kriterium (AIC), welche beide nützlich für den Vergleich von verschiedenen Modellen sind.

Analysiert man den Quellcode von `polr`, so stößt man auf mehrere Unterfunktionen. Zur Berechnung der Wahrscheinlichkeiten wird die entsprechende Verteilungsfunktion, welche in `method` angegeben wurde, benötigt. Die Funktion `pfun(t)` wählt die zugehörige Verteilungsfunktion aus und berechnet die Wahrscheinlichkeit, dass eine nach `method` verteilte Zufallsvariable kleiner oder gleich t ist.

```
pfun <- switch(method, logistic = plogis, probit = pnorm,
```

```
cloglog = pgumbel, cauchit = pcauchy)
```

Wird beispielsweise `method=logistic` angegeben, so ist

$$\texttt{pfun}(\alpha_j - \boldsymbol{\beta}^T x) = F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})}.$$

In der Funktion `fmin` wird `pfun` benötigt, um die Wahrscheinlichkeiten $P(Y = j \mid \boldsymbol{x})$ für alle $j = 1, \ldots, c$ zu berechnen

$$\begin{aligned}
P(Y = j \mid \boldsymbol{x}) &= F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}) - F(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}), \text{ mit} \\
P(Y = 1 \mid \boldsymbol{x}) &= F(\alpha_1 - \boldsymbol{\beta}^T \boldsymbol{x}) - F(-\infty) = F(\alpha_1 - \boldsymbol{\beta}^T \boldsymbol{x}) \text{ und} \\
P(Y = c \mid \boldsymbol{x}) &= F(\alpha_c - \boldsymbol{\beta}^T \boldsymbol{x}) - F(\alpha_{c-1} - \boldsymbol{\beta}^T \boldsymbol{x}) = 1 - F(\alpha_{c-1} - \boldsymbol{\beta}^T \boldsymbol{x}).
\end{aligned}$$

Mit diesen Wahrscheinlichkeiten wird die Log-Likelihood-Funktion der Stichprobe definiert. Dabei werden die α-Parameter so parametrisiert, dass sie monoton wachsend sind. Man setzt $\alpha_{0|1} = -\infty$ und wählt für das darauffolgende $\alpha_{1|2}$ einen Wert δ_1. Dann wird stets zum vorhergehenden $\alpha_{j|j+1}$ ein positiver Term in Form von $\exp(\delta_j) > 0$ mit $\delta_j \in \mathbb{R}, j = 1, \ldots, c-1$ dazu addiert und somit die Monotonie erreicht. Bei einem Modell mit einer Responsevariable mit c Stufen ist dann

$$\begin{aligned}
\alpha_0 &= \alpha_{0|1} = -\infty \\
\alpha_1 &= \alpha_{1|2} = \delta_1 \\
\alpha_2 &= \alpha_{2|3} = \alpha_{1|2} + \exp(\delta_2) \\
&\ldots \\
\alpha_{k-1} &= \alpha_{k-1|k} = \alpha_{k-2|k-1} + \exp(\delta_{k-1}) \\
&\ldots \\
\alpha_c &= \alpha_{c|c+1} = \infty
\end{aligned}$$

Es werden bei der Schätzung der Parameter mittels der Maximum-Likelihood-Methode in `polr` die partiellen Ableitungen nach $\boldsymbol{\beta}$ und $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{c-1})^T$ gebildet und bezüglich diesen maximiert. Somit erhält man auch die optimalen Schätzer für die Intercepts α, da sich diese mit Hilfe von $\boldsymbol{\delta}$ berechnen lassen.

In Analogie zu `pfun` gibt die Funktion `dfun(y)` den Wert der in `method` gewählten Dichtefunktion an der Stelle y zurück. Die Dichtefunktionen werden bei der Bestimmung der partiellen Ableitungen der Log-Likelihood-Funktion benötigt, was unter anderem in der Funktion `gmin` geschieht. Zunächst werden auch in `gmin` die Punktwahrscheinlichkeiten $P(Y = j \mid \boldsymbol{x})$ für $j = 1, \ldots, c$ berechnet. Beim Ableiten der Log-Likelihood-Funktion müssen die Wahrscheinlichkeiten $P(Y = j \mid \boldsymbol{x})$ abgeleitet werden. Da die Ableitung der Verteilungsfunktion der zugehörigen Dichtefunktion entspricht, kann dafür an dieser Stelle die Funktion `dfun` genutzt werden

$$F'(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}) = f(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}) = \texttt{dfun}(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}).$$

Damit werden die partiellen Ableitungen der Log-Likelihood-Funktion nach α und $\boldsymbol{\beta}$ wie in Abschnitt 3.5 bestimmt.

Schließlich werden `fmin` und `gmin` im Zusammenhang mit der Funktion `optim` verwendet. Es werden die Nullstellen der partiellen Ableitungen der Log-Likelihood-Funktion numerisch bestimmt. Diese minimieren die negative Log-Likelihood-Funktion und bilden die optimalen Maximum-Likelihood-Schätzer für $\boldsymbol{\alpha}$ und $\boldsymbol{\beta}$. Zudem wird die Hessematrix berechnet. Mittels dieser können die Standardfehler (so wie in 3.5.1 beschrieben) ermittelt werden.

3.7. Weitere Funktionen zur Analyse kumulativer Modelle

In diesem Abschnitt werden die Anwendungen weiterer möglicher `R`-Funktionen erläutert. Auf der Homepage des Programms http://www.r-project.org/ sind jeweils die detaillierten Beschreibungen aller Funktions-Argumente zu finden.

3.7.1. Die Funktion `vglm`

Die `R`-Funktion `vglm` aus dem Package `VGAM` ist zum Anpassen von Vektor-generalisierten linearen Modellen geeignet. Yee (2010) bietet dazu eine sehr ausführliche Beschreibung.

```
vglm(formula, family=...(link=... , parallel=TRUE, reverse=FALSE), data = list(),
weights = NULL, method = "vglm.fit", ...)
```

Diese Funktion erlaubt die Verwendung der generalisierten linearen Modelle und bietet unter anderem die Möglichkeit Parameter für kumulative Logit-Modelle zu schätzen, in dem man `family=cumulative` setzt. Dabei wird anders als bei `polr` vom Model mit einem positiven Vorzeichen vor dem Parameter $\boldsymbol{\beta}$ ausgegangen. Über `link=...` kann die Linkfunktion abgeändert werden, beispielsweise zur Normalverteilung (`probit`). Durch die Angabe von `parallel=TRUE` werden Parameter berechnet die auf der Annahme proportionaler Odds beruhen. Durch Weglassen dieser Option hat man auch die Möglichkeit Modelle zu schätzen, bei denen keine proportionalen Odds vorliegen, das heißt, dass für jede Stufe verschiedene β_j-Parameter existieren können. Des Weiteren erhält man mittels

```
family=sratio(reverse=FALSE,parallel=TRUE)
```

Schätzer für sequentielle Modelle, welche im nachfolgenden Kapitel erläutert werden. Dabei steht `sratio` für stopping ratios und wird für das Modell

$$\text{logit}(P(Y = j|Y \geq j)) = \alpha_j + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}$$

genutzt. Verwendet man stattdessen `cratio` (continuation ratios) wird vom Modell

$$\text{logit}(P(Y > j|Y \geq j)) = \alpha_j + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}$$

ausgegangen. Setzt man `reverse=FALSE` kehren sich die Relationen um, das heißt aus $Y \geq j$ wird $Y \leq j$ und bei `cratio` wird zudem die Wahrscheinlichkeit $Y < j$ betrachtet.

Durch die Vorgabe `familiy=acat(reverse=TRUE,parallel=TRUE)` können die in Abschnitt 6.3 beschriebenen Adjazent-Kategorie-Modelle analysiert werden. Eine Übersicht über einige mögliche Modelle ist in Tabelle 3.1 dargestellt (ohne spezielle Linkfunktion).

3.7.2. Die Funktion `lrm`

Die Funktion `lrm` aus dem Package `rms` schätzt die Parameter von binären und ordinalen logistischen Regressionsmodellen mit proportionalen Odds mittels der Maximum-Likelihood-Methode.

```
lrm(formula, data, subset, na.action=na.delete, method="lrm.fit", ...)
```

Angaben in `vglm`	Modell
`cumulative`	$P(Y \leq j)$
`cumulative(reverse=TRUE)`	$P(Y \geq j)$
`sratio`	$P(Y = j \mid Y \geq j)$
`sratio(reverse=TRUE)`	$P(Y = j \mid Y \leq j)$
`cratio`	$P(Y > j \mid Y \geq j)$
`cratio(reverse=TRUE)`	$P(Y < j \mid Y \leq j)$
`acat`	$P(Y = j + 1)/P(Y = j)$
`acat(reverse=TRUE)`	$P(Y = j)/P(Y = j + 1)$

Tabelle 3.1.: Modellspezifikationen für `vglm`

3.7.3. Die Funktion `clm`

Kumulative Link-Modelle können ebenfalls auch mittels der Funktion `clm` bearbeitet werden. Diese liefert mittels eines modifizierten Newton-Algorithmus‘ die Schätzer unter verschiedenen Linkfunktionen. Außerdem lässt sie verschiedene Einstellung für die Schwellenwerte (`threshold`) zu. Die standardmäßig unstrukturierten Schwellenwerte erhält man über `flexible`, `symmetric` sorgt für symmetrische Abstände zwischen den Schwellenwerten um den zentralen Schwellenwert. Dieser wird mittels `symmetric2` Null gesetzt und sorgt somit dafür, dass der Mittelwert der Schwellenwerte Null ist. Über `equidistant` erreicht man gleiche Abstände zwischen aufeinanderfolgenden Schwellenwerten. Christensen (2012) bietet eine umfangreiche Dokumentation über weitere Anwendungsmöglichkeiten.

```
clm(formula, scale, nominal, data, weights, start, subset, doFit = TRUE,
na.action, contrasts, model = TRUE, control=list(),
link = c("logit", "probit", "cloglog", "loglog", "cauchit"),
threshold = c("flexible", "symmetric", "symmetric2", "equidistant"), ...)
```

	Astrologie ist wissenschaftlich		
Schulbildung	gar nicht	mehr oder weniger	sehr
<High school	98 (48%)	84 (41%)	23 (11%)
High school	574 (63%)	286 (31%)	50 (5%)
Junior College	122 (72%)	44 (26%)	4 (2%)
Bachelor	268 (80%)	57 (17%)	11 (3%)
Graduate	148 (86%)	23 (13%)	1 (1%)

Tabelle 3.2.: Meinung über die Wissenschaftlichkeit von Astrologie in Abhängigkeit von der Bildung der Befragten

3.8. Vergleich der Funktionen anhand eines Beispiels

Beispiel 4. *Wir betrachten den Datensatz in Tabelle 3.2. Dieser wurde im Rahmen des General Social Survey 2006 erhoben. Dabei wurden 1793 Personen mit verschiedenem Bildungsniveau zu ihrer Meinung bezüglich der Astrologie befragt. Genauer gesagt ging es dabei um die Fragestellung „Würden Sie sagen, dass Astrologie sehr wissenschaftlich, mehr oder weniger wissenschaftlich oder gar nicht wissenschaftlich ist?“ Hier ist bereits ein Trend zu erkennen, der zeigt, dass Personen mit höherer Bildung eher dazu neigen, die Astrologie nicht als wissenschaftlich anzusehen, Personen mit geringerer Bildung jedoch eher.*

3.8.1. Analyse mit `polr`

Wir nehmen für die Daten ein kumulatives Logit-Modell an. Mithilfe der `R`-Funktion `polr`, welche in Abschnitt 3.6 genauer beschrieben ist, wollen wir die Schätzer berechnen. Damit können die geschätzten Punktwahrscheinlichkeiten für die einzelnen Ereignisse angegeben werden. Da die lexikografische Ordnung von `Einschaetzung` der gewünschten Reihenfolge entspricht, brauchen wir keine Umsortierung vorgeben und geben hier lediglich die Levels mittels der Funktion `ordered` hinzu.

```
> library(MASS)
```

```
Einschaetzung<-  factor(c(rep(c("gar nicht","mehr oder weniger", "sehr"), 5)))
Gewichte <- c(98,84,23,574,286,50,122,44,4,268,57,11,148,23,1)
Bildung_  <- factor(c(rep("<High school", 3),rep("High school", 3),
             rep("Junior college", 3),rep("Bachelor", 3),rep("Graduate", 3)))

> ordered(Einschaetzung)
 [1] gar nicht         mehr oder weniger sehr
 [4] gar nicht         mehr oder weniger sehr
 [7] gar nicht         mehr oder weniger sehr
[10] gar nicht         mehr oder weniger sehr
[13] gar nicht         mehr oder weniger sehr
Levels: gar nicht < mehr oder weniger < sehr
> fit_logistic <- polr(ordered(Einschaetzung) ~ Bildung_, weights=Gewichte)
> summary(fit_logistic)

Re-fitting to get Hessian

Call:
polr(formula = ordered(Einschaetzung) ~ Bildung_, weights = Gewichte)

Coefficients:
                         Value Std. Error t value
Bildung_Bachelor       -1.4746     0.1917  -7.692
Bildung_Graduate       -1.9439     0.2582  -7.529
Bildung_High school    -0.6497     0.1511  -4.300
Bildung_Junior college -1.0657     0.2164  -4.924

Intercepts:
                             Value   Std. Error t value
gar nicht|mehr oder weniger -0.1129  0.1360    -0.8301
mehr oder weniger|sehr       2.1797  0.1603    13.5986

Residual Deviance: 2656.409
AIC: 2668.409
```

Die Funktion `polr` wählt default-mäßig die logistische Verteilung, von welcher wir hier ausgehen. Die geschätzten Parameter können der obigen Ausgabe entnommen werden. Die geschätzten Intercepts haben die Werte $\hat{\alpha}_1 = -0.1129$ und $\hat{\alpha}_2 = 2.1797$. Weiters

erhalten wir

$$\begin{aligned}
\hat{\beta}(\text{<High school}) &= 0 \\
\hat{\beta}(\text{High school}) &= -0.6497 \\
\hat{\beta}(\text{Junior college}) &= -1.0657 \\
\hat{\beta}(\text{Bachelor}) &= -1.4746 \\
\hat{\beta}(\text{Graduate}) &= -1.9439.
\end{aligned}$$

Als Ausgangsreferenz wurde die Schulbildung „<High school“ gewählt. Die Wahrscheinlichkeit hierzu wird allein durch die Intercepts beschrieben. Wir erkennen, dass die Schätzer für die $\boldsymbol{\beta}$-Parameter mit höher werdender Bildung kleiner werden. Dies liefert die Interpretation, dass die geschätzten Wahrscheinlichkeiten für die Einschätzungen „sehr wissenschaftlich“ und „mehr oder weniger wissenschaftlich“ mit steigendem Bildungsniveau monoton fallend und für „gar nicht wissenschaftlich“ monoton wachsend sind. Dies bestätigt der nachfolgende Code. Es wird beim Schätzen der Wahrscheinlichkeiten mittels `fitted` eine Matrix gebildet, in der jede beobachtete Zeile so oft ausgegeben wird wie es Kategorien von der zu modellierenden Response gibt. Im vorliegenden Beispiel wird also verdreifacht. Die Wahrscheinlichkeiten sind übersichtlich noch einmal in Tabelle 3.3 dargestellt.

```
> fitted(fit_logistic)      # Punktwahrscheinlichkeiten
   gar nicht mehr oder weniger       sehr
1  0.4718110         0.4266039 0.10158508
2  0.4718110         0.4266039 0.10158508
3  0.4718110         0.4266039 0.10158508
4  0.6310845         0.3131639 0.05575160
5  0.6310845         0.3131639 0.05575160
6  0.6310845         0.3131639 0.05575160
7  0.7216840         0.2408249 0.03749105
8  0.7216840         0.2408249 0.03749105
9  0.7216840         0.2408249 0.03749105
10 0.7960471         0.1787281 0.02522477
11 0.7960471         0.1787281 0.02522477
12 0.7960471         0.1787281 0.02522477
```

	Astrologie ist wissenschaftlich		
Schulbildung	gar nicht	mehr oder weniger	sehr
<High school	47.2%	42.7%	10.2%
High school	63.1%	31.3%	5.6%
Junior College	72.2%	24.1%	3.7%
Bachelor	79.6%	17.9%	2.5%
Graduate	86.2%	12.2%	1.6%

Tabelle 3.3.: Geschätzte Wahrscheinlichkeiten für die Astrologiestudie

```
13 0.8618846          0.1221878 0.01592766
14 0.8618846          0.1221878 0.01592766
15 0.8618846          0.1221878 0.01592766
```

Schließlich wollen wir noch die Odds Ratios angeben.

```
> exp(-coef(fit_logistic))  # odds ratio
      Bildung_Bachelor       Bildung_Graduate
              4.369485               6.985993
   Bildung_High school Bildung_Junior college
              1.915058               2.902887
```

Die Odds Ratios sind jeweils im Vergleich zum Bildungsstand „<High school“ zu sehen. Man kann also sagen, dass die Chance, dass Astrologie als nicht wissenschaftlich, statt als eher wissenschaftlich angesehen wird bei Personen mit einem High school-Abschluss nur etwa doppelt so groß ist, wie bei Personen, die weniger als einen High school-Abschluss haben. Diese Chance steigt mit wachsendem Bildungsniveau bis sie schließlich bei Hochschulabsolventen sogar fast 7 mal so groß ist.

