Technische Universität München

Zentrum Mathematik

Gemeinsame Modellierung von Stornierungen und Abschlüssen in der Sachversicherung mithilfe des bivariaten Probit-Modells

Diplomarbeit

von

Sylvia Grain

Themensteller/in:	Prof. Dr. Claudia Czado
Betreuer:	Prof. Dr. Claudia Czado
	Christiane Belitz
Abgabetermin:	4. Mai 2005

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Garching, den 4. Mai 2005

Danksagung

Danken möchte ich der Abteilung PACS der Allianz Versicherungs-AG, die das von uns untersuchte Datenmaterial zur Verfügung gestellt hat, und vor allem Herrn Dr. Clemens Biller, der das Thema dieser Diplomarbeit angeregt und alle meine Fragen stets äußerst hilfsbereit beantwortet hat.

Ganz besonderer Dank gebührt Frau Prof. Dr. Claudia Czado für ihre eingehende und intensive Betreuung. Sie hatte immer ein offenes Ohr für meine Anliegen und motivierte mich stets aufs Neue. Herzlichen Dank auch an Christiane Belitz vom Institut für Statistik der Ludwig-Maximilians-Universität, die bei allen Problemen eine kompetente Ansprechpartnerin war und durch ihre wertvollen Anregungen und Impulse großen Anteil am Gelingen dieser Arbeit hatte.

Ein Dankeschön geht auch an Christoph Winter, Mathias Jais und Michael Schroll für den technischen Support und ihre Hilfe in allen Software- und Computerfragen und natürlich an meine Familie und Freunde für ihre Unterstützung.

S.G.

Inhaltsverzeichnis

1 Einleitung

In vielen Anwendungsbereichen dienen sogenannte Regressionsmodelle dazu, den Einfluss verschiedener Merkmale (bezeichnet als Kovariablen, Regressoren oder unabhängige Variablen) auf eine beobachtbare zufällige Größe (Ziel- oder abhängige Variable) zu bestimmen.

Häufig ist man dabei an erklärenden Modellen für binäre Zielgrößen interessiert, die nur zwei verschiedene Ausprägungen annehmen können. Beispielsweise untersucht man in den Wirtschaftswissenschaften das Entscheidungsverhalten von Kunden (Kauf oder Nichtkauf). Epidemiologische Studien dienen dazu, herauszufinden, welche Faktoren das Phänomen „Erkrankung“ bzw. „Nichterkrankung“ beeinflussen. In dieser Arbeit untersuchen wir das Verhalten von Versicherungskunden, die sich hinsichtlich der Stornierung bzw. Nichtstornierung eines laufenden Versicherungsvertrages sowie zwischen Abschluss bzw. Nichtabschluss eines neuen Vertrages zu entscheiden haben. Ziel ist es dabei, sowohl die wichtigsten Einflussfaktoren zu bestimmen als auch die Abhängigkeitsstruktur zwischen beiden Ereignissen zu analysieren. Die Identifikation hoch stornogefährdeter bzw. stark abschluss-affiner Kunden ermöglicht dem Versicherungsunternehmen die Entwicklung von zielgenau ausgerichteten Marketing-Kampagnen, die speziell auf solche Kundengruppen zugeschnitten sind. Des Weiteren versuchen wir die Frage zu klären, ob eine gemeinsame Modellierung besseren Einblick in das Gesamtverhalten des Kunden gewährt als die separate Modellierung von Stornierungen und Neuabschlüssen.

Im ersten Teil dieser Arbeit werden Möglichkeiten zur Modellierung einer einzelnen binären Zielvariablen erläutert. Dabei beschäftigen wir uns zunächst mit dem zugrunde liegenden Schwellenwertkonzept und dem allgemeinen theoretischen Hintergrund der Modelle. Danach stellen wir die häufigsten Modellformen vor. Die Anpassung solcher univariaten Modelle durch Maximum-Likelihood-Schätzung wird in Kapitel 3 beschrieben. In Kapitel 4 wenden wir uns der gemeinsamen Modellierung zweier binärer Zielgrößen zu. Wir zeigen, wie sich der den univariaten Modellen zugrunde liegende Ansatz auf den bivariaten Fall ausweiten lässt. Bei der Vorstellung verschiedener Modelltypen gehen wir besonders auf das von uns verwendete bivariate Probit-Modell und seine Eigenschaften ein. Die Maximum-Likelihood-Schätzung in diesem Modell wird in Kapitel 5 näher erläutert. Besonderes Augenmerk richten wir dabei auf Kriterien zu Existenz und Eindeutigkeit eines Maximum-Likelihood-Schätzers.

Eine anschauliche Möglichkeit zur Interpretation des im bivariaten Probit-Modells auftretenden Korrelationsparameters präsentieren wir in Kapitel 6. Dort führen wir den sogenannten Odds Ratio ein und zeigen, dass eine Schätzung der Korrelation mithilfe dieses Odds Ratios äquivalent ist zur Bestimmung der von Pearson (1901) entwickelten tetra-

chorischen Korrelation. Anschließend erläutern wir Strategien, die im Umgang mit sehr umfangreichem Datenmaterial hilfreich sein können. Speziell behandeln wir dabei einige Probleme, die sich oftmals bei der Modellanpassung ergeben und die auch bei den von uns betrachteten Daten aufgetreten sind. Ein eigenes Kapitel ist der Modellierung stetiger Kovariablen gewidmet. Dabei umreissen wir kurz verschiedene Möglichkeiten und gehen insbesondere auf die von uns verwendeten Polynomialsplines ein. Ein Überblick über verschiedene Arten zur Bewertung und zum Vergleich von Modellen schließt diesen Abschnitt ab.

Danach wenden wir uns der konkreten Modellentwicklung aus den von uns betrachteten Versicherungsdaten zu. Nach einer ausführlichen Datenbeschreibung und Exploration erläutern wir zunächst das schrittweise Vorgehen zur gesonderten Modellierung von Stornierungen und Abschlüssen. Aus diesen Ergebnissen erstellen wir danach bivariate Modelle. Besonders eingehend beschäftigen wir uns dabei mit der Interpretation und dem Vergleich der verschiedenen Modelle. Eine wichtige von uns gewonnene Erkenntnis ist hierbei, dass in bestimmten Kundengruppen eine Stornierung die Entscheidung für oder gegen einen Neuabschluss nicht negativ, sondern auf positive Art beeinflusst. Mit dem Ziel, solche unter Marketing-Gesichtspunkten besonders interessante Kunden zu identifizieren, untersuchen wir zum Abschluss dieser Arbeit einige Subpopulationen des Datensatzes auf die darin vorherrschende Korrelationsstruktur.

2 Univariate binäre Regressionsmodelle

In Kapitel 2 wird der theoretische Hintergrund für Regressionsmodelle binärer Daten bereitgestellt. Zunächst soll das allgemeine Konzept einer binären Regression erläutert werden. Die beiden wichtigsten und am häufigsten verwendeten Modelle werden gesondert betrachtet und einander vergleichend gegenübergestellt. Ein kurzer Überblick über alternative Möglichkeiten der Modellierung binärer Daten schließt das Kapitel ab.

Univariate binäre Regressionsmodelle werden in umfassender Form z.B. von Collett (1999) behandelt. Pindyck & Rubinfeld (1998) und Tutz (2000) geben einen guten Überblick über die verschiedenen Modelle. McCullagh & Nelder (1989) behandeln die Modellierung binärer Daten im Kontext der Theorie der generalisierten linearen Modelle.