3.8.2. Analyse mit `vglm`

Die Funktion `vglm` wollen wir nutzen um zum einen den Vergleich zur Funktion `polr` herzustellen und zum anderen um das Modell zu betrachten, wenn wir davon ausgehen,

dass keine proportionalen Odds vorliegen, das heißt, dass wir kategorieabhängige Parameter β_j zulassen. Die Schätzung des Modells mit proportionalen Odds liefert folgendes.

```
> library(VGAM)
> fit_vglm_proportional2 <- vglm(ordered(Einschaetzung) ~ Bildung_, weights=Gewichte,
                                 family=cumulative(link="logit",parallel=TRUE))
> summary(fit_vglm_proportional2)

Call:
vglm(formula = ordered(Einschaetzung) ~ Bildung_, family = cumulative(link = "logit",
    parallel = TRUE), weights = Gewichte)

Pearson residuals:
                    Min       1Q  Median     3Q     Max
logit(P[Y<=1]) -24.147 -10.6227 -1.7302 5.7077 17.8779
logit(P[Y<=2]) -28.943  -8.9594  1.7056 2.5421  6.6466

Coefficients:
                      Estimate Std. Error  z value
(Intercept):1          -0.11288    0.13519 -0.83493
(Intercept):2           2.17974    0.15959 13.65854
Bildung_Bachelor        1.47464    0.19077  7.72985
Bildung_Graduate        1.94391    0.25873  7.51321
Bildung_High school     0.64975    0.15041  4.31984
Bildung_Junior college  1.06571    0.21687  4.91396

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: 2656.409 on 24 degrees of freedom

Log-likelihood: -1328.205 on 24 degrees of freedom

Number of iterations: 4
```

Hier wurde der Zusammenhang zwischen Einschätzung und Bildung genauso modelliert wie in `polr`. In `vglm` hat man auch die Möglichkeit das Modell mittels einer Matrix und eines Vektors zu beschreiben. Dabei enthält die Matrix die Gewichte und der Vektor alle Stufen der erklärenden Variable. Es wird dabei davon ausgegangen, dass die erste Spalte in der Matrix der niedrigsten Stufe entspricht, die zweite der nächsthöheren usw. Wir sehen, dass auf diese Art und Weise dieselben Schätzer herauskommen.

```
wenigerHighschool <- c(98,84,23)
Highschool <- c(574,286,50)
JunCollege <- c(122,44,4)
Bachelor <- c(268,57,11)
Graduate <- c(148,23,1)
Matrix <- rbind(wenigerHighschool,Highschool,JunCollege,Bachelor,Graduate)
Bildung <-factor(c("<High school","High school",
"Junior college","Bachelor","Graduate"))

> fit_vglm_proportional <- vglm(Matrix ~ Bildung, family=
                                cumulative(link="logit",parallel=TRUE))
> summary(fit_vglm_proportional)

Call:
vglm(formula = Matrix ~ Bildung, family = cumulative(link = "logit",
    parallel = TRUE))

Pearson residuals:
                  logit(P[Y<=1]) logit(P[Y<=2])
wenigerHighschool       0.293607       -0.54248
Highschool             -0.045029        0.11130
JunCollege             -0.351374        0.99646
Bachelor                0.289290       -0.90800
Graduate               -0.319944        1.09080

Coefficients:
                     Estimate Std. Error  z value
(Intercept):1         -0.11288    0.13519 -0.83492
(Intercept):2          2.17974    0.15959 13.65841
BildungBachelor        1.47464    0.19077  7.72981
BildungGraduate        1.94391    0.25873  7.51334
```

```
BildungHigh school     0.64975    0.15041  4.31983
BildungJunior college  1.06571    0.21687  4.91396

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: 4.12754 on 4 degrees of freedom

Log-likelihood: -26.36407 on 4 degrees of freedom

Number of iterations: 4
> fitted(fit_vglm_proportional)
                         mu1       mu2        mu3
wenigerHighschool 0.4718109 0.4266042 0.10158493
Highschool        0.6310843 0.3131641 0.05575153
JunCollege        0.7216841 0.2408249 0.03749095
Bachelor          0.7960470 0.1787282 0.02522474
Graduate          0.8618846 0.1221878 0.01592762
```

Hier wird das Modell mit positiven Vorzeichen der Parameter verwendet. Dementsprechend haben die β-Parameter hier ein umgekehrtes Vorzeichen. Ansonsten ist im Vergleich zu den Schätzern von `polr` kein Unterschied zu erkennen. Die Standard Errors unterscheiden sich nur leicht.

Die Schätzer für das Modell mit kategorieabhängigen Parametern ergeben sich wie folgt.

```
> fit_vglm <- vglm(Matrix ~ Bildung, family=cumulative(link="logit",parallel=FALSE))
Warnmeldung:
In eval(expr, envir, enclos) :
  iterations terminated because half-step sizes are very small
> summary(fit_vglm)

Call:
vglm(formula = Matrix ~ Bildung, family = cumulative(link = "logit",
    parallel = FALSE))
```

```
Coefficients:
                          Estimate Std. Error  z value
(Intercept):1            -0.087861    0.13982 -0.62839
(Intercept):2             2.068512    0.22130  9.34718
BildungBachelor:1         1.459341    0.19490  7.48753
BildungBachelor:2         1.317417    0.37810  3.48432
BildungGraduate:1         1.907020    0.26072  7.31452
BildungGraduate:2         3.073151    1.02704  2.99223
BildungHigh school:1      0.623380    0.15578  4.00160
BildungHigh school:2      0.776397    0.26483  2.93167
BildungJunior college:1   1.020681    0.22041  4.63086
BildungJunior college:2   1.657181    0.55227  3.00070

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: -4.5325e-14 on 0 degrees of freedom

Log-likelihood: -24.3003 on 0 degrees of freedom

Number of iterations: 5
> fitted(fit_vglm)
                         mu1       mu2         mu3
wenigerHighschool 0.4780488 0.4097561 0.112195122
Highschool        0.6307692 0.3142857 0.054945055
JunCollege        0.7176471 0.2588235 0.023529412
Bachelor          0.7976190 0.1696429 0.032738095
Graduate          0.8604651 0.1337209 0.005813953
```

Es handelt sich hier um das saturierte Modell. Die Anzahl der Parameter entspricht der Anzahl der Zellen. Somit ist die Anzahl der Freiheitsgrade gleich 0. Wir sehen, dass jede Stufe der erklärenden Variable Bildung zwei Schätzer erhält. Einen zum Schätzen für $P(Y \leq 1 \mid \boldsymbol{x})$ und einen für $P(Y \leq 2 \mid \boldsymbol{x})$, wobei 1 für „sehr“, 2 für „mehr oder

weniger“ und 3 für „gar nicht“ steht. Bei den geschätzten Punktwahrscheinlichkeiten stellt man Unterschiede in einer Größenordnung kleiner als 0.02 im Vergleich zum Modell mit proportionalen Odds fest.

3.8.3. Analyse mit `lrm`

Mittels der Funktion `lrm` ergeben sich ebenfalls die gleichen Schätzwerte mit demselben Vorzeichen wie `polr`, sodass auch hier die Wahrscheinlichkeiten übereinstimmen. Die Warnmeldung soll uns hier nicht stören, da wir weder Model Validation noch Bootsstraping betreiben wollen.

```
library(rms)
> fit_lrm <- lrm(ordered(Einschaetzung) ~ Bildung_, weights=Gewichte)
Warnmeldung:
In lrm(ordered(Einschaetzung) ~ Bildung_, weights = Gewichte) :
  currently weights are ignored in model validation and bootstrapping lrm fits
> fit_lrm

Logistic Regression Model

lrm(formula = ordered(Einschaetzung) ~ Bildung_, weights = Gewichte)

Sum of Weights by Response Category

   0    1    2
1210  494   89

                        Model Likelihood     Discrimination    Rank Discrim.
                           Ratio Test            Indexes          Indexes
Obs               15    LR chi2     103.48    R2       0.071    C       0.627
 gar nicht         5    d.f.             4    g        0.808    Dxy     0.254
 mehr oder weniger 5    Pr(> chi2) <0.0001    gr       2.243    gamma   0.380
 sehr              5                          gp       0.162    tau-a   0.119
Sum of weights  1793                          Brier    0.208
max |deriv|    2e-07

                      Coef    S.E.   Wald Z Pr(>|Z|)
y>=mehr oder weniger  0.1129 0.1360   0.83 0.4065
```

```
y>=sehr                 -2.1797 0.1603 -13.60 <0.0001
Bildung_=Bachelor       -1.4746 0.1917  -7.69 <0.0001
Bildung_=Graduate       -1.9439 0.2582  -7.53 <0.0001
Bildung_=High school    -0.6497 0.1511  -4.30 <0.0001
Bildung_=Junior college -1.0657 0.2164  -4.92 <0.0001
```

3.8.4. Analyse mit `clm`

Auch `clm` liefert dieselben Schätzer wie `polr`. Es wir hier auch vom selben Modell ausgegangen, sodass die Vorzeichen ebenfalls übereinstimmen.

```
> library(ordinal)
> fit_clm <- clm(Einschaetzung ~ Bildung_, weights=Gewichte)
> summary(fit_clm)
formula: Einschaetzung ~ Bildung_

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  1793 -1328.20 2668.41 6(0)  5.68e-14 7.1e+01

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
Bildung_Bachelor       -1.4746     0.1917  -7.692 1.45e-14 ***
Bildung_Graduate       -1.9439     0.2582  -7.529 5.11e-14 ***
Bildung_High school    -0.6497     0.1511  -4.300 1.71e-05 ***
Bildung_Junior college -1.0657     0.2164  -4.924 8.46e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
                           Estimate Std. Error z value
gar nicht|mehr oder weniger  -0.1129     0.1360   -0.83
mehr oder weniger|sehr        2.1797     0.1603   13.60
```

3.9. Alternative Link-Funktionen

In 3.1 haben wir beim Schwellenwertansatz eine latente Zufallsvariable Z betrachtet, welche dem linearen Regressionsmodell

$$Z = \alpha^* + \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon$$

genügt, wobei α^* der Intercept und ϵ eine Zufallsvariable mit Verteilungsfunktion F ist. In 3.2 haben wir angenommen, dass F der logistischen Verteilungsfunktion entspricht, deren Dichte in Abbildung 3.4 dargestellt ist.

Wir wollen in diesem Abschnitt einige mögliche Alternativen aufzeigen. Beginnen wollen wir dabei mit der Normalverteilung, welche zum sogenannten Probit-Modell führt. Anschließend werden mit der Gompertz- und der Gumbel-Verteilung die kumulativen Extremwert-Modelle besprochen.

Probit-Modell

Beim Probit-Modell geht man davon aus, dass der Störterm ϵ sich gemäß einer Standardnormalverteilung verhält. Für die bedingte Wahrscheinlichkeit der ordinalen Response Y ergibt sich somit folgendes

$$P(Y \leq j \mid \boldsymbol{x}) = \Phi(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}).$$

Der Unterschied zum kumulativen Logit-Modell ist wie man auch in Abbildung 3.4 erkennen kann relativ gering. Allerdings lassen sich die Parameter hier nicht so leicht interpretieren wie beim Logit-Modell.

Kumulative Extremwert-Modelle

Weitere mögliche Verteilungen, die man für den Störterm annehmen kann, wären die Gompertz- und die Gumbel-Verteilung, welche in Abbildung 3.5 dargestellt sind.

Letztere entspricht einem kumulativen Maximum-Extremwert-Modell, Gompertz impliziert ein kumulatives Minimum-Extremwert-Modell, welches durch folgende Verteilungsfunktion beschrieben wird

$$F(\eta) = 1 - \exp(-\exp(\eta)).$$

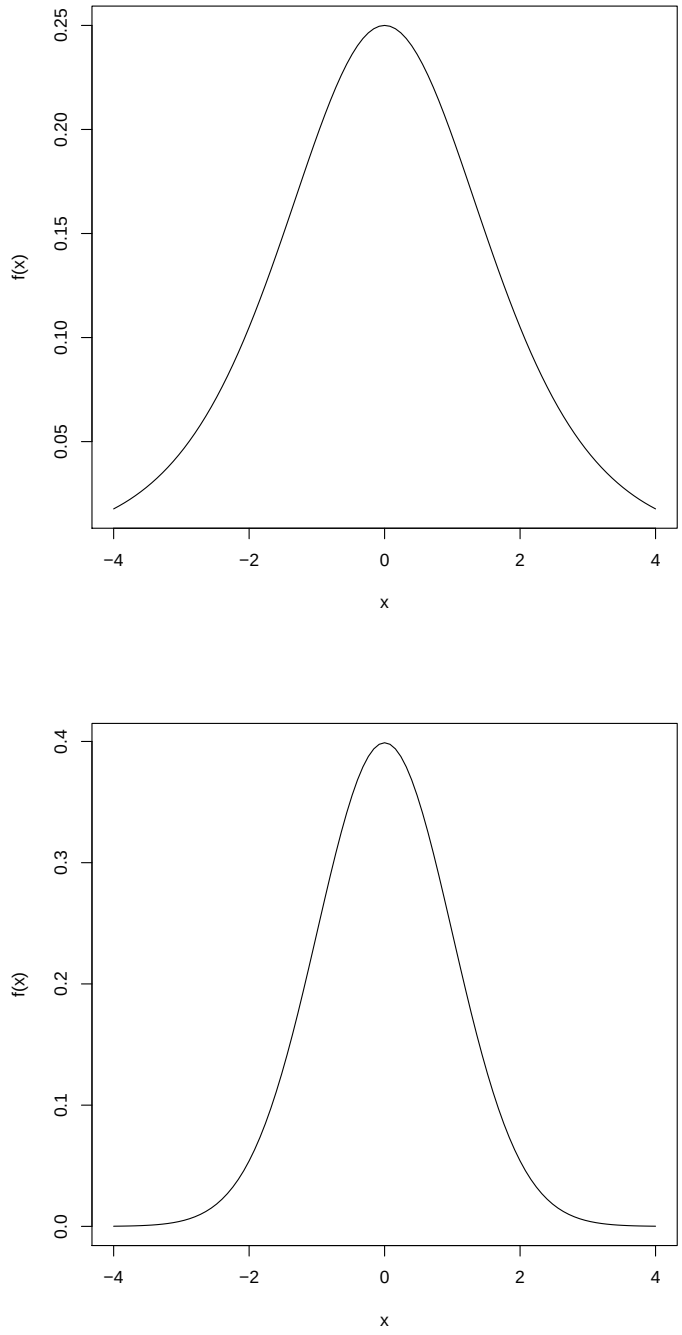

Abbildung 3.4.: Dichtefunktion der logistischen Verteilung (oben) und der Normalverteilung (unten)

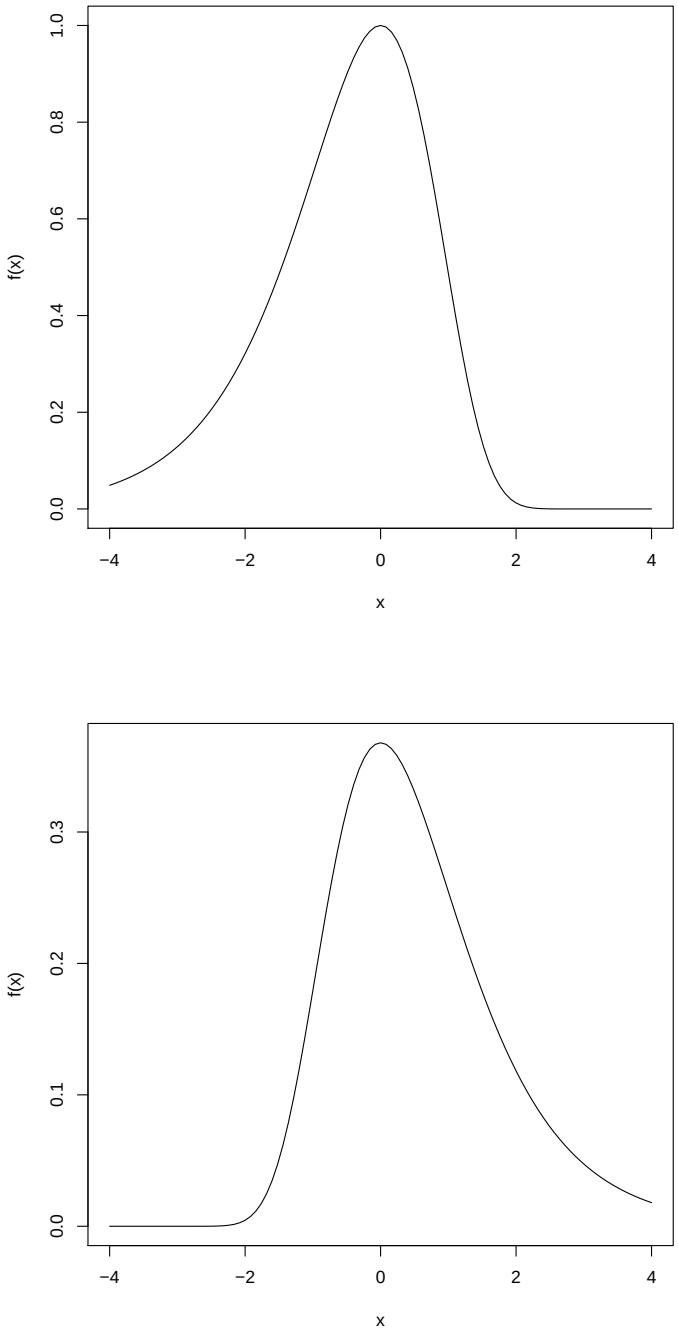

Abbildung 3.5.: Dichtefunktion der Gompertz- (oben) und Gumbelverteilung(unten)

Dies führt uns zu

$$P(Y \leq j \mid \boldsymbol{x}) = 1 - \exp(-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}))$$

und mit

$$\exp(-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})) = 1 - P(Y \leq j \mid \boldsymbol{x}) = P(Y \geq j+1 \mid \boldsymbol{x}) \tag{3.4}$$

folgt alternativ die Darstellung

$$\log(-\log(P(Y > j \mid \boldsymbol{x})) = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

Um den Zusammenhang zu sequentiellen Modellen darzustellen, wollen wir die bedingte Wahrscheinlichkeit $P(Y = j \mid Y \geq j, \boldsymbol{x})$ mit Hilfe von Ergebnis (3.4) ausdrücken. Nach dem Satz von Bayes gilt

$$\begin{aligned}
P(Y = j \mid Y \geq j, \boldsymbol{x}) &= \frac{P(Y = j, Y \geq j \mid \boldsymbol{x})}{P(Y \geq j \mid \boldsymbol{x})} = \frac{P(Y = j \mid \boldsymbol{x})}{P(Y \geq j \mid \boldsymbol{x})} \\
&= \frac{P(Y \geq j \mid \boldsymbol{x}) - P(Y \geq j+1)}{P(Y \geq j \mid \boldsymbol{x})} \\
&= 1 - \frac{P(Y \geq j+1 \mid \boldsymbol{x})}{P(Y \geq j \mid \boldsymbol{x})} \\
&\overset{(3.4)}{=} 1 - \frac{\exp(-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}))}{\exp(-\exp(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}))} \\
&= 1 - \exp(-\{\exp(\alpha_j)\exp(-\boldsymbol{\beta}^T \boldsymbol{x}) - \exp(\alpha_{j-1})\exp(-\boldsymbol{\beta}^T \boldsymbol{x})\}).
\end{aligned}$$

Mittels der Transformation $\exp(\tilde{\alpha}_j) = \exp(\alpha_j - \exp \alpha_{j-1})$ gelangen wir damit zur alternativen Modellbeschreibung

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = 1 - \exp(-\exp(\tilde{\alpha}_j - \boldsymbol{\beta}^T \boldsymbol{x})). \tag{3.5}$$

Durch diese Eigenschaft wird der enge Zusammenhang zu sequentiellen Modellen deutlich, welche wir im Kapitel 4 betrachten. Während die Parameter α_j monoton wachsend sein müssen, gelten für die Parameter $\tilde{\alpha}_j$ keine Restriktionen. Definiert man

$$\lambda(j, \boldsymbol{x}) = P(Y = j \mid Y \geq j, \boldsymbol{x})$$

als diskretes Risiko (Hazard-Rate) so kann Gleichung (3.5) auch als diskretes Cox- oder Hazard-Modell betrachtet werden. Dieses beschreibt einen Prozess, der in Kategorie j stoppt, gegeben, dass er Kategorie j erreicht hat. Der Cox-Prozess wird daher häufig für die Modellierung von Überlebenszeiten verwendet. Die Eigenschaft der strikten stochastischen Ordnung kann hier wie folgt dargestellt werden

$$\frac{\log P(Y > j \mid \boldsymbol{x}_1)}{\log P(Y > j \mid \boldsymbol{x}_2)} \overset{3.4}{=} \frac{-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x_1})}{-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x_2})}$$
$$= -\exp(-\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)).$$

Dieser Ausdruck ist unabhängig von der Kategorie j.

Mittels der Transformation $Y_{trans} = c + 1 - Y$ gelangen wir zum Gumbel- oder kumulativen Maximum-Extremwert-Modell, welches auf der Gumbel-Verteilung

$$F(\eta) = \exp(-\exp(\eta))$$

basiert. Das Modell ergibt sich daher zu

$$P(Y \leq j \mid \boldsymbol{x}) = \exp(-\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})) \quad \text{bzw.}$$
$$\log(-\log(P(Y \leq j \mid \boldsymbol{x})) = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

3.9.1. Vergleich anhand eines Beispiels

Wir betrachten erneut die Daten aus Beispiel 4, welche hier noch einmal in Tabelle 3.4 dargestellt sind. Wir wollen die Linkfunktion variieren und auftretende Unterschiede diskutieren. Durch Wahl der Normalverteilung ergeben sich folgende Schätzer und Wahrscheinlichkeiten.

```
> fit_probit<- polr(Einschaetzung ~ Bildung_, weights=Gewichte, method="probit")
> summary(fit_probit)

Re-fitting to get Hessian

Call:
polr(formula = Einschaetzung ~ Bildung_, weights = Gewichte,
    method = "probit")
```

	Astrologie ist wissenschaftlich		
Schulbildung	gar nicht	mehr oder weniger	sehr
<High school	98 (48%)	84 (41%)	23 (11%)
High school	574 (63%)	286 (31%)	50 (5%)
Junior College	122 (72%)	44 (26%)	4 (2%)
Bachelor	268 (80%)	57 (17%)	11 (3%)
Graduate	148 (86%)	23 (13%)	1 (1%)

Tabelle 3.4.: Meinung über die Wissenschaftlichkeit von Astrologie in Abhängigkeit von der Bildung

```
Coefficients:
                         Value Std. Error t value
Bildung_Bachelor       -0.8500    0.11037  -7.701
Bildung_Graduate       -1.1417    0.14276  -7.998
Bildung_High school    -0.3855    0.08975  -4.295
Bildung_Junior college -0.6444    0.12756  -5.052

Intercepts:
    Value   Std. Error t value
1|2 -0.0480  0.0815    -0.5894
2|3  1.1985  0.0877    13.6675

Residual Deviance: 2656.466
AIC: 2668.466
> fitted(fit_probit)
   gar nicht mehr oder weniger        sehr
1  0.4808496         0.4037927 0.115357708
2  0.4808496         0.4037927 0.115357708
3  0.4808496         0.4037927 0.115357708
4  0.6321069         0.3112937 0.056599344
5  0.6321069         0.3112937 0.056599344
6  0.6321069         0.3112937 0.056599344
7  0.7245504         0.2427816 0.032667956
8  0.7245504         0.2427816 0.032667956
```

```
9  0.7245504          0.2427816 0.032667956
10 0.7887180          0.1910275 0.020254531
11 0.7887180          0.1910275 0.020254531
12 0.7887180          0.1910275 0.020254531
13 0.8629615          0.1274033 0.009635125
14 0.8629615          0.1274033 0.009635125
15 0.8629615          0.1274033 0.009635125
```

Wählen wir die Gumbelverteilung so ergeben sich folgende Parameterschätzer und Wahrscheinlichkeiten.

```
> fit_gumbel<- polr(Einschaetzung  ~ Bildung_, weights=Gewichte, method="cloglog")
> summary(fit_gumbel)

Re-fitting to get Hessian

Call:
polr(formula = Einschaetzung ~ Bildung_, weights = Gewichte,
    method = "cloglog")

Coefficients:
                         Value Std. Error t value
Bildung_Bachelor       -1.1961     0.1561  -7.661
Bildung_Graduate       -1.6140     0.2266  -7.122
Bildung_High school    -0.4875     0.1124  -4.338
Bildung_Junior college -0.8250     0.1749  -4.717

Intercepts:
    Value   Std. Error t value
1|2  0.2880  0.0982     2.9332
2|3  2.3619  0.1366    17.2856

Residual Deviance: 2658.045
AIC: 2670.045
> fitted(fit_gumbel)
   gar nicht mehr oder weniger       sehr
1  0.4724965          0.4375674 0.08993606
2  0.4724965          0.4375674 0.08993606
```

```
3  0.4724965          0.4375674 0.08993606
4  0.6309996          0.3127649 0.05623546
5  0.6309996          0.3127649 0.05623546
6  0.6309996          0.3127649 0.05623546
7  0.7199651          0.2395770 0.04045787
8  0.7199651          0.2395770 0.04045787
9  0.7199651          0.2395770 0.04045787
10 0.7971547          0.1747505 0.02809480
11 0.7971547          0.1747505 0.02809480
12 0.7971547          0.1747505 0.02809480
13 0.8613426          0.1200699 0.01858743
14 0.8613426          0.1200699 0.01858743
15 0.8613426          0.1200699 0.01858743
```

Um die verschiedenen Werte für die Wahrscheinlichkeiten besser vergleichen zu können, wollen wir die Ergebnisse tabellarisch zusammenfassen. Tabelle 3.5 zeigt, dass die Werte nur sehr gering voneinander abweichen. Insgesamt liefern alle drei Verteilungen ähnliche Ergebnisse. Berechnungen mit der Funktion `vglm` liefern dieselben Ergebnisse.

3.10. Datenbeispiel: Studie über die Wirkung verschiedener Verbandsmaterialien

In diesem Abschnitt wollen wir die Erkenntnisse über kumulative Logit-Modelle auf eine aktuelle Studie der Medizinischen Universität Graz anwenden.

Es sollen durch Untersuchungen von insgesamt 63 Patienten die Verwendbarkeit von drei verschiedenen Verbandsmaterialien bei „Schürfwunden“ (Spalthautentnahmestellen) am Oberschenkel beurteilt werden. Dabei gelten folgende Abkürzungen

- S – Suprathel (künstlicher Hautersatz),
- M - MepilexAg (enthält Silberionen mit antiinfektiöser Wirkung) und
- B - Biatian Ibu (enthält Schmerzmittel; NSAD- non steroidal anti-inflammatoric drugs).

	Verteilung		
	Logistische	Normal	Gumbel
$P(Y = \text{sehr} \mid <\text{High school})$	10.2	11.5	9.0
$P(Y = \text{mehr oder weniger} \mid <\text{High school})$	42.7	40.4	43.8
$P(Y = \text{gar nicht} \mid <\text{High school})$	47.2	48.1	47.2
$P(Y = \text{sehr} \mid \text{High school})$	5.6	5.7	5.6
$P(Y = \text{mehr oder weniger} \mid \text{High school})$	31.3	31.1	31.3
$P(Y = \text{gar nicht} \mid \text{High school})$	63.1	63.2	63.1
$P(Y = \text{sehr} \mid \text{Junior college})$	3.7	3.3	4.0
$P(Y = \text{mehr oder weniger} \mid \text{Junior college})$	24.1	24.3	24.0
$P(Y = \text{gar nicht} \mid \text{Junior college})$	72.2	72.5	72.0
$P(Y = \text{sehr} \mid \text{Bachelor})$	2.5	2.0	2.8
$P(Y = \text{mehr oder weniger} \mid \text{Bachelor})$	17.9	19.1	17.5
$P(Y = \text{gar nicht} \mid \text{Bachelor})$	79.6	78.9	79.7
$P(Y = \text{sehr} \mid \text{Graduate})$	1.6	1.0	1.9
$P(Y = \text{mehr oder weniger} \mid \text{Graduate})$	12.2	12.7	12.0
$P(Y = \text{gar nicht} \mid \text{Graduate})$	86.2	86.3	86.1

Tabelle 3.5.: Wahrscheinlichkeiten der einzelnen Antwortmöglichkeiten in Prozent (gerundet)

Die beiden letzteren sind selbstklebende Polyurethanschaumstoffverbände. Bei manchen Patienten wurden nicht nur einer sondern zwei oder drei der Verbandsmaterialien getestet. In diesen Fällen gibt es für den entsprechenden Patient dann mehrere Datensätze und er fließt dann doppelt bzw. dreifach mit ein.

Von den Personen wurden zunächst verschiedene Daten ermittelt wie Alter, Geschlecht, Raucher, Diabetiker usw. Beim Alter wird lediglich zwischen jung (unter 55 Jahre alt) und alt (über 55 Jahre alt) unterschieden. Der ASA-Wert gibt eine von Anästhesisten vorgenommene Klassifikation zur Abschätzung des perioperativen Risikos an (1 bis 6, 1=gesunder Patient, 6=verstorben).

Nachdem den Patienten der Verband aufgetragen wurde, sind sie innerhalb von zwei

Wochen täglich nach ihrem Empfinden (Schmerz der Wunde) gefragt worden. Jeweils in Ruhe (VASR) und in Bewegung (VASB), wobei VAS für Visuelle Analog Skala zur Schmerzbestimmung steht (0 bis 10, 0 = keine Schmerzen, 10 = größtmögliche vorstellbare Schmerzen). Teilweise haben die Patienten während der Zeit Schmerzmittel verabreicht bekommen (Schmerzmittel: ja/nein; bei starken Schmerzen gemäß der VAS-Skala wurde Schmerzmittel verabreicht). Bei manchen wurde ein Verbandswechsel durchgeführt und der Schmerzwert erfragt (VW1: erster (von zwei) Verbandswechseln, 0=keine Schmerzen; 1=mittelmäßig; 2=sehr schmerzhaft), der 2. Verbandswechsel erfolgte am Ende der 10 bis 14 Tage. Einige Verbände sind schon vor dem 14. Tag abgenommen worden (Vab entspricht dem Tag der Abnahme).

Es geht darum einen Zusammenhang zwischen dem Schmerzwert und somit der Verheildauer der Wunde und dem verwendeten Verbandsmaterial zu untersuchen. Dabei sollen die ermittelten Größen (Alter etc.) bei Relevanz in das statistische Modell mit einfließen.

Wir wollen die ordinale Responsevariable Schmerzwert in Bewegung (VASB) in Abhängigkeit von den gegebenen Größen Verband, Alter, Geschlecht, Raucher, Tumor, Diabetiker, Allergiker, Hypertonus, vasc. Erkrankungen, ASA und Tag beschreiben, wobei der Tag linear und quadratisch einfließt. Der kubische Einfluss der Zeit ist nicht signifikant. Da es sich um eine ordinale Skala handelt, nehmen wir ein kumulatives Logit-Modell an und schätzen die Parameter mit den R-Funktionen `polr` und `vglm`. Durch eine schrittweise Variablenselektion gelangen wir zu einem Modell, welches die Parameter Verband, Alter, Raucher, Diabetiker, Hypertonus, ASA und Tag sowie Tag^2 enthält und erhalten die folgenden Parameterschätzer, welche für beide Funktionen übereinstimmen.

```
> verband1.mod.opt<-polr(as.ordered(VASB) ~ verband+alter+nikotin+diab
                         +hypertonus+asa+tag+I(tag^2), Hess=T)
> summary(verband1.mod.opt)

Coefficients:
                  Value Std. Error t value
verbandM        0.48726   0.163073   2.988
verbandS        1.15185   0.186999   6.160
alterjung      -0.43206   0.239298  -1.806
nikotinnein    -0.28667   0.154000  -1.861
diabnein       -0.83937   0.181789  -4.617
```

```
hypertonusnein  0.65381   0.171444   3.814
asa2           -0.97538   0.262591  -3.714
asa3           -0.95592   0.317096  -3.015
asa4           -1.59332   0.321207  -4.960
tag             0.13286   0.074571   1.782
I(tag^2)       -0.01202   0.005056  -2.377

Intercepts:
    Value   Std. Error t value
0|1 -0.9058  0.4382    -2.0669
1|2  0.2208  0.4361     0.5063
2|3  1.6271  0.4439     3.6656
3|4  2.4671  0.4611     5.3502
4|5  3.7832  0.5402     7.0039

Residual Deviance: 1967.162
AIC: 1999.162
```

Für die Interpretation der Parameterschätzer ist wichtig zu wissen, dass durch die Vergrößerung des linearen Prädiktors die Wahrscheinlichkeit für eine größere Stufe (stärkerer Schmerz) steigt. Somit tendieren alle Personen, die durch Faktorstufen mit großen, positiven Parameterwerten charakterisiert sind, zu stärkeren Schmerzen. Im Vergleich dazu weisen negative Parameterwerte auf geringere Schmerzen hin. Nicht angeführte Faktorstufen charakterisieren die Referenzklasse von Personen und sind als Nullen zu interpretieren. Deren Angaben werden durch die Intercepts beschrieben. Dazu zählen in diesem Fall ältere Patienten mit Hochdruck, die Raucher und Diabetiker sind, denen ein Verband B angelegt wurde und jene, die von Anästhesisten als gesund eingestuft wurden (ASA=1).

Konkret kann man sagen, dass beim Verband Biatian Ibu die geringsten Schmerzen aufgetreten sind. Es folgt MepilexAG, bei Suprathel sind die Schmerzen am stärksten. Junge Personen haben eine Tendenz zu geringeren Schmerzen als ältere Personen. Leidet eine Person an Diabetes oder ist diese ein Raucher, so empfindet sie stärkere Schmerzen, als wenn das nicht der Fall ist. Bei an Hypertonie erkrankten Personen ist dies umgekehrt. Eine Person mit Hochdruck empfindet weniger Schmerzen, als eine Person ohne diese Krankheit. Für den ASA-Faktor kann man in etwa sagen, dass eine als eher krank

eingestufte Person (4) weniger Schmerzen empfindet, als eine als gesund eingestufte Person (1). Setzt man die Schätzer für den Zeitverlauf (Tag) in die Modellformel ein, so erkennt man, dass die Schmerzen der Patienten im Laufe der Zeit geringer werden.

Diese Aussagen können wir mithilfe der Odds Ratios für die Faktorvariablen konkretisieren.

```
> matrix(exp(-coef(verband1.mod.opt)), ncol=1,
         dimnames=list(names(coef(verband1.mod.opt)),"odds ratio" ))
                odds ratio
verbandM         0.6143099
verbandS         0.3160527
alterjung        1.5404328
nikotinnein      1.3319780
diabnein         2.3149163
hypertonusnein   0.5200607
asa2             2.6521862
asa3             2.6010529
asa4             4.9200786
```

Es wird stets ein Vergleich zur Referenzklasse gemacht. So sehen wir, dass beim MepilexAG-Verband eine 0.6 mal so große Chance (Quote) besteht, eher geringere statt stärkere Schmerzen zu haben als beim Biatian Ibu Verband. Beim Suprathel ist sie sogar nur 0.32 mal so groß. Zum Alter kann man sagen, dass jüngere Personen eine 1.5 mal so große Chance haben eher geringere statt stärkere Schmerzen zu haben wie ältere Personen. Für einen Nichtraucher ist die Quote 1.3 mal so groß wie bei einem Raucher, bei einer Person, die nicht an Diabetes leidet 2.3 mal so groß wie bei einer an Diabetes erkrankten Person. Bei Personen, die an Hypertonie leiden, ist die Chance für geringere statt stärkere Schmerzen etwa doppelt so groß wie bei Personen ohne Hypertonie. Die Chance steigt mit wachsendem ASA-Faktor von einer 2.6 bei eher gesunden Personen bis hin zu einer fast 5 mal so großen Chance bei Personen, die schon als stark erkrankt eingestuft wurden.

Um nun einen Vergleich zwischen den drei Verbänden zu machen, betrachten wir die Veränderung der Wahrscheinlichkeit $P(Y \leq j \mid x)$ im Laufe der Zeit für die verschiedenen Verbände. Dabei gehen wir ohne Beschränkung der Allgemeinheit von einer älteren Person aus, die Raucher und Diabetiker ist, an Hypertonie erkrankt ist und als gesund

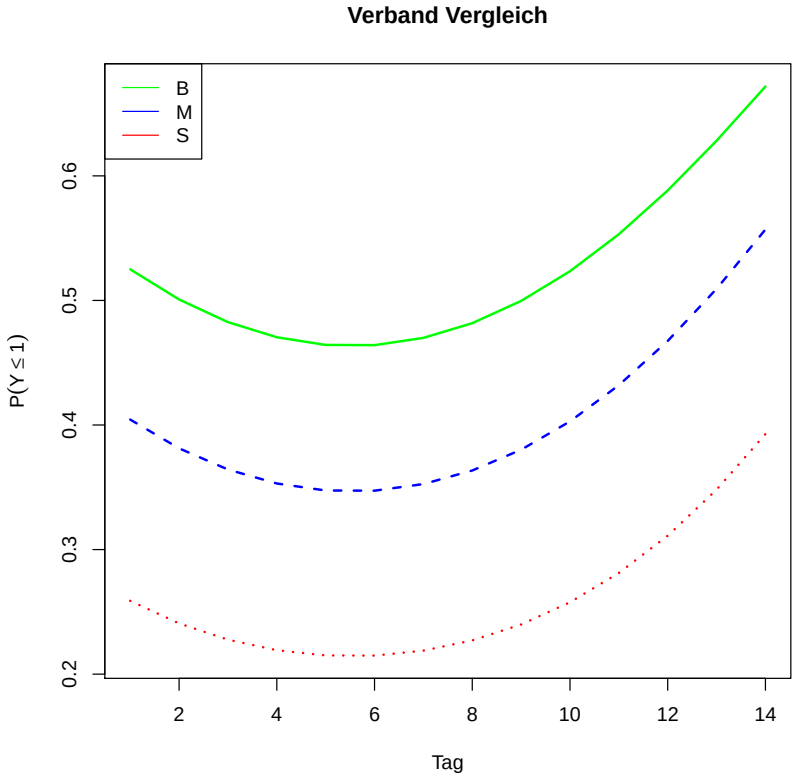

Abbildung 3.6.: Die kumulative Wahrscheinlichkeit $P(Y \leq 1 \mid \text{Tag})$ für verschiedene Verbände in Abhängigkeit vom zeitlichen Verlauf.

(ASA=1) eingestuft wurde. Dabei wählen wir $j = 1$, das heißt wir betrachten die Wahrscheinlichkeit für geringe Schmerzen und müssen den $1 \mid 2$ Intercept in die Berechnung mit einbeziehen. Eine Erhöhung von j bewirkt Verschiebungen nach oben. In Abbildung 3.6 ist zu erkennen, dass die Wahrscheinlichkeit für geringe Schmerzen für den Biatian Ibu Verband am höchsten ist, gefolgt von MepilexAG. Bei Suprathel ist die Wahrscheinlichkeit am geringsten während des gesamten Zeitverlaufs. Bei allen drei Verbandsarten fällt die Wahrscheinlichkeit bis etwa zum 5.-6. Tag und steigt danach wieder an. Dies bedeutet, dass die Schmerzen zunächst etwas zunehmen, aber dann weniger werden.