2.1 Das Schwellenwertkonzept

Eine beobachtbare binäre Größe Y erklärt man häufig durch die Existenz einer zugrunde liegenden latenten (also nicht beobachtbaren), stetigen Variablen Z.
$Y = 1$ wird genau dann beobachtet, wenn die latente Variable Z einen bestimmten Grenz- oder Schwellenwert θ unterschreitet (Tutz (2000)). In der Literatur findet sich eine Vielzahl weiterer Modelle mit latenten Variablen, nicht nur für binäre Daten. Einen Überblick über verschiedenste Anwendungen geben z.B. Bartholomew & Knott (1999).

Für binäre Daten erlaubt dieses Konzept eine höchst anschauliche Interpretation des Ein- bzw. Nichteintretens eines Ereignisses. So kommt ein Schwellenwertmodell häufig in der Biometrie zur Anwendung, z.B. bei der Betrachtung der Ereignisse „Überleben“ ($Y = 1$) und „Tod“ ($Y = 0$) eines Versuchstieres in Abhängigkeit von der verabreichten Dosis eines bestimmten Giftstoffes. Die latente stetige Variable Z, die das Überleben bzw. Nicht-Überleben bestimmt, wird hier als die durch das Gift verursachte (nicht direkt beobachtbare) Schädigung des Organismus aufgefasst. Bleibt diese unterhalb einer bestimmten Schwelle, so überlebt das Tier; wird diese Schranke jedoch überschritten, tritt der Tod ein. Diese allgemein als „Dose-Response Modell“ bezeichneten Modelle stellen eine der ersten Anwendungen von binären Regressionsmethoden dar (siehe z.B. Finney (1973) oder Dobson (1990)).

Im Kontext eines binären Regressionsmodells möchte man den Einfluss von gewissen (stetigen oder diskreten) Kovariablen $x_1, \ldots, x_p$ auf die binäre Größe Y untersuchen und quantifizieren. Um den Zusammenhang zwischen Y und den Kovariablen herzustellen, nimmt man an, dass Z sich durch $x_1, \ldots, x_p$ und eine stetige Störgröße ϵ beschreiben lässt.

Man betrachtet folgendes Modell für Y gegeben $x_1, \ldots, x_p$:

$$Y = 1 \,|\, x_1, \ldots, x_p \Longleftrightarrow Z = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + \epsilon = \boldsymbol{\alpha}'\boldsymbol{x} + \epsilon \leq \theta, \tag{2.1}$$

wobei $\boldsymbol{\alpha} := (\alpha_0, \alpha_1, \ldots, \alpha_p)'$ den zu schätzenden Parametervektor des Modells und $\boldsymbol{x} := (1, x_1, x_2, \ldots, x_p)'$ den Kovariablenvektor zur Beobachtung Y bezeichnen.

Sei nun F die Verteilungsfunktion der Störvariablen ϵ. Dann lässt sich die Erfolgswahrscheinlichkeit $P(Y = 1|\boldsymbol{x})$ in Abhängigkeit von den Kovariablen ausdrücken als

$$P(Y = 1|\boldsymbol{x}) = P(Z \leq \theta) = P(\boldsymbol{\alpha}'\boldsymbol{x} + \epsilon \leq \theta) = P(\epsilon \leq \theta - \boldsymbol{\alpha}'\boldsymbol{x}) = F(\theta - \boldsymbol{\alpha}'\boldsymbol{x}).$$

Mit $\beta_0 := \theta - \alpha_0,\ \beta_1 := -\alpha_1, \ldots, \beta_p := -\alpha_p$ und $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots \beta_p)'$ ergibt sich schließlich der Zusammenhang

$$P(Y = 1|\boldsymbol{x}) = F(\boldsymbol{\beta}'\boldsymbol{x}) \tag{2.2}$$

bzw.

$$F^{-1}\left[P(Y = 1|\boldsymbol{x})\right] = \boldsymbol{\beta}'\boldsymbol{x}.$$

Das Skalarprodukt $\boldsymbol{\beta}'\boldsymbol{x}$ wird als *linearer Prädiktor* bezeichnet. Die Funktion F^{-1}, die die Verbindung zwischen dem linearen Prädiktor und $P(Y = 1|\boldsymbol{x})$ herstellt, nennt man *Linkfunktion*.

Es ist sinnvoll, den Zusammenhang zwischen der Erfolgswahrscheinlichkeit $P(Y = 1|\boldsymbol{x})$ und den Kovariablen über eine Verteilungsfunktion zu definieren, da $P(Y = 1|\boldsymbol{x})$ auf das Intervall $[0, 1]$ beschränkt ist, während es für $\boldsymbol{\beta}'\boldsymbol{x}$ keine Einschränkung gibt und prinzipiell alle Werte zwischen $]-\infty, +\infty[$ angenommen werden können.

Setzt man einige wünschenswerte Eigenschaften von F wie Stetigkeit, Differenzierbarkeit und Monotonie voraus, so bildet F nicht nur die gesamte reelle Achse auf $[0, 1]$ ab, sondern gewährleistet zudem noch eine gute Interpretierbarkeit des Modells, da der Einfluss einer (stetigen) Kovariablen auf $P(Y = 1|\boldsymbol{x})$ aus (2.2) unmittelbar ersichtlich ist:

Es gilt

$$\frac{\partial}{\partial x_j} P(Y = 1|\boldsymbol{x}) = \underbrace{\frac{\partial F(\boldsymbol{\beta}'\boldsymbol{x})}{\partial x_j}}_{\geq 0} \cdot \beta_j\ , \quad j = 1, \ldots, p. \tag{2.3}$$

Für $\beta_j > 0$ bedeutet ein steigender Wert von x_j aufgrund des monotonen Wachstums von F also ein Zunehmen dieser Erfolgswahrscheinlichkeit; für $\beta_j < 0$ sinkt sie dagegen bei wachsenden Werten von x_j ab, während $P(Y = 0|\boldsymbol{x}) = 1 - F(\boldsymbol{\beta}'\boldsymbol{x})$ ansteigt.

Bei obiger Modellformulierung gilt es zu beachten, dass die beiden Parameter θ und α_0 nicht gleichzeitig identifiziert werden können:

Betrachtet man statt θ eine andere Schranke $\tilde{\theta} = \theta + \delta$ und parameterisiert man die latente Variable Z in (2.1) mit $\tilde{\alpha_0} = \alpha_0 + \delta$ anstatt mit α_0, so erhält man wieder Modell (2.2),

da die Differenz der beiden neuen Parameter unverändert bleibt: $\tilde{\theta} - \tilde{\alpha_0} = \theta - \alpha_0 = \beta_0$. Als Konsequenz lässt sich also lediglich der Abstand zwischen den beiden Parametern bestimmen, wodurch Formulierung (2.2) gerechtfertigt ist. O.B.d.A. könnte man deshalb auch $\theta = 0$ annehmen. Damit vereinfacht sich die Notation des Modells etwas, insbesondere gilt $\boldsymbol{\beta} = -\boldsymbol{\alpha}$.

Um das Entscheidungsverhalten eines Individuums zu motivieren, das zwischen zwei Alternativen zu wählen hat, findet man vor allem im Bereich der Ökonometrie auch häufig das sogenannte *Nutzenmaximierungsmodell* (Tutz (2000)).

Alternative 1 wird demnach genau dann gewählt, wenn deren individueller Nutzen (ausgedrückt über eine Zufallsvariable U_1) den von Alternative 2 übersteigt, also wenn $U_2 \leq U_1$ gilt. Diese Formulierung impliziert ein Schwellenwertmodell mit der latenten Variablen $Z := U_2 - U_1$ und dem Schwellenwert 0. Für Alternative 1 entscheidet sich ein Individuum also genau dann, wenn $Z \leq 0$. Bei $Z > 0$ wird dagegen Alternative 2 gewählt.