In einer weiteren interessanten Analyse wird versucht Interaktionen mit einzubeziehen um einen Zusammenhang zwischen dem Schmerzverlauf (Tag) und des Verbandsmaterials herzustellen.

```
> verband2.mod<-polr(as.ordered(VASB) ~  verband+alter+nikotin+diab
                     +hypertonus+asa+tag+I(tag^2)+verband:tag, Hess=T)
> summary(verband2.mod)

Coefficients:
                   Value Std. Error t value
verbandM       -0.227080   0.343392 -0.6613
verbandS        1.185314   0.364869  3.2486
alterjung      -0.445070   0.239885 -1.8553
nikotinnein    -0.279288   0.154542 -1.8072
diabnein       -0.846098   0.182493 -4.6363
hypertonusnein  0.648693   0.172277  3.7654
asa2           -0.996058   0.263863 -3.7749
asa3           -0.979289   0.317663 -3.0828
asa4           -1.646607   0.322510 -5.1056
tag             0.087681   0.079048  1.1092
I(tag^2)       -0.011235   0.005083 -2.2103
verbandM:tag    0.103032   0.043202  2.3849
verbandS:tag   -0.005209   0.045123 -0.1154

Intercepts:
    Value   Std. Error t value
0|1 -1.2088  0.4718    -2.5619
1|2 -0.0768  0.4695    -0.1635
2|3  1.3299  0.4767     2.7901
3|4  2.1698  0.4928     4.4025
4|5  3.4861  0.5677     6.1412

Residual Deviance: 1959.502
AIC: 1995.502
```

Der Großteil der Schätzer ist ähnlich wie zuvor. Lediglich die Schätzer vom Verband haben sich geändert, da für diesen die Interaktion ins Spiel gekommen ist. Wir erhalten damit den in Abbildung 3.7 abgebildeten Wahrscheinlichkeitsverlauf für die verschiedenen Verbandsmaterialien. Ein Likelihood-Quotienten-Test, der die beiden Modelle miteinander vergleicht liefert folgendes.

```
> anova(verband2.mod,verband1.mod.opt)
```

```
Likelihood ratio tests of ordinal regression models

Response: as.ordered(VASB)
                                                                                Model
1               verband + alter + nikotin + diab + hypertonus + asa + tag + I(tag^2)
2 verband + alter + nikotin + diab + hypertonus + asa + tag + I(tag^2) + verband:tag
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1       799   1967.162
2       797   1959.502 1 vs 2     2 7.660628 0.0217028
```

Das Modell mit Interaktion hat eine signifikant geringere Deviance als das Modell ohne Interaktionen bei einem Verlust von zwei Freiheitsgraden. Der p-Wert von 0.02 weist darauf hin, dass eher das Modell mit Interaktionen verwendet werden sollte.

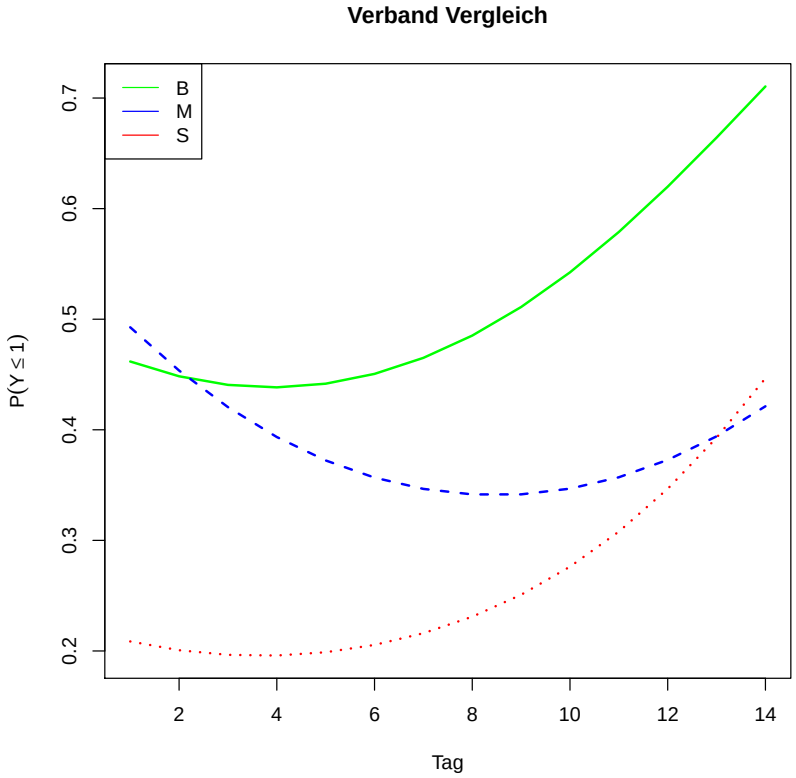

Abbildung 3.7.: Die kumulative Wahrscheinlichkeit $P(Y \leq 1 \mid \text{Tag})$ für verschiedene Verbände in Abhängigkeit vom zeitlichen Verlauf (mit Interaktionen).

4. Sequentielle Modelle

Es gibt ordinale Daten, bei denen das Eintreten der Stufe j voraussetzt, dass vorher die darunterliegenden Stufen durchlaufen worden sind. Dies ist oft bei Zeitangaben der Fall. Werden beispielsweise Personen gefragt, wie lange sie in ihrem Job tätig sind, wobei sie dabei mit „weniger als 1 Jahr“, „zwischen 1 und 5 Jahren“, „zwischen 5 und 10 Jahren“ und „länger als 10 Jahre“ antworten können, so kann jemand nur mit „zwischen 1 und 5 Jahren“ antworten, wenn er sich zuvor auch schon in dem Zustand befand, dass er weniger als 1 Jahr bei der Arbeitsstelle war. Auch der Dienstgrad im Militär stellt eine Rangordnung dar, bei der ein höherer Grad nur erreicht werden kann, wenn alle darunterliegenden durchlaufen wurden. Für solche Fälle sind sequentielle Modelle gut geeignet, welche wir in den folgenden Abschnitten aufbauend auf Tutz (2012) genauer erläutern wollen.

4.1. Das einfache sequentielle Modell und der Schwellenwertansatz

Man betrachtet latente Variablen $Z_1, Z_2, \ldots, Z_{c-1}$ und Schwellenwerte α_j, $j = 1, \ldots, c$. Mittels der ordinalen Responsevariable Y beschreiben wir einen Prozess, sodass

$$Y = 1, \text{ wenn } Z_1 \leq \alpha_1.$$

Ist $Z_1 > \alpha_1$, dann ist

$$Y = 2 \text{ gegeben, dass } Y \geq 2, \text{ wenn } Z_2 \leq \alpha_2.$$

Allgemein ist

$$Y = j \text{ gegeben, dass } Y \geq j, \text{ wenn } Z_j \leq \alpha_j.$$

Für Z_j wird die folgende Darstellung angenommen

$$Z_j = \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon_j,$$

wobei die ϵ_j unabhängig sind und der Verteilungsfunktion F genügen. Daraus ergibt sich folgender Zusammenhang

$$\begin{aligned} P(Y = j \mid Y \geq j, \boldsymbol{x}) &= P(Z_j \leq \alpha_j) \\ &= P(\boldsymbol{\beta}^T \boldsymbol{x} + \epsilon_j \leq \alpha_j) \\ &= P(\epsilon_j \leq \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}) \\ &= F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}). \end{aligned}$$

Einfaches sequentielles Modell

Im sequentiellen Modell gilt

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}).$$

Nimmt man an, dass F die logistische Verteilungsfunktion ist, so entspricht dies dem Modell, welches wir im nächsten Abschnitt besprechen wollen, und es gilt

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x})} \quad \text{bzw.}$$

$$\operatorname{logit}(P(Y = j \mid Y \geq j, \boldsymbol{x})) = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

Bemerkung 4. *In diesem Modell ist der Parametervektor $\boldsymbol{\beta}$ für alle Stufen gleich. Es gibt jedoch auch Modelle, bei denen man stattdessen Parametervektoren $\boldsymbol{\beta}_j$ betrachtet, die abhängig von der Stufe j sind. Allerdings ergeben sich somit oft Modelle mit zu vielen Parametern.*

Will man einen Zusammenhang zu einem binären Regressionsmodell herstellen, so erreicht man das, in dem man

$$W_j = 1 \Leftrightarrow Y = j \text{ gegeben, dass } Y \geq j,$$

$$W_j = 0 \Leftrightarrow Y > j \text{ gegeben, dass } Y \geq j$$

festlegt. Somit erhält man, dass

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = P(W_j = 1 \mid Y \geq j, \boldsymbol{x}).$$

Folglich ist das obige Modell ein binäres Regressionsmodell für $W_j \in \{0, 1\}$.

4.2. Das sequentielle Logit-Modell

Das sequentielle Logit-Modell verwendet als Verteilungsfunktion die Logistische Verteilungsfunktion, welche wie im kumulativen Modell zu den am häufigsten verwendeten Varianten gehört.

Definition 4.2.1 (Sequentieller Logit, sequential logit)**.** *Für c verschiedene Kategorien mit Wahrscheinlichkeiten $\pi_1, \ldots, \pi_c$ sind die sequentiellen Logits über*

$$\log\left(\frac{P\left(Y = j\right)}{P\left(Y > j\right)}\right) = \log\left(\frac{\pi_j}{\pi_{j+1} + \cdots + \pi_c}\right), \quad j = 1, \ldots, c-1$$

definiert. Alternativ kann man auch

$$\log\left(\frac{P\left(Y = j\right)}{P\left(Y < j\right)}\right) = \log\left(\frac{\pi_j}{\pi_1 + \cdots + \pi_{j-1}}\right), \quad j = 2, \ldots, c$$

betrachten.

Bemerkung 5. *Oft spricht man statt von einem sequential logit auch von einem continuation ratio logit und dem dazugehörigen continuation ratio logit model.*

Das sequentielle Logit-Modell beschreibt die Wahrscheinlichkeit in Kategorie j zu sein unter der Bedingung in einer Kategorie größer oder gleich j zu sein mittels

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x})}{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x})} \quad \text{bzw.}$$

$$\log \frac{P(Y = j \mid \boldsymbol{x})}{P(Y > j \mid \boldsymbol{x})} = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, c-1.$$

Die Äquivalenz der beiden Modelle erklärt sich über die logistische Funktion und deren Umkehrfunktion, denn es gilt

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x})}{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x})}$$

$$\Leftrightarrow \operatorname{logit}(P(Y = j \mid Y \geq j, \boldsymbol{x})) = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}$$

und mit

$$1 - P(Y = j \mid Y \geq j) = P(Y > j \mid Y \geq j) \tag{4.1}$$

folgt

$$\begin{aligned}
\text{logit}(P(Y = j \mid Y \geq j, \boldsymbol{x}) &= \log \frac{P(Y = j \mid Y \geq j, \boldsymbol{x})}{1 - P(Y = j \mid Y \geq j, \boldsymbol{x})} \\
&\overset{(4.1)}{=} \log \frac{P(Y = j \mid Y \geq j, \boldsymbol{x})}{P(Y > j \mid Y \geq j, \boldsymbol{x})} \\
&= \log \frac{P(Y = j \mid \boldsymbol{x})/P(Y \geq j \mid \boldsymbol{x})}{P(Y > j \mid \boldsymbol{x})/P(Y \geq j \mid \boldsymbol{x})} \\
&= \log \frac{P(Y = j \mid \boldsymbol{x})}{P(Y > j \mid \boldsymbol{x})} \\
&= \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}.
\end{aligned}$$

Im Vergleich zum kumulativen Modell gibt es hier keine Restriktionen an die Parameter.

4.3. Folgerungen und Eigenschaften

Im einfachen sequentiellen Modell gilt

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}), \quad j = 1, \ldots, c-1.$$

Ist $j = c$, was der höchsten Stufe entspricht, so haben wir ein sicheres Ereignis mit $P(Y = c \mid Y \geq c, \boldsymbol{x}) = 1$. Es lässt sich auch die Eintrittswahrscheinlichkeit von $Y = j$ modellieren. Wegen $P(Y \geq 1 \mid \boldsymbol{x}) = 1$ und mit (4.1) gilt

$$\begin{aligned}
P(Y \geq j \mid \boldsymbol{x}) &= \frac{P(Y \geq j \mid \boldsymbol{x})}{P(Y \geq j-1 \mid \boldsymbol{x})} \cdot \frac{P(Y \geq j-1 \mid \boldsymbol{x})}{P(Y \geq j-2 \mid \boldsymbol{x})} \cdots \frac{P(Y \geq 2 \mid \boldsymbol{x})}{P(Y \geq 1 \mid \boldsymbol{x})} \\
&= \frac{P(Y > j-1 \mid \boldsymbol{x})}{P(Y \geq j-1 \mid \boldsymbol{x})} \cdot \frac{P(Y > j-2 \mid \boldsymbol{x})}{P(Y \geq j-2 \mid \boldsymbol{x})} \cdots \frac{P(Y > 1 \mid \boldsymbol{x})}{P(Y \geq 1 \mid \boldsymbol{x})} \\
&= \prod_{k=1}^{j-1} P(Y > k \mid Y \geq k, \boldsymbol{x}) \\
&\overset{(4.1)}{=} \prod_{k=1}^{j-1} \left(1 - F(\alpha_k - \boldsymbol{\beta}^T \boldsymbol{x})\right).
\end{aligned}$$

Mit dem Satz von Bayes und erhält man

$$\begin{aligned} P(Y = j \mid \boldsymbol{x}) &= P(Y = j, Y \geq j \mid \boldsymbol{x}) \\ &= P(Y = j \mid Y \geq j, \boldsymbol{x}) P(Y \geq j \mid \boldsymbol{x}) \\ &= F(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}) \prod_{k=1}^{j-1} \left(1 - F(\alpha_k - \boldsymbol{\beta}^T \boldsymbol{x})\right). \end{aligned} \tag{4.2}$$

Strikte stochastische Ordnung

Beim sequentiellen Modell ergibt sich genauso wie beim kumulativen Modell die strikte stochastische Ordnung. Die kumulativen Odds Ratios sind unabhängig von α, denn es gilt

$$\begin{aligned} \frac{P\left(Y = j|\boldsymbol{x}_1\right) / P\left(Y > j|\boldsymbol{x}_1\right)}{P\left(Y = j|\boldsymbol{x}_2\right) / P\left(Y > j|\boldsymbol{x}_2\right)} &= \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}_1)}{\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}_2)} \\ &= \exp\left(+\boldsymbol{\beta}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right). \end{aligned}$$

Die Chance $Y = j$ bei $\boldsymbol{x} = \boldsymbol{x}_1$ zu erhalten, ist das $\exp\left(-\boldsymbol{\beta}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right)$-fache der Chance bei $\boldsymbol{x} = \boldsymbol{x}_2$ für alle $j = 1, \ldots, c-1$. Durch die Unabhängigkeit vom Intercept sind die Odds Ratios für alle Kategorien gleich und nur von $\boldsymbol{x}$ abhängig.

Zusammenhang zu diskreten Überlebens-Modellen (Survival Time Models)

Bei diskreten Überlebens-Modellen geht es darum, die Wahrscheinlichkeit des Eintritts eines gewissen Ereignisses oder Risikos zu einer bestimmten Zeit zu beschreiben. Dabei kann es sich beispielsweise um den Ausbruch einer Krankheit oder um die Rückkehr in ein Arbeitsverhältnis nach der Arbeitslosigkeit handeln. Die Zeitachse wird in c Intervalle $[t_0, t_1), \ldots, [t_{c-1}, t_c]$ unterteilt. Ist die Zufallsvariable $Y = j$, so ist das Ereignis im Zeitintervall $[t_{j-1}, t_j]$ eingetreten. Weiß man bereits, dass das Ereignis bis zum Zeitpunkt t_{j-1} nicht eingetreten ist (das heißt wir wissen, dass $Y \geq j$ ist), so interessiert man sich für die Wahrscheinlichkeit, dass das Ereignis im j-ten Zeitintervall eintritt, gegeben, dass es vorher noch nicht eingetreten ist. In solchem Zusammenhang spricht man auch von

Streptokokken	nicht vergrößert	vergrößert	stark vergrößert
ja	19	29	24
nein	497	560	269

Tabelle 4.1.: Mandelgröße

der diskreten Hazard-Rate, welche genau der gesuchten Wahrscheinlichkeit

$$\lambda(j, \boldsymbol{x}) = P(Y = j \mid Y \geq j, \boldsymbol{x}), \quad j = 1, \ldots, c$$

aus dem sequentiellen Modell entspricht.

4.4. Beispiele

Beispiel 5. *Wir betrachten die in Tabelle 4.1 vorliegenden Daten. Es wurden bei 1398 Kindern festgestellt, ob sie Eiter hervorrufende Streptokokken haben oder nicht und zugleich die Größe ihrer Mandeln gemessen. Die Größe wurde in drei Kategorien „nicht vergrößert“, „vergrößert“ und „stark vergrößert“ erfasst. Dieser Datensatz wurde unter anderen bereits von Agresti (2010) untersucht. Hierbei kann man davon ausgehen, dass der Zustand „stark vergrößert“ nur eintreten kann, wenn vorher die Stufen „nicht vergrößert“ und „vergrößert“ durchlaufen worden sind. Daher scheint hier eine sequentielle Modellannahme geeignet zu sein und wir verwenden daher* `family=sratio`*. Weiters wollen wir die Wahrscheinlichkeiten* $P(Y = j \mid Y \geq j)$ *betrachten, was wir mittels* `reverse=FALSE` *erreichen und gehen von proportionalen Odds aus (*`parallel=TRUE`*). Indem wir uns eine Matrix mit den Gewichten und einen Faktor mit den beiden Zuständen „ja“ oder „nein“ generieren, können wir mit der Funktion* `vglm`*, welche im Abschnitt 3.7.1 beschrieben ist, die Parameter schätzen.*

```
> g1<-c(19,29,24)
> g2<-c(497,560,269)
> Matrix<-rbind(g1,g2)
> traeger<-factor(c("1-ja","0-nein"))
> library(VGAM)
> fit_vglm_seq <- vglm(Matrix ~ traeger, family=sratio(reverse=FALSE,parallel=TRUE))
```

```
> summary(fit_vglm_seq)

Call:
vglm(formula = Matrix ~ traeger, family = sratio(reverse = FALSE,
    parallel = TRUE))

Pearson residuals:
   logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
g1           0.050963          -0.052368
g2          -0.010777           0.014092

Coefficients:
              Estimate Std. Error z value
(Intercept):1 -0.51102   0.056141 -9.1025
(Intercept):2  0.73218   0.072864 10.0486
traeger1-ja   -0.52846   0.197747 -2.6724

Number of linear predictors:  2

Names of linear predictors: logit(P[Y=1|Y>=1]), logit(P[Y=2|Y>=2])

Dispersion Parameter for sratio family:   1

Residual deviance: 0.00566 on 1 degrees of freedom

Log-likelihood: -11.76594 on 1 degrees of freedom

Number of iterations: 3
```

Mit den Schätzern können wir nun die Wahrscheinlichkeiten berechnen. Sei $x = 1$*, wenn ein Kind Träger von Streptokokken ist und* $x = 0$ *sonst, sowie* $j = 1$ *für nicht vergrößerte Mandeln,* $j = 2$ *für vergrößerte und* $j = 3$ *für stark vergrößerte Mandeln, so erhalten*

wir

$$P(Y = 1 \mid Y \geq 1, \boldsymbol{x} = 0) = \frac{\exp(-0.51)}{1 + \exp(-0.51)} = 0.37$$

$$P(Y = 2 \mid Y \geq 2, \boldsymbol{x} = 0) = \frac{\exp(0.73)}{1 + \exp(0.73)} = 0.68$$

$$P(Y = 1 \mid Y \geq 1, \boldsymbol{x} = 1) = \frac{\exp(-0.51 - 0.53)}{1 + \exp(-0.51 - 0.53)} = 0.26$$

$$P(Y = 2 \mid Y \geq 2, \boldsymbol{x} = 1) = \frac{\exp(0.73 - 0.53)}{1 + \exp(0.73 - 0.53)} = 0.55.$$

Daraus lassen sich nun die Punktwahrscheinlichkeiten wie in Formel 4.2 berechnen

$$P(Y = j \mid \boldsymbol{x}) = P(Y = j \mid Y \geq j, \boldsymbol{x}) P(Y \geq j \mid \boldsymbol{x})$$

und es ergibt sich

$$\begin{aligned}
P(Y = 1 \mid \boldsymbol{x} = 0) &= P(Y = 1 \mid Y \geq 1, \boldsymbol{x} = 0) \cdot 1 \\
&= 0.375 \\
P(Y = 2 \mid \boldsymbol{x} = 0) &= P(Y = 2 \mid Y \geq 2, \boldsymbol{x} = 0)(1 - P(Y = 1 \mid Y \geq 1, \boldsymbol{x} = 0)) \\
&= 0.422 \\
P(Y = 3 \mid \boldsymbol{x} = 0) &= 1 \cdot \prod_{j=1}^{2} (1 - P(Y = j \mid Y \geq j, \boldsymbol{x} = 0)) \\
&= 0.203.
\end{aligned}$$

Analog für den Fall, dass ein Kind Träger von Streptokokken ist. Diese Ergebnisse liefert uns auch der Aufruf `fitted`*.*

```
> fitted(fit_vglm_seq)
         mu1       mu2       mu3
g1 0.2612503 0.4068696 0.3318801
g2 0.3749547 0.4220828 0.2029625
```

Beispiel 6. *Wir wollen hier als Vergleich zum kumulativen Modell auch noch auf das Beispiel zurückgreifen, in dem es darum ging, ob man Astrologie als „sehr“, „mehr oder weniger“ oder „gar nicht“ wissenschaftlich einstufen würde (vergleiche Daten in Tabelle 3.2).*

```
> fit_vglm_seq <- vglm(Matrix ~ Bildung, family=sratio(reverse=FALSE,parallel=TRUE))
> summary(fit_vglm_seq)

Call:
vglm(formula = Matrix ~ Bildung, family = sratio(reverse = FALSE,
    parallel = TRUE))

Pearson residuals:
                  logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
wenigerHighschool          -0.411207            0.67328
Highschool                 -0.095543            0.21107
JunCollege                 -0.185324            0.51260
Bachelor                    0.733025           -2.38718
Graduate                   -0.031523            0.13737

Coefficients:
                      Estimate Std. Error  z value
(Intercept):1         -0.030374    0.12429 -0.24437
(Intercept):2          1.139228    0.15194  7.49809
BildungBachelor        1.302361    0.17520  7.43351
BildungGraduate        1.856469    0.24752  7.50016
BildungHigh school     0.572455    0.13531  4.23056
BildungJunior college  0.994769    0.20163  4.93352

Number of linear predictors:  2

Names of linear predictors: logit(P[Y=1|Y>=1]), logit(P[Y=2|Y>=2])

Dispersion Parameter for sratio family:   1

Residual deviance: 6.42466 on 4 degrees of freedom

Log-likelihood: -27.51263 on 4 degrees of freedom

Number of iterations: 4
> fitted(fit_vglm_seq)
                         mu1       mu2         mu3
wenigerHighschool 0.4924071 0.3845208 0.123072085
Highschool        0.6322964 0.3114650 0.056238645
```

```
JunCollege      0.7240008 0.2467885 0.029210708
Bachelor        0.7810827 0.2013917 0.017525605
Graduate        0.8612959 0.1320989 0.006605179
```

Die geschätzten Punktwahrscheinlichkeiten weichen hier lediglich um weniger als 0.01 vom kumulativen Modell ab.

4.5. Modellierung des sequentiellen Modells als binäres Modell

Das sequentielle Modell lässt sich auch als binäres Modell darstellen und mittels geeigneter Datentransformation als generalisiertes lineares Modell schätzen. Dazu müssen diese wie in Tutz (2000) beschrieben aufgebläht werden. Wie das genau geschieht, wollen wir anhand des Beispiels 5 erklären. Die ordinale Responsevariable, welche hier der Größe der Mandeln entspricht, kann drei Stufen annehmen. Wir wollen Y mit 0-1-Variablen kodieren. Eine 1 an der Stelle j bedeutet, dass Y in Kategorie j fällt und 0, dass sie in eine höhere Kategorie fällt. Für unsere drei Fälle ergibt sich also folgendes.

Kategorie 1	nicht vergrößerte Mandeln	$Y_1 = (1)$
Kategorie 2	vergrößerte Mandeln	$Y_2 = (0,1)^T$
Kategorie 3	stark vergrößerte Mandeln	$Y_3 = (0,0)^T$

Tabelle 4.2.: Kodierung von Y_1, Y_2 und Y_3

Da die 3. Kategorie die höchste ist, genügt es diese mit zwei Nullen zu kodieren. Die erste Null steht dafür, dass Y nicht in Kategorie 1 sondern in eine höhere Kategorie fällt, die zweite dafür, dass sie nicht in Kategorie 2 sondern in eine höhere fällt. Daraus folgt, dass sie in Kategorie 3 fällt, da es nur diese drei Stufen gibt. Insgesamt ist der Vektor für Y dann von folgender Form

$$Y = (1,0,1,0,0)^T.$$

Dieser muss abhängig von der Anzahl der Stufen der erklärenden Variable x vervielfacht werden. Da x bei uns nur zwei Stufen hat (Träger von Streptokokken: 1-ja oder 0-nein)

Y fällt in	Y	Träger	α_1	α_2	w
Kategorie 1	1	0	1	0	497
Kategorie 2	0	0	1	0	560
Kategorie 2	1	0	0	1	560
Kategorie 3	0	0	1	0	269
Kategorie 3	0	0	0	1	269
Kategorie 1	1	1	1	0	19
Kategorie 2	0	1	1	0	29
Kategorie 2	1	1	0	1	29
Kategorie 3	0	1	1	0	24
Kategorie 3	0	1	0	1	24

Tabelle 4.3.: Kodierung von Y, α_1, α_2 und den Gewichten w

wird Y verdoppelt und die zugehörige erklärende Variable sieht wie folgt aus

$$\text{traeger} = (0,0,0,0,0,1,1,1,1)^T \quad \text{mit}$$

$$Y = (1,0,1,0,0,1,0,1,0,0)^T \quad \text{und Gewichten}$$

$$w = (497, 560, 560, 269, 269, 19, 29, 29, 24, 24).$$

Dabei geben die Gewichte die jeweilige Anzahl an Kindern an, die in die entsprechende Kategorie fallen. Weiters müssen wir die Inercepts α_j angeben. Dabei ist α_j genau dann 1, wenn Kategorie j beschrieben wird und sonst 0. In Fall von drei Stufen wie im hier vorliegenden Beispiel benötigt man zwei Intercepts, welche im Zusammenhang zu Y, x und den Gewichten in der folgenden Tabelle 4.3 sind.

Zum besseren Verständnis geben wir in Tabelle 4.4 noch eine Übersicht an, die zeigt wie es für eine ordinale Response mit 5 Stufen und hier frei gewählten Gewichten w aussehen würde. Mit der Kodierung für das Beispiel mit den vergrößerten Mandeln können wir jetzt mit der Funktion `glm` eine Parameterschätzung durchführen.

```
> y<-c(1,0,1,0,0,1,0,1,0,0)
> alpha1<-c(1,1,0,1,0,1,1,0,1,0)
```

Y fällt in	Y	x	α_1	α_2	α_3	α_4	w
Kategorie 1	1	0	1	0	0	0	5
Kategorie 2	0	0	1	0	0	0	10
Kategorie 2	1	0	0	1	0	0	10
Kategorie 3	0	0	1	0	0	0	40
Kategorie 3	0	0	0	1	0	0	40
Kategorie 3	1	0	0	0	1	0	40
Kategorie 4	0	0	1	0	0	0	44
Kategorie 4	0	0	0	1	0	0	44
Kategorie 4	0	0	0	0	1	0	44
Kategorie 4	1	0	0	0	0	1	44
Kategorie 5	0	0	1	0	0	0	56
Kategorie 5	0	0	0	1	0	0	56
Kategorie 5	0	0	0	0	1	0	56
Kategorie 5	0	0	0	0	0	1	56
Kategorie 1	1	1	1	0	0	0	2
Kategorie 2	0	1	1	0	0	0	8
Kategorie 2	1	1	0	1	0	0	8
Kategorie 3	0	1	1	0	0	0	34
Kategorie 3	0	1	0	1	0	0	34
Kategorie 3	1	1	0	0	1	0	34
Kategorie 4	0	1	1	0	0	0	40
Kategorie 4	0	1	0	1	0	0	40
Kategorie 4	0	1	0	0	1	0	40
Kategorie 4	1	1	0	0	0	1	40
Kategorie 5	0	1	1	0	0	0	46
Kategorie 5	0	1	0	1	0	0	46
Kategorie 5	0	1	0	0	1	0	46
Kategorie 5	0	1	0	0	0	1	46

Tabelle 4.4.: Kodierung von Y, $\alpha_1, \ldots, \alpha_4$ und den Gewichten w

```
> alpha2<-c(0,0,1,0,1,0,0,1,0,1)
> traeger_b<-c(0,0,0,0,0,1,1,1,1,1)
> w<-c(497,560,560,269,269,19,29,29,24,24)
> matrix(c(traeger_b,y,alpha1,alpha2,w),ncol=5)
      [,1] [,2] [,3] [,4] [,5]
 [1,]    0    1    1    0  497
 [2,]    0    0    1    0  560
 [3,]    0    1    0    1  560
 [4,]    0    0    1    0  269
 [5,]    0    0    0    1  269
 [6,]    1    1    1    0   19
 [7,]    1    0    1    0   29
 [8,]    1    1    0    1   29
 [9,]    1    0    1    0   24
[10,]    1    0    0    1   24
> FIT <- glm( y~ alpha1+alpha2 + traeger_b -1,
family=binomial, weights=w); b <- FIT$coeff
> FIT$coeff
    alpha1     alpha2  traeger_b
-0.5110188  0.7321801 -0.5284613
```

Der Vergleich mit den Ergebnissen aus Beispiel 5 zeigt, dass es sich um exakt dieselben Schätzer handelt, welche wir mit der Funktion `vglm` erhalten haben. Beim Vergleich der Standard Errors, welche wir mittels `summary` erhalten, sehen wir, dass auch diese identisch sind.

```
> summary(FIT)

Call:
glm(formula = y ~ alpha1 + alpha2 + traeger_b - 1, family = binomial,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-24.600  -13.475   -4.002    6.827   31.226

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
alpha1    -0.51102    0.05614  -9.102  < 2e-16 ***
```

```
alpha2     0.73218    0.07286  10.049  < 2e-16 ***
traeger_b -0.52846    0.19790  -2.670  0.00758 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3160.8  on 10  degrees of freedom
Residual deviance: 2955.2  on  7  degrees of freedom
AIC: 2961.2

Number of Fisher Scoring iterations: 6
```

4.6. Alternative Link-Funktionen

Im Abschnitt 4.2 haben wir angenommen, dass die latente Zufallsvariable Z der logistischen Verteilung genügt. Jedoch können auch andere Verteilungsfunktionen eingesetzt werden wie zum Beispiel die Normalverteilung.

Einen besonders interessanten Fall liefert uns die Annahme einer Gompertz-Verteilung mit folgender Verteilungsfunktion

$$F(\eta) = 1 - \exp(-\exp(\eta)).$$

Sie führt uns zu einem sequentiellen Minimum-Extremwert-Modell

$$P(Y = j \mid Y \geq j, \boldsymbol{x}) = 1 - \exp(-\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x})),$$

welches äquivalent zum kumulativen Minimum-Extremwert-Modell aus Abschnitt 3.9 mit ist, bei dem

$$P(Y \leq j \mid \boldsymbol{x}) = 1 - \exp(-\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}))$$

unter dem Modell mit positiven Vorzeichen gilt. Für $j = 1$ ist

$$P(Y = 1 \mid Y \geq 1, \boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{x}) = P(y \leq 1 \mid \boldsymbol{x})).$$

Damit folgt für die Schätzung, dass in beiden Modellen die Werte für α_1 und $\boldsymbol{\beta}$ übereinstimmen müssen. Da mit derselben Funktion der Parameter verschiedene Wahrscheinlichkeiten geschätzt werden, ergeben sich für $\alpha_j, j = 2, \ldots, c-1$ verschiedene Werte.

4.6.1. Vergleich anhand eines Beispiels

Beispiel 7. *Zunächst betrachten wir erneut die Daten aus Beispiel 5. Wir wollen sehen, was uns die Normalverteilung als alternative Wahl der Linkfunktion liefert.*

```
> fit_vglm_probit <- vglm(Matrix ~ traeger, family=sratio(link="probit",
                          reverse=FALSE,parallel=TRUE))
> summary(fit_vglm_probit)

Call:
vglm(formula = Matrix ~ traeger, family = sratio(link = "probit",
    reverse = FALSE, parallel = TRUE))

Pearson residuals:
   probit(P[Y=1|Y>=1]) probit(P[Y=2|Y>=2])
g1            0.066621           -0.072046
g2           -0.014661            0.018905

Coefficients:
             Estimate Std. Error z value
(Intercept):1 -0.31862   0.034689 -9.1850
(Intercept):2  0.45433   0.044410 10.2304
traeger1-ja   -0.32336   0.120302 -2.6879

Number of linear predictors:  2

Names of linear predictors: probit(P[Y=1|Y>=1]), probit(P[Y=2|Y>=2])

Dispersion Parameter for sratio family:   1

Residual deviance: 0.01021 on 1 degrees of freedom

Log-likelihood: -11.76822 on 1 degrees of freedom

Number of iterations: 3
> fitted(fit_vglm_probit)
         mu1       mu2       mu3
g1 0.2604431 0.4083112 0.3312456
g2 0.3750064 0.4219990 0.2029947
```

Die Wahrscheinlichkeiten unterscheiden sich lediglich um Werte kleiner 0.02 von jenen unter dem Logit-Link-Modell.

Beispiel 8. *Mit Hilfe der Astrologie-Daten wollen wir noch zeigen, dass das sequentielle geschätzte Modell mit dem kumulativen Modell für die Gompertzverteilung übereinstimmt. Dazu wählen wir den* `cloglog`*-Link, welcher der Gumpelverteilung entspricht. Wir werden sehen, dass die geschätzten Wahrscheinlichkeiten übereinstimmen. Die Gumbelverteilung entspricht einer Transformation der Gompertzverteilung. Zunächst betrachten wir das sequentielle Modell.*

```
> fit_vglm_seq_cloglog <- vglm(Matrix ~ Bildung, family=sratio(link="cloglog",
                               reverse=FALSE,parallel=TRUE))
> summary(fit_vglm_seq_cloglog)

Call:
vglm(formula = Matrix ~ Bildung, family = sratio(link = "cloglog",
    reverse = FALSE, parallel = TRUE))

Pearson residuals:
                  cloglog(P[Y=1|Y>=1]) cloglog(P[Y=2|Y>=2])
wenigerHighschool             -1.02750              1.29792
Highschool                    -0.31635              0.50125
JunCollege                    -0.15667              0.30959
Bachelor                       1.22954             -2.82482
Graduate                       0.10925             -0.37667

Coefficients:
                     Estimate Std. Error z value
(Intercept):1         -0.32660   0.081696 -3.9977
(Intercept):2          0.27341   0.087026  3.1417
BildungBachelor        0.70976   0.100854  7.0375
BildungGraduate        0.99381   0.122330  8.1240
BildungHigh school     0.33664   0.086468  3.8932
BildungJunior college  0.57644   0.116784  4.9360

Number of linear predictors:  2
```

```
Names of linear predictors: cloglog(P[Y=1|Y>=1]), cloglog(P[Y=2|Y>=2])

Dispersion Parameter for sratio family:   1

Residual deviance: 11.64079 on 4 degrees of freedom

Log-likelihood: -30.12069 on 4 degrees of freedom

Number of iterations: 4
> fitted(fit_vglm_seq_cloglog)
                         mu1       mu2         mu3
wenigerHighschool 0.5139165 0.3555096 0.130573889
Highschool        0.6358155 0.3063761 0.057808362
JunCollege        0.7230242 0.2502767 0.026699106
Bachelor          0.7693636 0.2147107 0.015925670
Graduate          0.8575538 0.1383584 0.004087808
```

Beim kumulative Modell sehen wir, dass sich wie zuvor erläutert nur der Intercept α_2 von jenem des sequentiellen Modells unterscheidet. Die daraus resultierenden Wahrscheinlichkeiten sind identisch.

```
> fit_vglm_cloglog <- vglm(Matrix ~ Bildung, family=
cumulative(link="cloglog",parallel=TRUE))
> summary(fit_vglm_cloglog)

Call:
vglm(formula = Matrix ~ Bildung, family = cumulative(link = "cloglog",
    parallel = TRUE))

Pearson residuals:
                  cloglog(P[Y<=1]) cloglog(P[Y<=2])
wenigerHighschool         -1.25708          1.07707
Highschool                -0.39808          0.43915
JunCollege                -0.20243          0.28180
Bachelor                   1.61622         -2.62272
Graduate                   0.14972         -0.36248
```

```
Coefficients:
                      Estimate Std. Error z value
(Intercept):1         -0.32660   0.081696 -3.9977
(Intercept):2          0.71090   0.079739  8.9153
BildungBachelor        0.70976   0.100854  7.0375
BildungGraduate        0.99381   0.122330  8.1240
BildungHigh school     0.33664   0.086468  3.8932
BildungJunior college  0.57644   0.116784  4.9360

Number of linear predictors:  2

Names of linear predictors: cloglog(P[Y<=1]), cloglog(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: 11.64079 on 4 degrees of freedom

Log-likelihood: -30.12069 on 4 degrees of freedom

Number of iterations: 5
> fitted(fit_vglm_cloglog)
                          mu1       mu2         mu3
wenigerHighschool 0.5139164 0.3555097 0.130573879
Highschool        0.6358155 0.3063762 0.057808330
JunCollege        0.7230241 0.2502768 0.026699084
Bachelor          0.7693637 0.2147107 0.015925625
Graduate          0.8575538 0.1383584 0.004087799
```

5. Log-lineare Poisson-Modelle mit fixen Scores

In diesem Kapitel soll ein weiterer Zugang zur Analyse ordinaler Daten beschrieben werden. Es geht dabei um Modelle, die Strukturen in Tabellen erlauben und die versuchen, diese Strukturen, insbesondere die Abhängigkeiten zwischen zwei Variablen, zu modellieren. Dafür ist das Vorliegen einer Kontingenztabelle erforderlich. Zunächst sollen einfache log-lineare Modelle für Kontingenztafeln betrachtet werden. Durch Abwandlungen gelangt man zum Linear-by-Linear-Association-Model, welches den Zusammenhang zweier ordinaler Variablen erklärt. Schließlich werden wir uns mit dem Row-Effect-Association-Modell befassen, durch welches sich der Einfluss einer nominalen auf eine ordinale Variable modellieren lässt.