Auch die in dieser Arbeit betrachteten Versicherungsdaten legen eine solche Interpretation nahe: Den beiden Zielvariablen `storno` und `abschluss` liegt eine Entscheidung des Versicherungsnehmers zwischen Tätigung und Nichttätigung eines Stornos bzw. eines Neuabschlusses zugrunde. Der Kunde hat dabei abzuwägen, welches Vorgehen für ihn nützlicher ist: eine Stornierung bzw. ein Neuabschluss (Alternative 1), oder die Beibehaltung des Status Quo (Alternative 2).

Ausgehend von dem in diesem Abschnitt beschriebenen Grundkonzept ergeben sich aus der Verwendung unterschiedlicher Verteilungsfunktionen F verschiedene Modellierungsmöglichkeiten binärer Daten. Meist geht man von einer Standardnormalverteilung oder einer logistischen Verteilung aus. Diese beiden speziellen Modelle werden im Folgenden näher betrachtet.

2.2 Das Probit-Modell

Beim univariaten Probit-Modell wird die Verteilung F der Störgröße ϵ als Standardnormalverteilung spezifiziert mit Dichte $\phi(x) = \frac{1}{2\pi}\, e^{-\frac{1}{2}x^2}$ und Verteilungsfunktion $\Phi(x)$.

Übertragen auf n unabhängige Beobachtungen eines binären Responses $Y_i,\ i = 1, \ldots, n$, lautet das Modell somit

$$Y_i = 1|\boldsymbol{x} \Longleftrightarrow Z_i = \boldsymbol{\alpha}'\boldsymbol{x_i} + \epsilon_i \leq \theta$$

mit

$$\epsilon_i \overset{iid}{\sim} N(0,1), \quad i = 1, \ldots, n\ ,$$

wobei $\boldsymbol{x_i} := (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ den Kovariablenvektor zur i-ten Beobachtung bezeichnet.

Für die Erfolgswahrscheinlichkeit $\pi_1(\boldsymbol{x_i}) := P(Y_i = 1|\boldsymbol{x_i})$ und die Wahrscheinlichkeit eines Misserfolgs $\pi_0(\boldsymbol{x_i}) := P(Y_i = 0|\boldsymbol{x_i})$ ergibt sich nach (2.2) in diesem Spezialfall

$$\pi_1(\boldsymbol{x_i}) = P(\epsilon_i \leq \boldsymbol{\beta'x_i}) = \Phi(\boldsymbol{\beta'x_i}) \tag{2.4}$$
$$\pi_0(\boldsymbol{x_i}) = P(\epsilon_i > \boldsymbol{\beta'x_i}) = 1 - \Phi(\boldsymbol{\beta'x_i}) \tag{2.5}$$

für $i = 1, \ldots, n$.

Die Linkfunktion im Probit-Modell ist Φ^{-1}, und es gilt

$$\Phi^{-1}[\pi_1(\boldsymbol{x_i})] = \boldsymbol{\beta'x_i}.$$

Die transformierte Wahrscheinlichkeit $\Phi^{-1}[\pi_1(\boldsymbol{x_i})]$ wird auch *Probit* genannt.

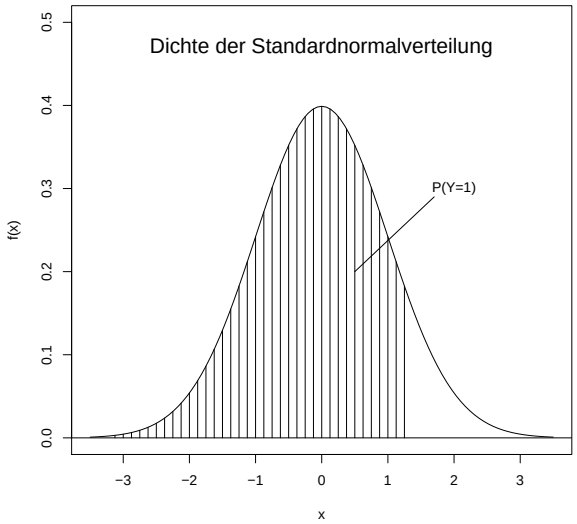


Abbildung 2.1: $\pi_1(x) = P(Y = 1|x)$ *an der Stelle* $x = 1.25$ *für* $\beta = 1$

Abbildung 2.1 verdeutlicht den Zusammenhang zwischen Dichtefunktion und Erfolgswahrscheinlichkeit für nur eine stetige Einflussgröße x und $\beta = 1$. $\pi_1(x)$ ergibt sich als Integral über die Dichte zwischen den Grenzen $-\infty$ und x. Für $x = 1.25$ ist dies in Abb. 2.1 exemplarisch als schraffierte Fläche gekennzeichnet.

Ein Nachteil dieses Modells ist die Integraldarstellung der Erfolgswahrscheinlichkeit, wodurch sich die Interpretation des Einflusses einzelner Kovariablen schwierig gestaltet. Zwar gilt für stetige Kovariablen nach (2.3)

$$\frac{\partial}{\partial x_{ij}} P(Y_i = 1|\boldsymbol{x_i}) = \phi(\boldsymbol{\beta'x_i}) \cdot \beta_j \ , \quad j = 1, \ldots, p,$$

woraus der marginale Effekt von x_{ij} leicht ersichtlich ist. Für kategoriale Kovariablen in

Dummy-Kodierung ist es jedoch schwierig, deren Einfluss zu quantifizieren:

Sei x eine kategoriale Variable mit M Kategorien. Dann gilt

$$P(Y = 1|x) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_{M-1} x_{M-1}),$$

wobei Kategorie M als Referenzkategorie verwendet wird und

$$x_i := \begin{cases} 1, & \text{falls } x = i \\ 0, & \text{falls } x \neq i \end{cases} \quad , \; i = 1, \ldots, M-1 \tag{2.6}$$

die Dummyvariablen bezeichnen.

Der Unterschied in den Erfolgswahrscheinlichkeiten beim Übergang von Kategorie M zu Kategorie $I, I \in \{1, \ldots, M-1\}$ beträgt

$$\Phi(\beta_0 + \beta_I) - \Phi(\beta_0)$$

und ist schwer zu interpretieren.

Das sogenannte Logit-Modell dagegen gewährleistet aufgrund der besonderen Struktur der Linkfunktion eine einfache Interpretierbarkeit aller Kovariablen.

2.3 Das Logit-Modell

Für das Logit-Modell geht man von einer logistischen Verteilung mit Dichte bzw. Verteilungsfunktion

$$f(x) = \frac{\exp(x)}{[1 + \exp(x)]^2} \quad , \quad F(x) = \frac{\exp(x)}{1 + \exp(x)} \tag{2.7}$$

aus. Mit (2.7) gilt nach (2.2)

$$\pi_1(\boldsymbol{x_i}) = F(\boldsymbol{\beta}'\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x_i})}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x_i}),} \tag{2.8}$$

woraus sich für die Linkfunktion im Logit-Modell

$$F^{-1}[\pi_1(\boldsymbol{x_i})] = \log \frac{\pi_1(\boldsymbol{x_i})}{1 - \pi_1(\boldsymbol{x_i})} = \boldsymbol{\beta}'\boldsymbol{x_i} \tag{2.9}$$

ergibt.

Abbildung 2.2 verdeutlicht den Zusammenhang zwischen logistischer Dichte und Erfolgswahrscheinlichkeit für nur eine stetige Kovariable x und $\beta = 1$. $\pi_1(x)$ ergibt sich als Integral über die Dichte zwischen den Grenzen $-\infty$ und x. Für $x = 1.75$ ist dies in Abb. 2.2 exemplarisch als schraffierte Fläche gekennzeichnet.