Wir gehen von einer Stichprobe mit n Beobachtungen aus, die sich in eine $r \times c$ - Kontingenztabelle (siehe Tabelle 5.1) eintragen lassen, welche 2 Merkmale X und Y mit r bzw. c Ausprägungen beschreibt. Die Zellanzahlen n_{ij} mit $\sum_i \sum_j n_{ij} = n$ geben

	Y_1	Y_2	$\dots$	Y_j	$\dots$	Y_c
X_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1c}
X_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2c}
$\dots$			$\dots$			
X_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ic}
$\dots$			$\dots$			
X_r	n_{r1}	n_{r2}	$\dots$	n_{rj}	$\dots$	n_{rc}

Tabelle 5.1.: Kontingenztabelle mit beobachteten Häufigkeiten n_{ij}

an, wie viele der Beobachtungen Ausprägung i im einen und Ausprägung j im anderen Merkmal aufweisen und seien multinomialverteilt. Die zugehörigen erwarteten Anzahlen μ_{ij} sollen geschätzt werden. Für die multinomialen Zellwahrscheinlichkeiten gilt

$$\pi_{ij} = \frac{\mu_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{ij}} = \frac{\mu_{ij}}{n}.$$

Die folgenden Abschnitte 5.1, 5.2 und 5.3 basieren auf Simonoff (2003).

5.1. Einfache log-lineare Modelle

Man versucht die logarithmierten erwarteten Anzahlen mittels einem globalen Parameter λ, einen zeilenabhängigen Parameter λ_i^X und einen spaltenabhängigen Parameter λ_j^Y zu beschreiben. Das log-lineare Modell bei statistischer Unabhängigkeit einer $r \times c$ - Tabelle ist

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \quad i = 1, \dots r,\ j = 1, \dots, c.$$

An die Parameter müssen Bedingungen geknüpft werden, da sonst nicht alle Parameter schätzbar sind. Beispielsweise kann man verlangen, dass

$$\sum_{i=1}^{r} \lambda_i^X = 0,$$

$$\sum_{j=1}^{c} \lambda_j^Y = 0$$

gilt. Eine alternative Bedingung wäre, dass die Parameter für die letzte Zeile und die letzte Spalte 0 sind

$$\lambda_r^X = 0,$$

$$\lambda_c^Y = 0.$$

In diesem Modell gibt es $(r-1)(c-1)$ freie Parameter. Es werden mögliche Zusammenhänge zwischen den Merkmalen X und Y außer Acht gelassen. Um diese mit einzubeziehen benötigt man Zeilen- und Spalten-abhängige Parameter. Das generelle log-lineare Modell

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \dots r,\ j = 1, \dots, c.$$

	Y_1	$\dots$	Y_j	$\dots$	Y_c
X_1					
$\dots$					
X_i			μ_{ij}	$\longleftrightarrow$	μ_{ic}
$\dots$			$\updownarrow$		$\updownarrow$
X_r			μ_{rj}	$\longleftrightarrow$	μ_{rc}

Tabelle 5.2.: Log-Odds Ratio

beinhaltet zusätzlich weitere $(r-1)(c-1)$ linear unabhängige λ_{ij}^{XY}-Terme (Bedingung: $\lambda_{ij}^{XY} = 0$, wenn $i = r$ oder $j = c$ ist). Diese entsprechen genau den Log-Odds Ratios eines 2×2 - Abschnitts, bei dem die letzte Zeile und die letzte Spalte als Vergleich genommen werden (siehe Tabelle 5.2), denn

$$\begin{aligned}
\log\left(\frac{\mu_{ij}/\mu_{ic}}{\mu_{rj}/\mu_{rc}}\right) &= \log\left(\frac{\mu_{ij}/\mu_{rj}}{\mu_{ic}/\mu_{rc}}\right) \\
&= \log\mu_{ij} + \log\mu_{rc} - \log\mu_{ic} - \log\mu_{rj} \\
&= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda + \lambda_r^X + \lambda_c^Y + \lambda_{rc}^{XY} \\
&- \lambda - \lambda_i^X - \lambda_c^Y - \lambda_{ic}^{XY} - \lambda - \lambda_r^X - \lambda_j^Y - \lambda_{rj}^{XY} \\
&= \lambda_{ij}^{XY}.
\end{aligned}$$

In diesem Modell sind insgesamt $r \cdot c$ Parameter zu schätzen. Dies entspricht gerade der Anzahl der Zellen in der Kontingenztabelle. Es handelt sich also um das saturierte Modell und daher ist die Anzahl der Freiheitsgrade gleich 0. Bei diesen Ansätzen werden die Daten als nominal angenommen. Liegen ordinale Daten vor, so wird diese Struktur ignoriert. Somit sind diese Modelle für die Analyse ordinaler Daten nur im beschränkten Maße sinnvoll. Für ordinale Daten werden Terme benötigt, welche die Trends in den Variablen erlauben und darstellen. Solche Modelle sind oft komplexer. Parameter, welche die Zusammenhänge einfach erklären, sind dabei erwünscht. Somit wird die Aussagekraft erhöht und statistische Schlussfolgerungen für die gefundenen Effekte werden verstärkt.

5.2. Linear-by-linear-association-Modell

Für dieses Modell wird angenommen, dass die beiden vorliegenden Variablen X und Y eine ordinale Struktur in einer $r \times c$ - Kontingenztabelle aufweisen. Mittels nur einem Parameter mehr als beim Modell, in dem X und Y als unabhängig angenommen wurden, kann man den positiven oder negativen Zusammenhang zwischen den beiden Variablen beschreiben. Dazu teilt man den Zeilen und Spalten geordnete Scores $u_1 \leq u_2 \leq \cdots \leq u_r$ (für die Zeilen, X) und $v_1 \leq v_2 \leq \cdots \leq v_c$ (für die Spalten, Y) zu. Diese repräsentieren die Ordnung. Das Modell

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j, \quad i = 1, \ldots r,\, j = 1, \ldots, c.$$

mit der möglichen Bedingung

$$\lambda_r^X = \lambda_c^Y = 0$$

ist ein Spezialfall des in Abschnitt 5.1 besprochenen Modells, wobei $\lambda_{ij}^{XY} = \beta u_i v_j$ ist. Das Unabhängigkeits-Modell benötigte $(r-1)(c-1)$ Parameter um den Zusammenhang zu beschreiben. Hier wird dazu nur noch ein Parameter benötigt. Demzufolge liegt die Anzahl der Freiheitsgrade hier bei $(r-1)(c-1)-1$. Der Term $\beta u_i v_j$ beschreibt die Veränderung von der logarithmierten erwarteten Anzahl für abhängige X und Y im Vergleich zum Vorliegen von Unabhängigkeit zwischen X und Y. Diese Veränderung ist linear in den X-Scores bei einem fixen Level von Y und ist linear in den Y-Scores bei einem fixen Level von X. Unabhängigkeit liegt genau dann vor, wenn $\beta = 0$ ist.

Für das Log-Odds Ratio einer 2×2 - Untertabelle mit Zeilen a und b $(a < b)$ und Spalten c und d $(c < d)$ gilt

$$\begin{aligned}
\log\left(\frac{\mu_{ac}/\mu_{ad}}{\mu_{bc}/\mu_{bd}}\right) &= \log \mu_{ac} + \log \mu_{bd} - \log \mu_{ad} - \log \mu_{bc} \\
&= \beta(u_a v_c + u_b v_d - u_a v_d - u_b v_c) \\
&= \beta(u_b - u_a)(v_d - v_c).
\end{aligned}$$

Daraus kann man entnehmen, dass β die Stärke und die Richtung des Zusammenhangs beschreibt. Ist β positiv, so steigt Y, wenn auch X steigt. Bei negativem β fällt Y, wenn X steigt. Je größer $|\beta|$, desto stärker ist der Zusammenhang.

In diesem Modell gibt es keine Vorgabe für die Scores u_i und v_j. Durch eine passende Wahl kann man sich jedoch die Interpretationen erleichtern. Man kann die Scores so wählen, dass $u_i - u_{i-1}$ konstant ist für alle $i = 2, \ldots, r$ und $v_j - v_{j-1}$ konstant ist für alle $j = 2, \ldots, c$ und man erreicht, dass die Log-Odds Ratios benachbarter Zellen konstant sind. Damit sind auch die Odds Ratios konstant. Ein Spezialfall wäre die Wahl $u_i = i$ und $v_j = j$. Es ergibt sich für die Log-Odds Ratios benachbarter Zellen

$$\log\left(\frac{\mu_{a,c}/\mu_{a,c+1}}{\mu_{a+1,c}/\mu_{a+1,c+1}}\right) = \beta(u_{a+1} - u_a)(v_{c+1} - v_c) = \beta \tag{5.1}$$

und die zugehörigen Odds Ratios sind demzufolge $\exp(\beta)$.

Eine ordinale Datenstruktur kann auch durch das Teilen einer Skala in Teilbereiche entstehen. Betrachtet man zum Beispiel die Anzahl der Kinder in einer Familie $0, 1, 2, 3, 4, \ldots$, so kann man diese in drei Gruppen teilen: *0 bis 2 Kinder*, *3 bis 5 Kinder* und *mehr als 5 Kinder*. Hier ist es sinnvoll die Scores so zu wählen, dass diese in der Mitte der Bereiche liegen, z.B. $(1, 4, 7)$.

Hat man Daten vorliegen, bei dem ein positiver oder negativer Zusammenhang zu erkennen ist, so ist das Linear-by-linear-Association-Modell besser geeignet als das Modell, bei dem von Unabhängigkeit ausgegangen wird.

5.3. Row-effect-association-Modell

Das Linear-by-linear-Association-Modell kann durch Parametrisierung der Scores verallgemeinert werden. Behandelt man die Scores wie Parameter, anstatt sie als fix zu betrachten, kann man zu besseren Schätzungen gelangen. Die Scores müssen in diesem Fall nicht bestimmt werden. Ändert man allerdings die Ordnung der Kategorien, erhält man den gleichen Fit (mit permutierten Werten für die geschätzten Scores). Die Daten werden also als nominal angenommen.

Verwendet man die Parametrisierung sowohl für die Zeilen- als auch für die Spaltenscores, so spricht man vom Row+Column-Effect-Modell (vergleiche Simonoff, 2003). Haben wir eine nominale Zeilenvariable X und eine ordinale Spaltenvariable Y vorliegen, so ist es möglich für Y fixe monotone Scores anzugeben und für X Parameterscores schätzen zu lassen. Modelliert man diesen nominal-ordinalen Zusammenhang gelangt

man zum Row-Effect-Association-Modell (kurz Row-Effect-Modell). Dabei wird der geordnete Parameter βu_i durch einen ungeordneten Parameter κ_i ersetzt und man erhält

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \kappa_i v_j$$

mit der Nebenbedingung $\lambda_r^X = \lambda_c^Y = \kappa_r = 0$. Es hat $r - 1$ Parameter mehr als das Unabhängigkeitsmodell. Unabhängigkeit liegt vor, wenn $\kappa_i = 0$ für alle $i = 1, \ldots, r$. Möchte man zwei Zeilen a und b für benachbarte Spalten vergleichen und betrachtet das zugehörige Log-Odds Ratio

$$\begin{aligned} \log \left(\frac{\mu_{aj}/\mu_{a,j+1}}{\mu_{bj}/\mu_{b,j+1}} \right) &= \kappa_a v_j + \kappa_b v_{j+1} - \kappa_a v_{j+1} - \kappa_b v_j \\ &= (\kappa_b - \kappa_a)(v_{j+1} - v_j), \end{aligned}$$

so kann man erkennen, dass die Log-Odds Ratios (für alle j) dasselbe Vorzeichen besitzen wie $\kappa_b - \kappa_a$. Wählt man die Scores mit Abstand von 1 ist das Log-Odds Ratio für benachbarte Spalten allein durch $\kappa_b - \kappa_a$ bestimmt. Die κ_i werden Row-Effects genannt. Vertauscht man Zeilen und Spalten, hat man also eine ordinale Zeilenvariable mit fixen Scores u_i und eine nominale Spaltenvariable mit Parametern ν_j, so spricht man vom Column-Effect-Association-Modell, welches zum Row-Effect-Modell äquivalent ist.

5.4. Beispiel

Beispiel 9. *Wir nehmen noch ein Mal unser Astrologie-Beispiel her, welches wir bereits in Abschnitt 3.8 mit der Annahme eines kumulativen Modells untersucht haben. Die geschätzten Erwartungswerte berechnen wir hier mit der Funktion* `glm`*. Die geschätzten Parameter für das Unabhängigkeitsmodell sehen wie folgt aus.*

```
> fit_glm<-glm(Gewichte ~ Bildung_+Einschaetzung_,family=poisson)
> summary(fit_glm)

Call:
glm(formula = Gewichte ~ Bildung_ + Einschaetzung_, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-3.9828  -2.4943  -0.4189   2.4204   3.4444

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      4.92974    0.07174  68.716  < 2e-16 ***
Bildung_Bachelor                 0.49410    0.08862   5.575 2.47e-08 ***
Bildung_Graduate                -0.17552    0.10340  -1.697   0.0896 .
Bildung_High school              1.49043    0.07731  19.279  < 2e-16 ***
Bildung_Junior college          -0.18721    0.10373  -1.805   0.0711 .
Einschaetzung_mehr oder weniger -0.89584    0.05339 -16.778  < 2e-16 ***
Einschaetzung_sehr              -2.60974    0.10982 -23.763  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2202.66  on 14  degrees of freedom
Residual deviance:  107.61  on  8  degrees of freedom
AIC: 207.64

Number of Fisher Scoring iterations: 4
> fitted(fit_glm)
         1          2          3
138.343558  56.480759  10.175683
         4          5          6
614.110429 250.719465  45.170106
         7          8          9
114.723926  46.837702   8.438371
        10         11         12
226.748466  92.573341  16.678193
        13         14         15
116.073620  47.388734   8.537646
```

Die `glm`-Funktion beachtet die Ordnung der Variablen nicht, sondern sieht diese als nominal an. Dementsprechend haben wir hier etwas stärkere Abweichungen von den Modellen, bei denen die ordinale Struktur mit eingeflossen ist.

Fügen wir Scores hinzu, so können wir diese nutzen um das ordinale Verhalten der Variablen mit einfließen zu lassen. Man kann beispielsweise einen Zeilenscore $v = (1, 2, 3, 4, 5)$ *und einen Spaltenscore* $u = (1, 2, 3)$ *einfügen um damit Zeilen- und Spalten-Effekte einfließen zu lassen. Es ergeben sich deutliche Änderungen im Vergleich zur klassischen* `glm`*-Schätzung von zuvor.*

```
> u<-c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5) #Zeilen - Bildung
> v<-c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3) #Spalten - Einschätzung

> fit_glm_linear<-glm(Gewichte ~ Bildung_+Einschaetzung_+u:v,family=poisson)
> summary(fit_glm_linear)

Call:
glm(formula = Gewichte ~ Bildung_ + Einschaetzung_ + u:v, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.69382  -0.48694  -0.04418   0.24369   2.08966

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     5.03999    0.07489  67.297  < 2e-16 ***
Bildung_Bachelor                2.11088    0.19632  10.752  < 2e-16 ***
Bildung_Graduate                1.90228    0.24230   7.851 4.13e-15 ***
Bildung_High school             2.08063    0.10382  20.041  < 2e-16 ***
Bildung_Junior college          0.93750    0.16346   5.735 9.74e-09 ***
Einschaetzung_mehr oder weniger 0.10128    0.11701   0.866 0.386749
Einschaetzung_sehr             -0.77887    0.20964  -3.715 0.000203 ***
u:v                            -0.38982    0.04210  -9.260  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2202.6574  on 14  degrees of freedom
Residual deviance:    6.8389  on  7  degrees of freedom
AIC: 108.87
```

```
Number of Fisher Scoring iterations: 4
> fitted(fit_glm_linear)
         1          2          3
104.602431  78.384087  22.013482
         4          5          6
567.349542 287.898212  54.752246
         7          8          9
122.488644  42.090707   5.420649
        10         11         12
268.158230  62.399866   5.441904
        13         14         15
147.401153  23.227129   1.371718
```

Der Parameter β beschreibt die Log-Odds Ratios (siehe 5.1), somit ergibt sich das geschätzte Odds Ratio aus

$$\widehat{OR} = \exp(\hat{\beta}) = \exp(-0.38982) = 0.68,$$

das heißt die Chance eher Einschätzungen in Richtung gar nicht wissenschaftlich zu bekommen im Vergleich zu eher Antworten in Richtung sehr wissenschaftlich zu bekommen, ist bei geringerer Bildung 0.68 *mal so groß wie bei höher Bildung. Je höher die abgeschlossene Ausbildung ist, desto eher wird Astrologie als gar nicht wissenschaftlich eingestuft. Dies wird durch das negative Vorzeichen vom Parameter β bestätigt. Die geschätzten Anzahlen sind hier deutlich besser als beim Unabhängigkeitsmodell.*

Wir wollen außerdem noch zeigen, was sich für Änderungen ergeben, wenn nur Zeilen- bzw. Spalteneffekte in das Modell einbezogen werden.

```
> fit_glm_column<-glm(Gewichte ~ Bildung_+Einschaetzung_+
                      Einschaetzung_:u,family=poisson)
> summary(fit_glm_column)

Call:
glm(formula = Gewichte ~ Bildung_ + Einschaetzung_ + Einschaetzung_:u,
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-0.89763  -0.45247  -0.05916   0.35177   1.45713

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                       3.9756     0.1614  24.627  < 2e-16 ***
Bildung_Bachelor                 -1.0502     0.3603  -2.915 0.003561 **
Bildung_Graduate                 -2.3142     0.4835  -4.787 1.70e-06 ***
Bildung_High school               1.0254     0.1350   7.598 3.00e-14 ***
Bildung_Junior college           -1.1702     0.2505  -4.671 2.99e-06 ***
Einschaetzung_mehr oder weniger   0.1803     0.1348   1.338 0.180919
Einschaetzung_sehr               -1.0015     0.2817  -3.555 0.000378 ***
Einschaetzung_gar nicht:u         0.6680     0.1231   5.425 5.79e-08 ***
Einschaetzung_mehr oder weniger:u 0.2448     0.1277   1.918 0.055151 .
Einschaetzung_sehr:u                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2202.6574  on 14  degrees of freedom
Residual deviance:    5.4676  on  6  degrees of freedom
AIC: 109.5

Number of Fisher Scoring iterations: 4
> fitted(fit_glm_column)
         1          2          3
103.916598  81.511352  19.572050
         4          5          6
565.104460 290.322970  54.572570
         7          8          9
122.654649  41.272061   6.073289
        10         11         12
269.710931  59.441557   6.847512
        13         14         15
148.613363  21.452059   1.934579


> fit_glm_row<-glm(Gewichte ~Bildung_+Einschaetzung_+
                   Bildung_:v,family=poisson)
```

```
> summary(fit_glm_row)

Call:
glm(formula = Gewichte ~ Bildung_ + Einschaetzung_ + Bildung_:v,
    family = poisson)


Coefficients: (1 not defined because of singularities)
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                      3.7204     0.2354  15.807  < 2e-16 ***
Bildung_Bachelor                 2.1419     0.2409   8.890  < 2e-16 ***
Bildung_Graduate                 2.0788     0.3093   6.721 1.81e-11 ***
Bildung_High school              2.2598     0.2008  11.257  < 2e-16 ***
Bildung_Junior college           1.0936     0.2760   3.963 7.40e-05 ***
Einschaetzung_mehr oder weniger -1.0871     0.1492  -7.288 3.16e-13 ***
Einschaetzung_sehr              -3.1607     0.3067 -10.305  < 2e-16 ***
Bildung_<High school:v           0.8801     0.1799   4.893 9.92e-07 ***
Bildung_Bachelor:v              -0.2892     0.1877  -1.541   0.1234
Bildung_Graduate:v              -0.8001     0.2520  -3.175   0.0015 **
Bildung_High school:v            0.3754     0.1569   2.392   0.0168 *
Bildung_Junior college:v             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2202.66  on 14  degrees of freedom
Residual deviance:    5.31  on  4  degrees of freedom
AIC: 113.34

Number of Fisher Scoring iterations: 4

> fitted(fit_glm_row)
         1          2          3
 99.534834  80.930333  24.534834
         4          5          6
575.712620 282.574760  51.712620
         7          8          9
123.224163  41.551674   5.224163
```

```
        10         11         12
263.259572  66.480857   6.259572
        13         14         15
148.268812  22.462377   1.268812
```

Der Schätzer für die Interaktion mit dem Zeilenscore u ist für die Einschätzung „sehr“ nicht angegeben. Gleiches gilt bei dem Spaltenscore v und der Bildungsstufe „Junior college“. Wie bereits beschrieben, ist ein Parameter Null. Diese Stufe bildet dann eine Art Referenzklasse, wobei die Parameter für die anderen Stufen die entsprechenden Abweichungen von dieser Referenzklasse darstellen.