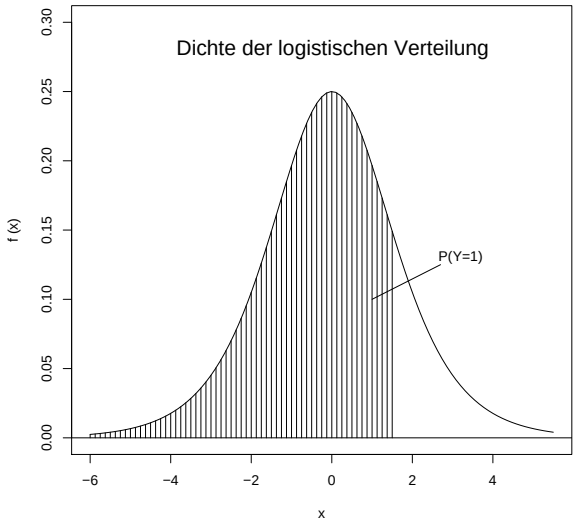


Abbildung 2.2: $\pi_1(x) = P(Y = 1|x)$ *an der Stelle* $x = 1.75$ *für* $\beta = 1$

Der in der Linkfunktion auftretende Bruch (2.9) stellt das Verhältnis der Wahrscheinlichkeiten von Erfolg zu Misserfolg (bzw. von Eintreten zu Nichteintreten eines Ereignisses) dar:

$$\frac{\pi_1(\boldsymbol{x_i})}{1-\pi_1(\boldsymbol{x_i})} = \frac{\pi_1(\boldsymbol{x_i})}{\pi_0(\boldsymbol{x_i})} = \frac{P(Y_i = 1|\boldsymbol{x_i})}{P(Y_i = 0|\boldsymbol{x_i})}$$

Daher wird dieser Ausdruck als Angabe der *Chancen* interpretiert: Ist beispielsweise $\pi_1(\boldsymbol{x_i}) = \frac{4}{5}$ und somit $\pi_0(\boldsymbol{x_i}) = \frac{1}{5}$, so stehen die Chancen 4 zu 1, einen Erfolg bzw. Misserfolg zu haben. Alternativ kann man auch die logarithmierten Chancen betrachten, die als *Logits* bezeichnet werden. Insbesondere gilt $\text{Logit}(\boldsymbol{x_i}) := \log\left(\frac{\pi_1(\boldsymbol{x_i})}{1-\pi_1(\boldsymbol{x_i})}\right)$.

Da der lineare Prädiktor nach (2.9) gerade die logarithmierten Chancen beschreibt, erlaubt das Logit-Modell – im Gegensatz zum Probit-Modell – eine relativ einfache Interpretation der Kovariablen:

Bei nur einer stetigen Kovariablen x und Interzept β_0 gilt

$$\pi_1(x) = \frac{\exp(\beta_0 + \beta x)}{1 + \exp(\beta_0 + \beta x)} \ ,$$

$$\frac{\pi_1(x)}{1-\pi_1(x)} = \left(\frac{\exp(\beta_0 + \beta x)}{1 + \exp(\beta_0 + \beta x)}\right) \Big/ \left(\frac{1}{1 + \exp(\beta_0 + \beta x)}\right) = \exp(\beta_0 + \beta x)$$

und somit $\quad \text{Logit}(x) = \beta_0 + \beta x.$

Die Veränderung der Logits bei Anwachsen von x um eine Einheit auf $x+1$ lässt sich damit wie folgt quantifizieren:

$$\text{Logit}(x+1) - \text{Logit}(x) = \beta_0 + \beta(x+1) - (\beta_0 + \beta x) = \beta.$$

Die Veränderung der Logits ist also konstant und hängt nicht vom Anfangswert x ab. Für die Veränderung des Verhältnisses der Chancen ergibt sich entsprechend

$$\frac{\left(\frac{\pi_1(x+1)}{1-\pi_1(x+1)}\right)}{\left(\frac{\pi_1(x)}{1-\pi_1(x)}\right)} = \frac{\exp(\beta_0 + \beta(x+1))}{\exp(\beta_0 + \beta x)} = e^{\beta}.$$

Bei einer kategorialen Einflussgröße x mit M verschiedenen Ausprägungen $\{1, \ldots, M\}$ gilt in Dummy-Kodierung

$$\pi_1(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{M-1} x_{M-1})}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{M-1} x_{M-1})},$$

$$\frac{\pi_1(x)}{1-\pi_1(x)} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{M-1} x_{M-1})$$

$$\text{und somit} \quad \text{Logit}(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{M-1} x_{M-1}\ ,$$

wobei Kategorie M als Referenzkategorie verwendet wird und $x_i,\ i = 1, \ldots, M-1$ wieder die Dummyvariablen in (2.6) bezeichnen.

Die Veränderung der Logits beim Übergang von der Referenz- zu einer anderen Kategorie I ist daraus sofort ersichtlich:

$$\text{Logit}(I) - \text{Logit}(M) = \beta_0 + \beta_I - \beta_0 = \beta_I, \quad I \in \{1, \ldots, M-1\}.$$

Für die Veränderung der Chancen ergibt sich entsprechend

$$\frac{\left(\frac{\pi_1(I)}{1-\pi_1(I)}\right)}{\left(\frac{\pi_1(M)}{1-\pi_1(M)}\right)} = \frac{\exp(\beta_0 + \beta_I)}{\exp(\beta_0)} = e^{\beta_I}.$$

Abgesehen von der besseren Interpretierbarkeit der Kovariablen gleichen sich Probit- und Logit-Modell stark und liefern auch ähnliche Schätzergebnisse. Die Gründe hierfür werden im folgenden Abschnitt dargelegt.

2.4 Vergleich von Probit- und Logit-Modell

Wie Abbildung 2.3 zeigt, sind die Dichten und Verteilungsfunktionen von Probit- und Logit-Modell grundsätzlich ähnlich, jedoch weist die logistische Dichte „schwerere“ Tails auf als die der Standardnormalverteilung.

Beide Verteilungen sind symmetrisch um 0, d.h. es gilt $1 - F(x) = F(-x)$. Betrachtet man also $P(Y = 0|\,\boldsymbol{x})$ anstatt von $P(Y = 1|\,\boldsymbol{x})$, so erhält man ein Regressionsmodell mit Parametern $-\boldsymbol{\beta}$ anstatt $\boldsymbol{\beta}$, da $P(Y = 0|\,\boldsymbol{x}) = 1 - P(Y = 1|\boldsymbol{x}) = 1 - F(\boldsymbol{\beta}'\boldsymbol{x}) = F(-\boldsymbol{\beta}'\boldsymbol{x})$.

Demzufolge spielt es bei symmetrischen Verteilungen für die Modellierung zunächst keine Rolle, wie man die Ereignisse „Erfolg“ und „Nichterfolg“ definiert, da man stets dasselbe Grundmodell erhält, das sich nur in der Parameterisierung unterscheidet. Ein Schwellenwertansatz ist mit $P(Y = 0) = P(Z > \theta)$ also grundsätzlich auch für das Überschreiten eines Schwellenwertes möglich. Speziell für den Fall $\theta = 0$ ergibt sich unmittelbar $P(Y = 0|\,\boldsymbol{x}) = P(Z > 0) = F(-\boldsymbol{\beta}'\boldsymbol{x}) = F(\boldsymbol{\alpha}'\boldsymbol{x})$.