Die geschätzten Anzahlen für das Column-effect-Modell oder das Row-effect-Modell sind auch deutlich besser als für das Unabhängigkeitsmodell und ähnlich zum Linear-by-linear-association-Modell. Alle drei liefern gute Deviance-Werte in Relation zur Anzahl freier Parameter. Das Linear-by-linear-association-Modell scheint hier etwas besser als die anderen zu sein (leicht geringeren AIC). Welches im Allgemeinen das beste Modell ist, ist jedoch abhängig vom vorliegenden Beispiel.

6. Weitere Modelle

In diesem Kapitel soll ein Ausblick auf weitere Modellen besprochen werden, die geeignet sind um ordinale Daten zu beschreiben. Für genauere Ausführungen sei dazu auf Agresti (2010) und Tutz (2012) hingewiesen.

6.1. Kumulative Modelle mit Skalierungsparameter

Hat man Daten vorliegen, bei denen sich für verschiedene erklärende Variablen deutliche Unterschiede bezüglich der Ausprägungen in der Responsevariable zeigen, so ist es vielleicht von Vorteil einen Skalierungsparameter und damit das von McCullagh (1980) eingeführte Modell

$$P(Y \leq j \mid \boldsymbol{x}) = F\left(\frac{\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}}{\tau_{\boldsymbol{x}}}\right)$$

zu verwenden. Tutz (2012) hat sich ebenfalls mit diesem Modell auseinandergesetzt. Im Vergleich zum einfachen kumulativen Modell, bei welchem angenommen wurde, dass die Verteilungsfunktion unabhängig von $\boldsymbol{x}$ ist, wird hier durch den Parameter $\tau_{\boldsymbol{x}}$ ein Zusammenhang zwischen $\boldsymbol{x}$ und der Verteilungsfunktion hergestellt. Dadurch können im Falle von unterschiedlichen Streuungen für verschiedene $\boldsymbol{x}$ diese Streuungen besser modelliert werden. Im einfachen kumulativen Modell wird das Wahrscheinlichkeitsmaß bei Veränderung von $\boldsymbol{x}$ lediglich verschoben. Daher ist es in solchen Fällen nicht ganz so gut geeignet.

Die Skalierungsparameter werden zusätzlich geschätzt. Die Funktion `clm` bietet die Möglichkeit über `scale=..` anzugeben, für welche erklärende Variable ein Skalierungsparameter einfließen soll. Anhand des p-Werts kann entschieden werden, ob es sinnvoll ist den Skalierungsparameter ins Modell einfließen zu lassen oder nicht.

	sehr viel	viel	eher weniger	gar keine
Alter 6-20	30	13	5	27
Alter 20-40	40	27	16	9

Tabelle 6.1.: Freude am Sport

Beispiel 10. *Fragt man Personen verschiedener Altersklassen, nach ihrer Lust und Freude daran Sport zu treiben, so können sich beispielsweise die in Tabelle 6.1 dargestellten Anzahlen ergeben. Hier ist zu erkennen, dass die jüngeren Personen mehr in Randbereichen „sehr viel“ und „gar keine“ konzentriert sind, während bei den älteren Personen eine Monotonie erkennbar ist. Da beim einfachen kumulativen Modell die Verteilungsfunktion nur verschoben wird, ist diese für solche Fälle wie hier dargestellt ungeeignet und eine Einführung eines Skalierungsparameters wäre sinnvoll.*

6.2. Hierarchisch strukturierte Modelle

Hierarchisch strukturierte Modelle können zur Anwendung kommen, wenn es möglich ist die Kategorien der Responsevariable in Untergruppen einzuteilen. Ähneln sich die Verteilungen in den Untergruppen, so kann für diese jeweils das gleiche Modell angenommen werden (mit verschiedenen Parametern). Das Modell für die Gruppen selbst wird noch zusätzlich aufgestellt. Teilen wir also die c Kategorien in Untergruppen $S_1, \dots, S_t$, wobei $S_i = \{m_{i-1} + 1, \dots, m_i\}, m_0 = 0, m_t = c, i = 1, \dots, t$ und bezeichnen deren Vereinigungen mit $T_i = S_1 \cup \dots \cup S_i$ so ergibt sich das Modell zu

$$P(Y \leq j \mid Y \in S_i, \boldsymbol{x}) = F(\theta_{ij} - \boldsymbol{\beta}_i^T \boldsymbol{x}), \quad j \in S_i$$

$$P(Y \in T_i \mid \boldsymbol{x}) = F(\theta_i - \boldsymbol{\beta}_0^T \boldsymbol{x}).$$

Dabei sind die θ-Parameter monoton wachsend, das heißt

$$\theta_1 < \theta_2 < \dots < \theta_t = \infty$$

$$\theta_{i,m_{i-1}+1} < \dots < \theta_{i,m_i} = \infty, \quad i = 1, \dots, t.$$

Das hierarchisch strukturierte Modell ist also in zwei Schritten definiert. Zum einen wird innerhalb der Gruppen ein kumulatives Modell gebildet und schließlich über die Gruppen ebenfalls. Ein Vorteil ist, dass für die verschiedenen Gruppen unterschiedliche Parameter angenommen werden können.

6.3. Adjazent-Kategorie-Modelle

Hierfür betrachten wir die Logits von benachbarten (adjazenten) Zellen für $j = 1, \ldots, c-1$

$$\text{logit}(P(Y = j \mid Y \in \{j, j+1\}, \boldsymbol{x}) = \log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}$$

und modellieren diese mittels der Parameter α_j und $\boldsymbol{\beta}$

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, c-1.$$

Hierbei hat die erklärende Variable $\boldsymbol{x}$ für alle Logits denselben Effekt. Spezifiziert man das Modell über kategorie-abhängige Parameter $\boldsymbol{\beta}_j$ und erklärt die Logits über

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x}, \quad j = 1, \ldots, c-1,$$

so hat man ein Adjazent-Kategorie Logit Modell, bei dem die erklärenden Variablen einen unterschiedlich starken Einfluss auf die einzelnen Kategorien haben. Dieses Modell ist in gleichen Situationen anwendbar wie das kumulative Logit-Modell, da beide Modelle eine stochastische Ordnung der Intercepts und somit der Verteilung von Y für verschiedene Prädiktorwerte voraussetzen.

Beispiel 11. *Zum Vergleich liefern wir hier die Schätzung der Modellparameter für das Astrologie-Beispiel.*

```
> fit_vglm_adj <- vglm(Matrix ~ Bildung, family=acat(reverse=TRUE,parallel=TRUE))
> summary(fit_vglm_adj)

Call:
vglm(formula = Matrix ~ Bildung, family = acat(reverse = TRUE,
    parallel = TRUE))
```

```
Pearson residuals:
                  log(P[Y=1]/P[Y=2]) log(P[Y=2]/P[Y=3])
wenigerHighschool          -0.274763            0.40077
Highschool                 -0.148736            0.28497
JunCollege                 -0.260871            0.61256
Bachelor                    0.766950           -2.10694
Graduate                   -0.070489            0.25561

Coefficients:
                     Estimate Std. Error z value
(Intercept):1         0.20692    0.11355  1.8222
(Intercept):2         1.19349    0.13921  8.5730
BildungBachelor       1.16931    0.15769  7.4153
BildungGraduate       1.68027    0.23128  7.2652
BildungHigh school    0.50475    0.11669  4.3255
BildungJunior college 0.88015    0.17987  4.8932

Number of linear predictors:  2

Names of linear predictors: log(P[Y=1]/P[Y=2]), log(P[Y=2]/P[Y=3])

Dispersion Parameter for acat family:   1

Residual deviance: 5.31001 on 4 degrees of freedom

Log-likelihood: -26.9553 on 4 degrees of freedom

Number of iterations: 4
> fitted(fit_vglm_adj)
                         mu1       mu2         mu3
wenigerHighschool 0.4855358 0.3947821 0.119682115
Highschool        0.6326512 0.3105217 0.056827055
JunCollege        0.7248480 0.2444216 0.030730371
Bachelor          0.7835106 0.1978597 0.018629677
Graduate          0.8620280 0.1305952 0.007376812
```

Die Schätzer liegen recht nahe bei den wahren Werten. Es scheint für dieses Beispiel gut

geeignet zu sein.

6.4. Stereotypische Modelle

Um eine Zwischenlösung zwischen einem Adjazent-Kategorie-Modell mit kategorieabhängigen Parametern und dem Modell mit einem globalen Parameter $\boldsymbol{\beta}$ für alle Kategorien zu finden wurde das stereotypische Modell eingeführt (vergleiche Agresti, 2010). Das erstere ist auch äquivalent zum sogenannten Baseline-Category-Logit-Modell

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_c(\boldsymbol{x})} = \alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x}, \quad j = 1, \ldots, c-1,$$

welches zu viele zu schätzende Parameter beinhaltet und zusätzlich noch eine nominale Skala für die Responsevariable annimmt. Um die Anzahl der zu schätzenden Parameter zu verringern, wird ein kategorieabhängiger Parameter ϕ_j eingeführt und das stereotypische Modell beschrieben durch

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_c(\boldsymbol{x})} = \alpha_j - \phi_j \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, c-1.$$

Die Wahrscheinlichkeiten ergeben sich in diesem Modell mit $\alpha_c = \phi_c = 0$ zu

$$P(Y = j \mid \boldsymbol{x}) = \pi_j(\boldsymbol{x}) = \frac{\exp(\alpha_j - \phi_j \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \sum_{k=1}^{c-1} \exp(\alpha_k - \phi_k \boldsymbol{\beta}^T \boldsymbol{x})}, \quad j = 1, \ldots, c.$$

Die Parameter sind nicht eindeutig, da $\phi_j \boldsymbol{\beta} = (\phi_j / K) K \boldsymbol{\beta}$ für jede Konstante $K \neq 0$, es sei denn man führt zusätzliche Restriktionen wie beispielsweise $\phi_1 = 1$ ein.

Das stereotypische Modell kann auch im Zusammenhang zu adjazenten Kategorie-Logits aufgestellt werden

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j - \nu_j \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, c-1,$$

wobei folgender Zusammenhang zwischen den Parameter besteht

$$\phi_j = \sum_{k=j}^{c-1} \nu_k, \quad j = 1, \ldots, c-1 \text{ bzw.}$$

$$\nu_j = \phi_j - \phi_{j+1}.$$

7. Zusammenfassung

Wir haben in dieser Arbeit einige Modelle für die Analyse ordinaler Daten besprochen, ihre Analyse in `R` gesehen und dies anhand einiger Beispiele erläutert. Je nach Situation ist das entsprechend passende Modell zu wählen.

Die kumulativen Modelle sind in den meisten Situationen, in denen ordinale Daten vorliegen, anwendbar, sowohl mit globalem Parameter β, der für alle Kategorien gleich ist, als auch für den Fall, dass man kategoriespezifische Parameter benötigt. Ein Test auf proportionale Odds, kann hier sinnvoll sein. Die Wahl der Linkfunktion verändert die Ergebnisse nur sehr gering. Jedoch ist der Logit-Link auf Grund der einfachen Interpretation der Odds Ratios zu bevorzugen. Die Odds Ratios sind unabhängig vom Intercept und damit für alle Stufen gleich. In `R` gibt es einige Funktionen, die die Parameter für ein kumulatives Modell schätzen. Die Funktion `polr` schätzt wie `clm` und `lrm` das Modell mit negativem Vorzeichen vor dem Parameter β. Alle liefern somit dieselben Schätzer. Bei `vglm` wird im Modell vor β ein positives Vorzeichen angenommen. Diese Funktion hat jedoch den Vorteil, dass man auch das Modell mit kategorieabhängigen Parametern schätzen kann. Außerdem ist es für das sequentielle Modell und für Adjazent-Kategorie-Modelle anwendbar.

Die Anwendung eines sequentiellen Modells erweist sich als sinnvoll, wenn die Stufen der ordinalen Response durchlaufen werden müssen, das heißt, dass man nur in einer höheren Kategorie sein kann, wenn man vorher die darunterliegenden durchlaufen hat. Hier wird nämlich stets die Wahrscheinlichkeit verglichen, dass man in einer bestimmten Stufe ist unter der Gegebenheit, dass man in dieser oder einer höheren Stufe wäre. Wir haben gezeigt, wie sich ein sequentielles Modell als binäres Modell approximieren lässt. In unseren Beispielen, die wir mit Hilfe der Funktion `vglm` analysiert haben, war zu erkennen, dass sich die Punktwahrscheinlichkeiten letztendlich jedoch nur wenig von denen des kumulativen Modells unterscheiden.

Die log-linearen Poisson-Modelle wurden ebenfalls betrachtet. Dort sind wir von einem Unabhängigkeitsmodell ausgegangen, welches die vorliegenden Daten als nominal angenommen hat. Wir haben gesehen, dass durch eine geeignete Modellierung mit fixen Scores die ordinale Struktur mit Hilfe von nur einem weiteren Parameter modelliert werden kann (Linear-by-linear-association-Modell). Das Row-effect-association-Modell ist für die Modellierung des Zusammenhangs zwischen einer ordinalen und einer nominalen Merkmalsausprägung geeignet.

Darüber hinaus wurden noch ein paar weitere Modelle beleuchtet, welche sich in gewissen Situation als nützlich erweisen können.

A. Anhang

A.1. Herleitung des Schätzers $\hat{\pi}_{ij}$

Sei Y ein k-stufiger Faktor, der in I Situationen beobachtet werden kann. Die i-te Situation ($i = 1, \ldots, I$) kann durch einen Vektor von erklärenden Größen $\boldsymbol{x}_i$ beschrieben werden. In jeder Situation wird nun eine Zufallsstichprobe festen Umfangs gezogen. Bezeichne $Y_{i1}, \ldots, Y_{in_i}$ die Zufallsstichprobe in der i-ten Situation und wir interessieren uns für die Anzahl all dieser Stichprobenelemente, welche in Stufe j realisieren (absolute Häufigkeit der Stufe j in der Situation i). Dieses Experiment kann durch die Tabelle A.1 vollständig beschrieben werden. In Tabelle A.1 bezeichnet die nicht-negative, diskrete Zufallsvariable N_{ij} die Anzahl der Variablen $Y_{i1}, \ldots, Y_{in_i}$, die in der Situation i auf Stufe j realisieren.

Es könnte beispielsweise sein, dass x die Haarfarbe einer Person und $Y_{i1}, \ldots, Y_{in_i}$ die beobachteten Augenfarben von n_i Personen mit der gleichen Haarfarbe beschreiben. Oder $x = x_i$ sei ein bestimmter Monat eines Jahres und die $Y_{i1}, \ldots, Y_{in_i}$ beschreiben das Wetter der einzelnen Tage im i-ten Monat, wobei es k verschiedene Wettersituationen gibt (wie im Abschnitt 2.1). Die Anzahl der sonnigen Tage wäre dann z.B. N_{i1}, die Anzahl regnerischer Tage N_{i2} usw. und für den Juni wäre $n_i = 30$, da es 30 Beobachtungstage gibt.

Wir interessieren uns jetzt in der i-ten Situation für die Wahrscheinlichkeit π_{ij}, dass eine der Zufallsvariablen $Y_{i1}, \ldots, Y_{in_i}$ auf der j-ten Stufe realisiert. Hierfür ist es ausreichend zu wissen, wie viele dieser n_i Zufallsvariablen auf der Stufe j realisieren und wie viele dies nicht tun. Somit entspricht dies einem Bernoulli-Experiment und wir haben einen Erfolg, falls $Y_{il} = j$, für $l = 1, \ldots, n_i$. Die Anzahl dieser Erfolge ist gerade N_{ij} und wir interpretieren $n_i - N_{ij}$ als die Anzahl der erzielten Misserfolge. Dementsprechend genügt N_{ij} einer Binomialverteilung mit Parameter π_{ij} (Erfolgswahrscheinlichkeit) bei

Situation		Experimentausgang 1	...	j	...	k	Absolute Häufigkeit
1	$\mathbf{x}_1$	N_{11}	...	N_{1j}	...	N_{1k}	n_1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
i	$\mathbf{x}_i$	N_{i1}	...	N_{ij}	...	N_{ik}	n_i
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
I	$\mathbf{x}_I$	N_{I1}	...	N_{Ij}	...	N_{Ik}	n_I

Tabelle A.1.: Tabellarische Darstellung der Anzahlen und Häufigkeiten

n_i Versuchen.

Sei nun $Z_{il} = 1$ falls $Y_{il} = j$ und $Z_{il} = 0$ sonst. Für die i-te Situation ergibt sich damit die Likelihood-Funktion

$$\begin{aligned} L(\pi_{ij}|N_{ij}) &= L(\pi_{ij}|Z_{i1}, \dots Z_{in_i}) \\ &= \prod_{l=1}^{n_i} \pi_{ij}^{Z_{il}} (1-\pi_{ij})^{1-Z_{il}} \\ &= \pi_{ij}^{N_{ij}} (1-\pi_{ij})^{n_i - N_{ij}} \,, \end{aligned}$$

da für die i-te Situation $\sum_{l=1}^{n_i} Z_{il} = N_{ij}$ gilt. Für die Log-Likelihood-Funktion folgt damit

$$\log L(\pi_{ij}|N_{ij}) = N_{ij} \log \pi_{ij} + (n_i - N_{ij}) \log(1-\pi_{ij}) \,.$$

Wir suchen die Nullstelle der Score-Funktion

$$\frac{\partial}{\partial \pi_{ij}} \log L(\pi_{ij}|N_{ij}) = \frac{N_{ij}}{\pi_{ij}} - \frac{n_i - N_{ij}}{1-\pi_{ij}}$$

und erhalten durch Nullsetzen den Maximum-Likelihood-Schätzer für die Erfolgswahrscheilichkeit (in die j-te Stufe zu fallen)

$$\hat{\pi}_{ij} = \frac{1}{n_i} N_{ij}$$

was gerade der relativen Häufigkeit für dieses Experiment entspricht. Dieses Ergebnis hält für jede Situation und für jede Stufe des Faktors.

	Anzahl sonniger Tage	Anzahl nicht sonniger Tage	Gesamtanzahl
Februar	$n_{11} = 8$	$n_{12} = 20$	$n_1 = 28$
Juli	$n_{21} = 20$	$n_{22} = 11$	$n_2 = 31$
August	$n_{31} = 24$	$n_{32} = 7$	$n_3 = 31$

Tabelle A.2.: Beobachtete absolute Häufigkeiten

	geschätzte W! für sonnige Tage	geschätzte W! für nicht sonnige Tage
Februar	$\hat{\pi}_{11} = 8/28$	$\hat{\pi}_{12} = 20/28$
Juli	$\hat{\pi}_{21} = 20/31$	$\hat{\pi}_{22} = 11/31$
August	$\hat{\pi}_{31} = 24/31$	$\hat{\pi}_{32} = 7/31$

Tabelle A.3.: ML-Schätzer der Wahrscheinlichkeit für einen sonnigen Tag

Beispiel 12. *Für eine Kontingenztabelle über sonnige und nicht sonnige Tage für verschiedene Monate mit Anzahlen wie in Tabelle A.2 dargestellt ergeben sich die in Tabelle A.3 geschätzten Wahrscheinlichkeiten. Hierbei sei n_{ij} die Realisierung der Zufallsvariable N_{ij}.*

A.2. Generalisierte lineare Modelle

Bei generalisierten linearen Modellen handelt es sich um eine Verallgemeinerung der klassischen linearen Regressionsmodelle. In bestimmten Fällen ergeben sich bei den linearen Modellen unzulässige Vorhersagen. Außerdem beschränken sie sich auf Zufallsvariablen, die aus der Normalverteilung stammen. Bei den generalisierten linearen Modellen sind hingegen auch andere Verteilungen zulässig. Die betrachtete Responsevariable muss aus der einparametrigen Exponentialfamilie stammen. Sie beschreibt eine Klasse von Wahrscheinlichkeitsverteilungen, die einer bestimmten Form entsprechen. Dazu zählt die Normalverteilung, aber beispielsweise auch die Binomialverteilung, die Poissonverteilung, die Gammaverteilung oder die Invers-Gaußverteilung. Wir definieren sie hier wie in McCullagh & Nelder (1989).

Definition A.2.1 (Exponentialfamilie)**.** *Die einparametrige Exponentialfamilie ist eine Menge von Wahrscheinlichkeitsverteilungen, deren zugehörige Dichte- oder Wahrscheinlichkeitsfunktion sich durch*

$$f(y \mid \theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

darstellen lassen. Hierbei sind $a(\cdot), b(\cdot)$ und $c(\cdot)$ spezielle bekannte Funktionen mit $a(\phi) > 0$ und ϕ eine feste Größe.

Mit dem folgenden Satz, können wir den Erwartungswert und die Varianz von einer Zufallsvariable Y aus der Exponentialfamilie berechnen.

Satz A.2.2 (Eigenschaften der Score-Funktion)**.** *Sei $f(y \mid \theta)$ die Dichte- oder Wahrscheinlichkeitsfunktion von Y und θ ein unbekannter Parameter. Dann gilt für die Score-Funktion (Ableitung der Log-Likelihood-Funktion)*

$$E\left(\frac{\partial \log f(Y \mid \theta)}{\partial \theta}\right) = 0, \tag{A.1}$$

$$E\left(\frac{\partial \log f(Y \mid \theta)}{\partial \theta}\right)^2 + E\left(\frac{\partial^2 \log f(Y \mid \theta)}{\partial \theta^2}\right) = 0. \tag{A.2}$$

Beweis: Siehe Casella & Berger (2002).

Für die Exponentialfamilie ergibt sich damit

$$\frac{\partial \log f(y \mid \theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) = \frac{y - b'(\theta)}{a(\phi)}$$

$$\frac{\partial^2 \log f(y \mid \theta)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$$

und mit (A.1) erhalten wir

$$E\left(\frac{\partial \log f(Y \mid \theta)}{\partial \theta}\right) = \frac{E(Y - b'(\theta))}{a(\phi)} = 0,$$

womit

$$E(Y) = b'(\theta)$$

folgt. Ergebnis (A.2) liefert uns

$$\begin{aligned} E\left(\frac{\partial^2 \log f(Y \mid \theta)}{\partial \theta^2}\right) + E\left(\frac{\partial \log f(Y \mid \theta)}{\partial \theta}\right)^2 &= -\frac{b''(\theta)}{a(\phi)} + \frac{E(Y - b'(\theta))^2}{a^2(\phi)} \\ &= -\frac{b''(\theta)}{a(\phi)} + \frac{E(Y - E(Y))^2}{a^2(\phi)} \\ &= -\frac{b''(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)} = 0 \end{aligned}$$

und daraus

$$Var(Y) = a(\phi)b''(\theta).$$

Hierbei wird $\phi > 0$ als Dispersionsparameter und $b(\theta)$ als Kumulantenfunktion bezeichnet. Schreiben wir $Var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$, so sehen wir, dass die Varianz ein Produkt zweier unabhängiger Funktionen ist. Die sogenannte Varianzfunktion $V(\mu)$ ist nur vom Erwartungswert μ abhängig und $a(\phi)$ ist unabhängig von μ.

Das generalisierte lineare Modell ist nun durch drei Komponenten definiert.

1. Die stochastische Komponente ist der Vektor der Responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, wobei die Y_i unabhängig sind und aus der Exponentialfamilie stammen. Wir nehmen an, dass $E(Y_i)$ existiert und $E(Y_i) = \mu_i = \mu(\theta_i)$ gilt.

2. Die systematische Komponente wird durch den Vektor der linearen Prädiktoren $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ und

   $$\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i, \quad i = 1, \ldots, n$$

   beschrieben, wobei $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ der p-elementige Vektor der Prädiktorvariablen, der bekannten erklärenden Variablen, ist. Diese fassen wir zur $(n \times p)$ Designmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ zusammen. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ ist der Vektor der unbekannten zu schätzenden Parameter. Der lineare Zusammenhang zwischen diesem und $\boldsymbol{\eta}$ gibt dem Modell seinen Namen.

Verteilung	kanonischer Link	Varianzfkt.	`family`	`link`	`variance`
Normal	μ	1	`gaussian`	`identity`	`constant`
Binomial	$\log(\mu/(1-\mu))$	$\mu(1-\mu)$	`binomial`	`logit`	`mu(1-mu)`
Poisson	$\log\mu$	μ	`poisson`	`log`	`mu`
Gamma	$1/\mu$	μ^2	`Gamma`	`inverse`	`mu^2`
Invers-Gauß	$1/\mu^2$	μ^3	`inverse.gaussian`	`1/mu^2`	`mu^3`

Tabelle A.4.: Übersicht über die Link- und Varianzfunktionen

3. Die Linkfunktion $g(\cdot)$ verknüpft den Erwartungswert μ mit dem linearen Prädiktor mittels

$$g(\mu_i) = \eta_i.$$

Hierbei ist zu erkennen, dass der Zusammenhang zwischen erklärenden Variablen und Erwartungswert nicht linear sein muss, da $g(\cdot)$ beliebig sein kann. Die Linkfunktion muss monoton und zweimal stetig differenzierbar sein. Sie ist abhängig von der Verteilung der Response-Variable.

Die Schätzung des Parameters $\boldsymbol{\beta}$ erfolgt mittels der Maximum-Likelihood-Methode. Aus dieser resultiert ein nicht-lineares Gleichungssystem, welches numerisch zu lösen ist. Ein möglicher Ansatz ist die Fisher Scoring Technik (für eine genauere Beschreibung siehe McCullagh & Nelder, 1989, oder Tutz, 2012). In `R` hilft uns dabei die Funktion `glm()`. Mittels `familiy= ...` kann man die Link- und die Varianzfunktion spezifizieren. Weitere Anpassungen sind möglich. Eine Übersicht einiger möglicher Angaben für `family` zeigt Tabelle A.4. Der kanonische Link $g(\mu) = \eta = \theta$ wird in `R` defaultmäßig verwendet. Er kann jedoch auch abgeändert werden beispielsweise über `glm(y~` $\boldsymbol{x}$`, family=gaussian(link=log))`.

A.3. Logistische Regression

Die logistische Regression betrachtet den Fall eines generalisierten linearen Modells, bei dem die Binomialverteilung vorliegt. Für ein besseres Verständnis wollen wir dazu

herleiten, dass diese zur Exponentialfamilie gehört. Da die Binomialverteilung sich durch die Hintereinanderausführung von Bernoulli-Experimenten ergibt, sollen auch diese nicht außer Acht gelassen werden. Sei daher Y zunächst eine diskrete Responsevariable, die nur zwei Werte annehmen kann, 0 und 1. Oft geht es dabei um Misserfolg und Erfolg eines Zufallsexperiments. Die zugehörige Wahrscheinlichkeitsfunktion versuchen wir so umzuschreiben, dass wir erkennen können, dass sie zur Exponentialfamilie gehört und welche Parameter sie besitzt. Sei dazu $P(Y = 1) = \pi$ die Erfolgswahrscheinlichkeit, dann gilt

$$\begin{aligned} f(y \mid \pi) &= \pi^y (1-\pi)^{1-y} \\ &= \exp(y \log \pi) \exp((1-y)\log(1-\pi)) \\ &= \exp(y \log \pi + \log(1-\pi) - y \log(1-\pi)) \\ &= \exp\left(y \log \frac{\pi}{1-\pi} + \log(1-\pi)\right). \end{aligned}$$

Wir sehen, dass dies der Form einer Dichte aus der Exponentialfamilie in Definition A.2.1 entspricht, wobei $\pi = \mu$ und

$$\begin{aligned} \theta &= \log \frac{\pi}{1-\pi} \\ a(\phi) &= 1 \\ b(\theta) &= -\log(1-\pi) = \log(1+\exp\theta) \\ c(y,\phi) &= 0. \end{aligned}$$

Mit der kanonischen Linkfunktion $\eta = g(\pi) = \theta$ resultiert

$$\begin{aligned} \eta &= \log \frac{\pi}{1-\pi} \quad \text{bzw.} \\ \pi &= \frac{\exp\eta}{1+\exp\eta}. \end{aligned}$$

Wiederholen wir nun ein solches Bernoulli-Experiment m mal unabhängig hintereinander und betrachten Y als die Anzahl der Erfolge, so kann Y alle Werte in $\{0, \ldots, m\}$ annehmen und folgt einer Binomialverteilung, welche wir auch wieder in die Form der

Wahrscheinlichkeitsfunktion der Exponentialfamilie umschreiben können. Es gilt

$$\begin{aligned} f(y \mid \pi) &= \binom{m}{y} \pi^y (1-\pi)^{m-y} \\ &= \exp\left(\log \binom{m}{y} + y \log \pi + (m-y) \log(1-\pi) \right) \\ &= \exp\left(\frac{\frac{y}{m} \log \frac{\pi}{1-\pi} + \log(1-\pi)}{1/m} + \log \binom{m}{y} \right). \end{aligned}$$

Hier beschreibt Y die Anzahl der Erfolge. Dies entspricht der absoluten Häufigkeit. Dividieren wir Y durch m so gelangen wir zur relativen Häufigkeit und zur sogenannten standardisierten oder skalierten Binomialverteilung. Es sei $\tilde{Y} = Y/m$ und daher $\tilde{Y} \in \{0, \frac{1}{m}, \ldots, \frac{m}{m}\}$. Wir wissen, dass $Y = m\tilde{Y}$ binomialverteilt ist und daher folgt

$$\begin{aligned} f(\tilde{y} \mid \pi) &= P(\tilde{Y} = \tilde{y}) = P(m\tilde{Y} = m\tilde{y}) \\ &= \exp\left(\frac{\tilde{y} \log \frac{\pi}{1-\pi} + \log(1-\pi)}{1/m} + \log \binom{m}{m\tilde{y}} \right). \end{aligned}$$

Wir sehen, dass obiges der Form einer Wahrscheinlichkeitsfunktion aus der Exponentialfamilie entspricht, wobei

$$\begin{aligned} \theta &= \log \frac{\pi}{1-\pi} \\ a(\phi) &= \frac{1}{m} \\ \phi &= 1 \\ b(\theta) &= \log \frac{1}{1-\pi} = \log(1 + \exp(\theta)) \\ c(y, \phi) &= \log \binom{m}{my} = \log \binom{1/\theta}{y/\theta}. \end{aligned}$$

Da $E(Y) = m \cdot \pi$ folgt

$$\mu = E(\tilde{Y}) = E(Y)/m = \pi = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Da eine Bernoulli-Zufallsvariable ein Spezialfall einer binomialverteilten Zufallsvariable ist, resultiert mit der kanonischen Linkfunktion $\eta = g(\pi) = \theta$ wie beim Bernoulli-Modell

der Logit-Link

$$\eta = \log \frac{\pi}{1-\pi} = \log \frac{\mu}{1-\mu} = \text{logit}(\mu)$$

und daher

$$\mu = \frac{\exp \eta}{1 + \exp \eta}.$$

Wie bereits im vorherigen Abschnitt beschrieben entspricht η dem linearen Prädiktor, das heißt $\eta = \boldsymbol{\beta}^T \boldsymbol{x}$ und $\boldsymbol{x}$ stellt den Vektor der erklärenden Variablen dar. Sind nur Faktoren im linearen Prädiktor enthalten, so spricht man von Logit-Modellen. Bei komplexen logistischen Modellen können sowohl Faktoren als auch Variablen enthalten sein.

A.4. Die Delta-Methode

Der folgende Abschnitt baut auf den Beschreibungen von Agresti (2002) und Casella & Berger (2002) auf. Man betrachtet eine Statistik $T_n = T(Y_1, \ldots, Y_n)$. Diese ist abhängig von der Anzahl n der Stichprobenelemente und wir nehmen an, dass sie sich mit wachsendem n normalverteilt verhält, das heißt T_n erfülle

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Außerdem sei $g(\cdot)$ eine zweimal differenzierbare Funktion im Punkt θ. Wir wollen die Grenzverteilung von $g(T_n)$ ermitteln. Dafür betrachten wir die Taylorentwicklung von $g(t)$ mit $g'(\theta) \neq 0$ und nehmen an, dass θ^* ein Punkt ist, der zwischen t und θ liegt.

$$\begin{aligned} g(t) &= g(\theta) + (t-\theta)g'(\theta) + (t-\theta)^2 \frac{g''(\theta^*)}{2} \\ &= g(\theta) + (t-\theta)g'(\theta) + O(\mid t - \theta \mid^2) \end{aligned}$$

Setzen wir für t die Statistik T_n ein, so erhalten wir

$$\sqrt{n}(g(T_n) - g(\theta)) = \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}O(\mid T_n - \theta \mid^2).$$

An dieser Stelle wollen wir die Notation $O_p(Z_n)$ nutzen. Als Abwandlung der gewöhnlichen O-Notation bezieht sie sich auf Folgen von Zufallsvariablen (p-probability). Mit

$O_p(Z_n)$ beschreiben wir eine Zufallsvariable, sodass für alle $\epsilon > 0$ ein K und ein $n_0 \in \mathbb{N}$ existiert, sodass $P(\mid O_p(Z_n)/Z_n \mid < K) > 1 - \epsilon \; \forall n > n_0$ gilt. Somit ergibt sich

$$\sqrt{n}O(\mid t - \theta \mid^2) = \sqrt{n}O\left(O_p\left(\frac{1}{n}\right)\right) = O_p\left(\frac{1}{n}\right)$$

und wir bekommen

$$\sqrt{n}(g(T_n) - g(\theta)) = \sqrt{n}(T_n - \theta)g'(\theta) + O_p\left(\frac{1}{\sqrt{n}}\right),$$

wobei $O_p\left(\frac{1}{\sqrt{n}}\right)$ für $n \to \infty$ vernachlässigbar ist. Daraus folgt, dass

$$\sqrt{n}(g(T_n) - g(\theta)) \sim \sqrt{n}(T_n - \theta)g'(\theta).$$

Da wir wissen, dass $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ gilt, folgt

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, \sigma^2(g'(\theta))^2).$$

Dies entspricht der Delta-Methode für asymptotische Verteilungen. Nimmt man an, dass $\sigma^2 = \sigma^2(\theta)$ und $g'(\theta)$ vom zu schätzenden Parameter θ abhängig sind, so ist die Varianz unbekannt. Als Schätzer von θ nutzen wir T_n. Sind die Funktionen $\sigma(\cdot)$ und $g'(\cdot)$ stetig bei θ, so ist $\sigma(T_n)g'(T_n)$ ein konsistenter Schätzer für $\sigma(\theta)g'(\theta)$. Um ein $(1-\alpha)100\%$-Konfidenzintervall für $g(\theta)$ anzugeben, nutzen wir daher aus, dass

$$\frac{\sqrt{n}(g(T_n) - g(\theta))}{\sigma(T_n) \mid g'(T_n) \mid} \xrightarrow{d} N(0,1)$$

und erhalten als Grenzen des Konfidenzintervalls

$$g(T_n) \pm z_{1-\alpha/2} \cdot \sigma(T_n)\frac{\mid g'(T_n) \mid}{\sqrt{n}}.$$

A.4.1. Die Delta-Methode für Funktionen von Zufallsvektoren

Die Delta-Methode kann auch auf Vektoren verallgemeinert werden. $\boldsymbol{T}_n$ sei nun also ein Vektor der Form $\boldsymbol{T}_n = (T_{1n}, \ldots, T_{pn})^T$, welcher asymptotisch multivariat normalverteilt ist mit Erwartungswert $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ und $p \times p$ Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}/n$. Wir nehmen an, dass das Differential der Funktion $g(t_1, \ldots, t_p)$ ausgewertet bei $\boldsymbol{\theta}$ nicht $\mathbf{0}$ ist. Wir schreiben $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$, wobei

$$\phi_j = \left.\frac{\partial g}{\partial t_j}\right|_{\boldsymbol{t}=\boldsymbol{\theta}}.$$

Ähnlich wie bei der eindimensionalen Delta-Methode ergibt sich für ein wachsendes n

$$\sqrt{n}(g(\boldsymbol{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\phi}^T \boldsymbol{\Sigma} \boldsymbol{\phi}),$$

was anders ausgedrückt bedeutet, dass sich $g(\boldsymbol{T}_n)$ für große n normalverteilt mit Erwartungswert $g(\boldsymbol{\theta})$ und Varianz $\boldsymbol{\phi}^T \boldsymbol{\Sigma} \boldsymbol{\phi}/n$ verhält. Für weitere Ausführungen sei hier auf Casella & Berger (2002) hingewiesen.

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis* (2. Aufl.). New Jersey: Wiley.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data* (2. Aufl.). New Jersey: Wiley.

Casella, G. & Berger, R. L. (2002). *Statistical Inference* (2. Aufl.). Californien: Wadsworth Group.

Christensen, R. H. B. (2012). *Analysis of ordinal data with cumulative link models - estimation with the R-package ordinal.*

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*, 109-142.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2. Aufl.). New York, London: Chapman and Hall.

Simonoff, J. S. (2003). *Analizing Categorical Data.* New York: Springer.

Tutz, G. (2000). *Die Analyse kategorialer Daten.* München, Wien: Oldenbourg Wissenschaftsverlag GmbH.

Tutz, G. (2012). *Regression for Categorical Data.* New York: Cambridge University Press.

Winship, C. & Mare, R. D. (1984). Regression models with ordninal variables. *American Sociological Review*, *42*, 109-142.

Yee, T. W. (2010). *The VGAM package for categorical data analysis.* Journal of Statistical Software.