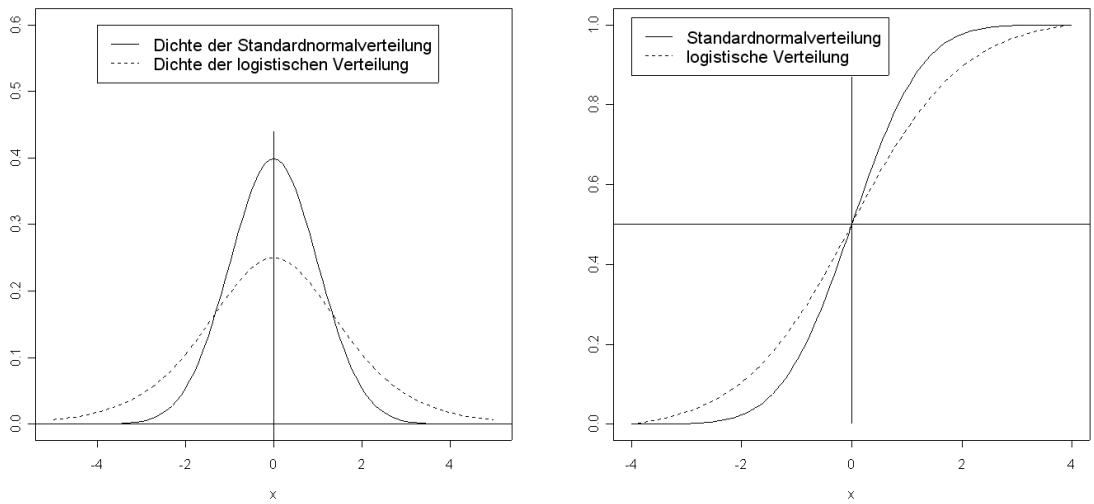


Abbildung 2.3: *Dichten bzw. Verteilungsfunktionen von Probit- und Logit-Modell*

Des Weiteren ist zu beachten, dass die logistische Verteilung eine Varianz von $\pi^2/3$ aufweist, während sie bei der Standardnormalverteilung 1 beträgt. Die geschätzten Parameter beider Modelle sind also nicht direkt miteinander vergleichbar. Um dies zu ermöglichen, ist eine Standardisierung der logistischen Verteilung nötig. Die Parameter $\tilde{\boldsymbol{\beta}}$ im standardisierten Modell unterscheiden sich von jenen des nicht standardisierten Modells um den Faktor $\pi/\sqrt{3}$, also $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}\,\pi/\sqrt{3}$.

Dieser Zusammenhang ist folgendermaßen ersichtlich: Im Logit-Modell gilt mit der bisher verwendeten Notation

$$\pi_1(\boldsymbol{x}) = P(Y = 1|\,\boldsymbol{x}) = F(\boldsymbol{\beta}'\boldsymbol{x}) = P(Z \leq \boldsymbol{\beta}'\boldsymbol{x}) = P\left(\frac{Z}{\pi/\sqrt{3}} \leq \frac{\boldsymbol{\beta}'\boldsymbol{x}}{\pi/\sqrt{3}}\right) = \widetilde{F}\left(\frac{\boldsymbol{\beta}'\boldsymbol{x}}{\pi/\sqrt{3}}\right), \tag{2.10}$$

wobei F die logistische Verteilungsfunktion der latenten Variablen Z und $\widetilde{F}$ die Verteilungsfunktion der standardisierten latenten Variablen $\frac{Z}{\pi/\sqrt{3}}$ bezeichnet.

Setze $\frac{\sqrt{3}}{\pi}\boldsymbol{\beta} := \tilde{\boldsymbol{\beta}}$, dann folgt aus (2.10)

$$\pi_1(\boldsymbol{x}) = F(\boldsymbol{\beta}'\boldsymbol{x}) = \widetilde{F}(\tilde{\boldsymbol{\beta}}'\boldsymbol{x}).$$

Das nicht-standardisierte Logit-Modell mit Parametervektor $\boldsymbol{\beta}$ ist also äquivalent zu einem standardisierten Modell mit den Parametern $\tilde{\boldsymbol{\beta}}$.

Abbildung 2.4 zeigt, dass die Verteilungsfunktionen des Probit- und des standardisierten Logit-Modells ähnlich verlaufen, so dass beide Modelle auch ähnliche Schätzergebnisse liefern. Damit sind die Parameterschätzer des nicht-standardisierten Logit-Modells also in etwa um den Faktor $\pi/\sqrt{3}$ größer als die des Probit-Modells.

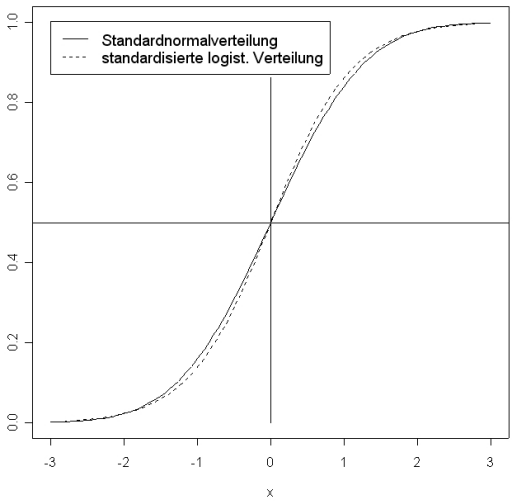


Abbildung 2.4: *Verteilungsfunktionen von Probit- und standardisiertem Logit-Modell*

Eine Unterscheidung zwischen Probit- und Logit-Modell anhand eines Goodness-of-Fit Kriteriums ist daher äußerst schwierig, wie z.B. Chambers & Cox (1967) näher ausführen.

2.5 Alternative Modelle

Neben der Standardnormalverteilung und der logistischen Verteilung, die am häufigsten verwendet werden, gibt es noch eine Vielzahl weiterer Verteilungsfunktionen, die zur Modellierung herangezogen werden können. Einige davon werden in diesem Abschnitt kurz vorgestellt.

- **Lineares Wahrscheinlichkeits-Modell**:

  Beim linearen Wahrscheinlichkeits-Modell wird eine Gleichverteilung auf dem Intervall $[0\,,1]$ verwendet, mit Verteilungsfunktion

$$F(x) = \begin{cases} 0, & \text{falls } x < 0 \\ x, & \text{falls } 0 \leq x \leq 1 \\ 1, & \text{falls } x > 1\ . \end{cases}$$

  Setzt man voraus, dass $\boldsymbol{\beta}'\boldsymbol{x_i}$ stets in $[0\,,\,1]$ liegt, so ergibt sich ein direkter, linearer Zusammenhang zwischen $\pi_1(\boldsymbol{x_i})$ und dem Prädiktor:

$$\pi_1(\boldsymbol{x_i}) = \boldsymbol{\beta}'\boldsymbol{x_i}, \quad i = 1, \ldots, n.$$

  Dies stellt eines der ersten Regressionsmodelle für binäre Daten dar. Für eine eingehendere Behandlung, auch anhand eines Beispiels, siehe z.B. Pindyck & Rubinfeld (1998). Da die getroffene Einschränkung $\boldsymbol{\beta}'\boldsymbol{x_i} \in [0\,,\,1]$ für alle $i = 1, \ldots, n$ in der Praxis aber nur schwer einzuhalten ist, ist man inzwischen von der Verwendung dieses Modells abgekommen.

- **Minimum-Extremwertmodell:**

  Beim Minimum-Extremwertmodell verwendet man

$$F(x) = 1 - \exp[-\exp(x)]$$

  als Verteilung (auch als Gompertz-Verteilung bekannt). Dieses Modell wird auch als komplementäres log-log Modell bezeichnet, da für die Linkfunktion $F^{-1}[\pi_1(\boldsymbol{x_i})] = \log[-\log\big(1 - \pi_1(\boldsymbol{x_i})\big)] = \boldsymbol{\beta}'\boldsymbol{x_i}$ gilt.

- **Exponentialmodell:**

  Beim Exponentialmodell wählt man als Verteilungsfunktion die der Exponentialverteilung mit Parameter $\lambda = 1$,

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

  In Tutz (2000) findet sich hierfür eine spezielle Motivation über ein Wartezeitmodell. Des Weiteren lässt sich dieses Modell auch über die Reduktion einer Poisson-verteilten Beobachtung Y auf eine dichotome Variable Y^* erklären mit $Y^* = 0$ für $Y = 0$, und $Y^* = 1$ sonst.

3 Parameterschätzung im univariaten Modell

In diesem Kapitel wird dargestellt, wie sich binäre Regressionsmodelle in den Kontext der Theorie der generalisierten linearen Modelle einfügen. Die GLM-Darstellung der Bernoulli- bzw. Binomialdichten wird zur Herleitung der Log-Likelihood und der Score-Gleichungen verwendet. Anschließend erläutern wir den für GLMs standardmäßig verwendeten Schätzalgorithmus für den Spezialfall der Binomialverteilung. Eine Einführung in die Theorie der generalisierten linaren Modelle geben z.B. Dobson (1990) oder Fahrmeir & Tutz (1994). Eine umfassende Diskussion der GLMs findet sich in McCullagh & Nelder (1989).

3.1 Binäre Regression als generalisiertes lineares Modell

Die GLM-Theorie behandelt Regressionsmodelle für Daten mit Response-Dichten der Form

$$f(y, \xi, \phi) = \exp\left[\frac{y\,\xi - b(\xi)}{a(\phi)} + c(y, \phi)\right] \tag{3.1}$$

mit verteilungsspezifischem Parameter ξ, Dispersionsparameter ϕ sowie bekannten Funktionen $a(\cdot), b(\cdot)$ und $c(\cdot)$. Dazu gehört die Normalverteilungsdichte genauso wie die Poisson- oder die Gammadichte. Auch die Bernoulli- und die Binomialdichte mit Erfolgswahrscheinlichkeit π fallen in diese Klasse:

Sei $Y \sim \text{Bernoulli}(\pi)$, dann gilt

$$\begin{aligned} P(Y = y) &= \pi^{\,y}(1-\pi)^{1-y} = \exp\left[y\ \log\pi + (1-y)\log(1-\pi)\right] \\ &= \exp\Big[y\ \log\Big(\frac{\pi}{1-\pi}\Big) + \log(1-\pi)\Big]. \end{aligned} \tag{3.2}$$

Hierbei ist $\xi = \log\big(\frac{\pi}{1-\pi}\big)$, $a \equiv 1$, $b = -\log(1-\pi) = \log(1+e^{\xi})$ und $c \equiv 0$.

Für $Y \sim \text{Bin}(n, \pi)$ gilt

$$\begin{aligned} P(Y = y) &= \binom{n}{y}\pi^{\,y}(1-\pi)^{n-y} = \exp\Big[y\ \log\pi + (n-y)\log(1-\pi) + \log\binom{n}{y}\Big] \\ &= \exp\Big[y\ \log\Big(\frac{\pi}{1-\pi}\Big) + n\log(1-\pi) + \log\binom{n}{y}\Big] \\ &= \exp\left[\frac{y/n\ \log\big(\frac{\pi}{1-\pi}\big) + \log(1-\pi)}{1/n} + \log\binom{n}{y}\right]. \end{aligned} \tag{3.3}$$

Eine GLM-Dichte ergibt sich demnach nur für $\widetilde{Y} := Y/n \sim \frac{\mathrm{Bin}(n,\pi)}{n}$, also für die relative Häufigkeit des Erfolges. Dabei ist
$\xi = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = n$, $a(\phi) = 1/n$, $b = -\log(1-\pi) = \log(1+e^{\xi})$ und $c(\phi, y) = \binom{n}{y}$.

Die in McCullagh & Nelder (1989) für generalisierte lineare Modelle entwickelten Likelihoodgleichungen und der dort beschriebene Schätzalgorithmus werden nun speziell für binäre Daten betrachtet.

3.2 Log-Likelihood und Scoregleichungen

Die Log-Likelihood für n unabhängige binäre Beobachtungen Y_i mit jeweils zugehörigem Kovariablenvektor $\boldsymbol{x_i}$ und Erfolgswahrscheinlichkeit $\pi_1(\boldsymbol{x_i})$, $i = 1, \ldots, n$, ergibt sich aus der GLM-Darstellung (3.2) einfach als

$$l(\boldsymbol{\beta}) = \log \prod_{i=1}^{n} P(Y_i = y_i | \boldsymbol{x_i}) = \sum_{i=1}^{n} \left[y_i \ \log\left(\frac{\pi_1(\boldsymbol{x_i})}{1-\pi_1(\boldsymbol{x_i})} \right) + \log\big(1 - \pi_1(\boldsymbol{x_i})\big) \right]$$

mit $\boldsymbol{\pi_1} := \big(\pi_1(\boldsymbol{x_1}), \pi_1(\boldsymbol{x_2}), \ldots, \pi_1(\boldsymbol{x_n})\big)'$ und $y_i \in \{0, 1\}$, $i = 1, \ldots, n$. Dabei hängt $l(\boldsymbol{\beta})$ nur durch $\pi_1(\boldsymbol{x_i})$ von $\boldsymbol{\beta}$ ab.

Im Kontext eines Regressionsmodells mit kategorialen Kovariablen bietet es sich jedoch an, die n Einzelbeobachtungen zu gruppieren, d.h. Beobachtungen, die denselben Kovariablenvektor $\boldsymbol{x_i}$, also auch dieselbe Erfolgswahrscheinlichkeit $\pi_1(\boldsymbol{x_i})$ aufweisen, zusammenzufassen. Auf diese Weise entstehen $k \leq n$ verschiedene Gruppen.

Sei n_j, $j = 1, \ldots, k$ die Gesamtanzahl von Beobachtungen in der j-ten Gruppe. Es gilt $\sum_{j=1}^{k} n_j = n$. Bezeichne ferner Y_j^* die Anzahl von Beobachtungen in der j-ten Gruppe mit $Y_i = 1$. Dann sind die Y_j^* jeweils binomialverteilt mit Erfolgswahrscheinlichkeit $\pi_1(\boldsymbol{x_j})$ und n_j Versuchen, $j = 1, \ldots, k$. Für die Log-Likelihood gilt daher nach (3.3)

$$l(\boldsymbol{\beta}) = \log \prod_{j=1}^{k} P(Y_j^* = y_j | \boldsymbol{x_j}) = \sum_{j=1}^{k} \left[\frac{(y_j/n_j) \ \log\left(\frac{\pi_1(\boldsymbol{x_j})}{1-\pi_1(\boldsymbol{x_j})}\right) + \log\big(1 - \pi_1(\boldsymbol{x_j})\big)}{1/n_j} + \log \binom{n_j}{y_j} \right] \tag{3.4}$$

mit $\boldsymbol{\pi_1} := \big(\pi_1(\boldsymbol{x_1}), \pi_1(\boldsymbol{x_2}), \ldots, \pi_1(\boldsymbol{x_k})\big)'$ und $y_j \in \{0, \ldots, n_j\}$, $j = 1, \ldots, k$.

Der Maximum-Likelihood-Schätzer für gruppierte Daten ist jener Parametervektor $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, der $l(\boldsymbol{\beta})$ in (3.4) maximiert.

Der gesuchte Vektor $\hat{\boldsymbol{\beta}}$ muss also notwendigerweise

$$\left. \frac{\partial l}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = 0 \tag{3.5}$$

erfüllen. Es gilt

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{j=1}^{k}\left[y_j\,\frac{\partial}{\partial \pi_1}\cdot\log\left(\frac{\pi_1(\boldsymbol{x_j})}{1-\pi_1(\boldsymbol{x_j})}\right)+n_j\cdot\frac{\partial}{\partial \pi_1}\log[1-\pi_1(\boldsymbol{x_j})]\right]\cdot\frac{\partial \pi_1}{\partial \boldsymbol{\beta}}\\ &= \sum_{j=1}^{k}\left[\frac{y_j-n_j\,\pi_1(\boldsymbol{x_j})}{\pi_1(\boldsymbol{x_j})[1-\pi_1(\boldsymbol{x_j})]}\right]\cdot\frac{\partial \pi_1}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{k}\left[\frac{y_j-\mathbb{E}(Y_j)}{\pi_1(\boldsymbol{x_j})[1-\pi_1(\boldsymbol{x_j})]}\right]\cdot\frac{\partial \pi_1}{\partial \boldsymbol{\beta}}. \qquad (3.6)\end{aligned}$$

Setzt man (3.6), die sogenannten Scoregleichungen, gleich Null, so erhält man ein nichtlineares Gleichungssystem, das nur iterativ über ein numerisches Verfahren zu lösen ist. Die Maximum-Likelihood-Schätzer der Parameter im univariaten Probit- und auch Logit-Modell bestimmt man im Kontext der Theorie der generalisierten linearen Modelle (GLM) unter Verwendung des sogenannten IWLS-Algorithmus, den wir im folgenden Abschnitt vorstellen.

3.3 Der IWLS-Algorithmus

Im Allgemeinen werden die Scoregleichungen mittels der Fisher-Scoring Methode (einer speziellen Form des bekannten Newton-Verfahrens zur Minimierung von Funktionen) gelöst. Für Dichten der Form (3.1) stellt man jedoch fest, dass dieses Vorgehen äquivalent zu einer iterativen gewichteten Kleinste-Quadrate Schätzung (*Iterative Weighted Least Squares*) ist. Dabei ist die Zielvariable eine linearisierte Form der Linkfunktion; als Gewichte werden Funktionen der angepassten Werte benützt.

Iterativ ist dieses Vorgehen deshalb, weil sowohl die Zielvariable als auch die Gewichte von den angepassten Werten abhängen, für die es jeweils nur eine aktuelle Schätzung gibt. Ausgehend von diesen Werten bestimmt man neue Parameterschätzer, aus denen sich wiederum eine neue Zielvariable und neue Gewichte ergeben.

Für k unabhängig binomialverteilte Beobachtungen $Y_j \sim \mathrm{Bin}(n_j, \pi_1(\boldsymbol{x_j}))$, $j = 1, \ldots, k$ stellt sich der Prozess folgendermaßen dar:

Sei $\hat{\boldsymbol{\beta}}_0$ der aktuelle Parameterschätzer, $\hat{\eta}^{0j} = \hat{\boldsymbol{\beta}}_0^{'}\boldsymbol{x_j}$ die aktuelle Schätzung des linearen Prädiktors der j-ten Beobachtung und $\hat{\pi}_1^{0j} = F(\hat{\boldsymbol{\beta}}_0^{'}\boldsymbol{x_j})$ bezeichne die aktuelle Schätzung der Erfolgswahrscheinlichkeit der j-ten Beobachtung, $j = 1, \ldots, k$.

- **Schritt 1:**

  Berechne die momentanen Zielvariable $\boldsymbol{z^0} := (z_1^0, z_2^0, \ldots, z_k^0)'$ mit

$$z_j^0 := \hat{\eta}_j^0 + (y_j/n_j - \hat{\pi}_1^{0j})\left(\frac{d\eta}{d\pi_1}\right)_{0j},$$

wobei $\left(\frac{\partial\eta}{\partial\pi_1}\right)_{0j}$ die Ableitung der Linkfunktion an der Stelle $\hat{\pi}_1^{0j}$ bezeichnet.
Im Falle des Logit-Links ergibt sich die (am Punkt $\hat{\pi}_1^{0j}$ auszuwertende) Ableitungsfunktion als

$$\frac{\partial\eta}{\partial\pi_1} = \frac{1}{\pi_1(1-\pi_1)}\ , \tag{3.7}$$

für das Probit-Modell dagegen als

$$\frac{\partial\eta}{\partial\pi_1} = \frac{1}{\phi\left[\Phi^{-1}(\pi_1)\right]}\ .$$

- **Schritt 2:**

  Berechne die momentanen Gewichte

$$w_j^0 := \left[\frac{\hat{\pi}_1^{0j}(1-\hat{\pi}_1^{0j})}{n_j}\cdot\left(\frac{d\eta}{d\pi_1}\right)^2_{0j}\right]^{-1} = \left[V_{0j}\cdot\left(\frac{d\eta}{d\pi_1}\right)^2_{0j}\right]^{-1}$$

  mit $V_{0j} := \frac{\hat{\pi}_1^{0j}(1-\hat{\pi}_1^{0j})}{n_j}$. Da

$$\mathrm{Var}(Y_j/n_j) = \frac{1}{n_j^2}\cdot\mathrm{Var}(Y_j) = \frac{1}{n_j^2}\cdot n_j\,\pi_1(1-\pi_1) = \frac{\pi_1(1-\pi_1)}{n_j}$$

  gilt, handelt es sich bei V_{0j} also um die aktuelle geschätzte Varianz von Y_j/n_j. Im Logit-Modell vereinfachen sich die zu berechnenden Gewichte unter Verwendung von (3.7) zu

$$w_j^0 = n_j\,\hat{\pi}_1^{0j}(1-\hat{\pi}_1^{0j}).$$

- **Schritt 3:**

  Regressiere $\boldsymbol{z^0}$ auf die Kovariablenvektoren $\boldsymbol{x_1},\ldots,\boldsymbol{x_k}$ unter Verwendung der Gewichtsmatrix $\boldsymbol{W^0} := \mathrm{diag}(w_1^0, w_2^0, \ldots, w_k^0)$. Nenne die resultierenden Parameterschätzer $\hat{\boldsymbol{\beta}}_1$. Es gilt $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X'W^0X})^{-1}\boldsymbol{X'W^0z^0}$. Dabei ist die Designmatrix $\boldsymbol{X}$ definiert als $\boldsymbol{X} := (\boldsymbol{x_1},\ldots,\boldsymbol{x_k})'$.

- **Schritt 4:**

  Berechne die neuen geschätzten linearen Prädiktoren $\hat{\eta}_j^1 := \hat{\boldsymbol{\beta}}_1'\boldsymbol{x_j}$ und die neuen geschätzten Erfolgswahrscheinlichkeiten $\hat{\pi}_1^{1j} := F(\hat{\boldsymbol{\beta}}_1'\boldsymbol{x_j})$.

- **Schritt 5:**

  Gehe zu Schritt 1. Wiederhole solange, bis die Veränderung in den Schätzern klein genug ist.

Um geeignete Startwerte zu erhalten, benützt man oftmals die Daten y_j und n_j, aus denen man über die Linkfunktion $\hat{\eta}_j^0 := F^{-1}(y_j/n_j),\ j = 1, \ldots, k$ berechnen kann.

Im Allgemeinen konvergiert dieser Prozess, allerdings können in bestimmten Fällen Probleme hinsichtlich Existenz und Eindeutigkeit von $\hat{\boldsymbol{\beta}}$ auftreten. Diesbezüglich gibt es zahlreiche Resultate (siehe z.B. Haberman (1977) oder Wedderburn (1976)). Dabei kommt es nicht nur auf die Gestalt der Log-Likelihood bzw. der Linkfunktion an, sondern auch auf die jeweilige Datenkonstellation. Für ungruppierte Daten nennt Silvapulle (1981) hinreichende und notwendige Kriterien, den Fall gruppierter Daten behandeln z.B. Fahrmeir & Tutz (1994). Auf diesen Fall wollen wir hier näher eingehen:

Seien die Daten also im Folgenden wieder gruppiert mit k Beobachtungen $Y_j = y_j$, wobei $Y_j \sim \mathrm{Bin}(n_j, \pi_1(\boldsymbol{x_j}))$, $j = 1, \ldots, k$. Ferner gelte für die dem Modell zugrunde liegende Verteilungsfunktion F, dass F und $(1-F)$ log-konkav sind, d.h. dass $\log F$ und $\log(1-F)$ konkav sind. Dies erfüllen z.B. der Probit-, aber auch der Logit-Link sowie die komplementäre log-log-Funktion. Zudem setze man voraus, dass die Designmatrix $\boldsymbol{X}$ vollen Rang hat. Unter diesen Bedingungen existiert ein endliches und eindeutiges $\hat{\boldsymbol{\beta}}$, das die Log-Likelihood maximiert, genau dann, wenn das Ungleichungssystem

$$\begin{aligned} y_j \cdot \boldsymbol{\beta}'\boldsymbol{x_j} &\geq 0, \\ (n_j - y_j) \cdot \boldsymbol{\beta}'\boldsymbol{x_j} &\leq 0\ ,\ j = 1, \ldots, k \end{aligned} \tag{3.8}$$

nur die triviale Lösung $\boldsymbol{\beta} = 0$ besitzt. Anschaulich bedeutet dies, dass kein endlicher ML-Schätzer existiert, wenn es eine Hyperebene $\boldsymbol{\beta}'\boldsymbol{x} = 0$ gibt, die die Daten der Erfolge von den Daten der Nichterfolge trennt. Den Beweis hierfür führen Albert & Anderson (1984) bzw. Santner & Duffy (1986).

Abbildung 3.1 zeigt einen solchen Datensatz mit zwei Kovariablen. Die Kovariablenkonstellationen $\boldsymbol{x_j}$ der Erfolge liegen stets oberhalb (im mit „+“ gekennzeichneten Bereich) oder auf der Geraden $\boldsymbol{\beta}'\boldsymbol{x} = 0$, während die Kovariablenwerte $\boldsymbol{x_j}$ aller Nichterfolge unterhalb (im mit „−“ gekennzeichneten Bereich) oder genau auf der Geraden liegen. In Abbildung 3.2 ist es dagegen unmöglich, die Erfolge und Nichterfolge durch eine Gerade räumlich voneinander zu trennen. Die Daten zu Erfolgen und Misserfolgen überlappen sich hier; sie sind sozusagen gut „durchmischt“. In diesem Fall existiert ein eindeutiger ML-Schätzer $\hat{\boldsymbol{\beta}}$.

Da der ML-Schätzer zumindest asymptotisch existieren muss (Fahrmeir & Tutz (1994)), kann das Problem der Nichtexistenz in großen Stichproben überwunden werden. Falls der betrachtete Datensatz genügend groß ist, erhält man also meist gute Ergebnisse.

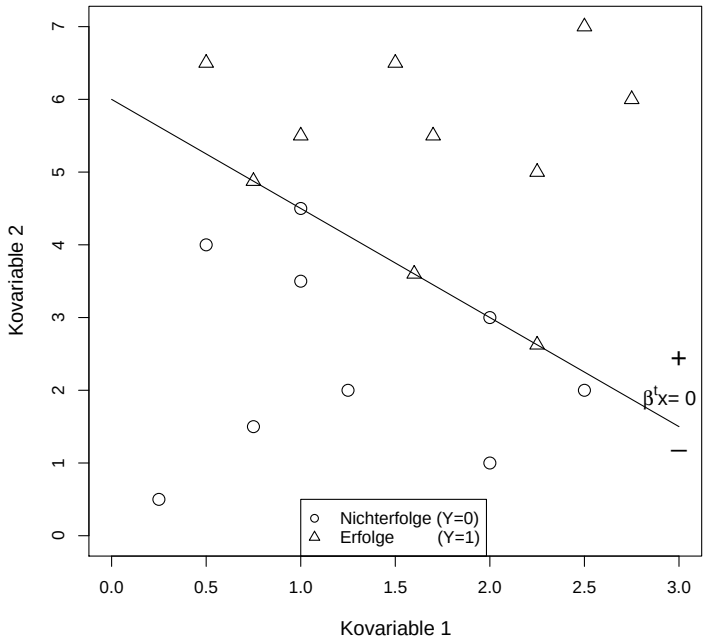


Abbildung 3.1: *Datenkonstellation, bei der Erfolge und Misserfolge durch die Gerade* $\boldsymbol{\beta}'\boldsymbol{x} = 0$ *getrennt werden .*

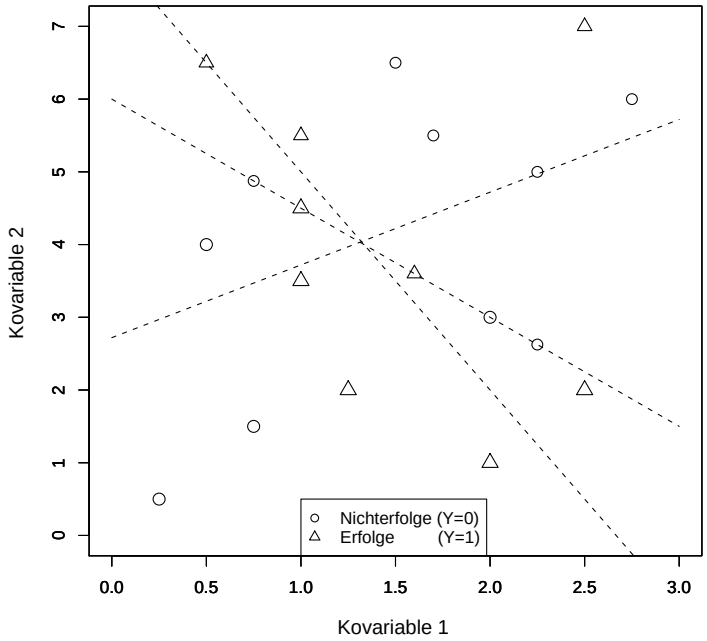


Abbildung 3.2: *Datenkonstellation, bei der Erfolge und Misserfolge nicht separabel sind.*

4 Bivariate binäre Regressionsmodelle

Für die Modellierung einer einzelnen binären Zielvariablen gibt es also, wie im vorherigen Kapitel gezeigt wurde, eine Vielzahl von Möglichkeiten, die in der Praxis äußerst brauchbare Ergebnisse liefern.

Häufig ist man allerdings nicht nur an einem, sondern an mehreren Zielmerkmalen gleichzeitig interessiert: Ein multivariater binärer Response ergibt sich beispielsweise bei Langzeitstudien, in denen das Auf- oder Nichtauftreten eines bestimmten Sachverhaltes zu mehreren Zeitpunkten untersucht wird. Solche Daten betrachten z.B. Glonek & McCullagh (1995): in vier aufeinanderfolgenden Jahren wurden 537 Kinder hinsichtlich des Auftretens von Kurzatmigkeit in Abhängigkeit vom Rauchverhalten der Mutter untersucht. Als Beispiel aus der Ökonomie sei die von Morimune (1979) verwendete „Panel Study of Income Dynamics“ des Survey Research Center, University of Michigan, erwähnt. Hierbei wurden in Abhängigkeit vom Familieneinkommen zwei Zielgrößen betrachtet, von denen die erste den Wert 1 annimmt, falls die Familie ein eigenes Haus besitzt, und 0 sonst, während die zweite Variable angibt, ob das Haus mehr als 5 Zimmer hat oder nicht. Auch in der Biometrie finden sich solche multivariaten Daten, z.B. bei der „Belgian Interuniversity Research on Nutrition and Health (BIRNH) study“ , in der die Rauch- und Trinkgewohnheiten von 10341 Testpersonen gleichzeitig untersucht wurden (Lesaffre & Molenberghs (1991)).

Wie später in diesem Kapitel erläutert wird, ist es im Allgemeinen keinesfalls ausreichend, für jede einzelne Komponente des Responses ein separates univariates Modell aufzustellen. Deshalb müssen für multivariate binäre Zielvariablen ganz eigene Methoden zur Modellierung bereitgestellt werden.

Hier setzen wir uns mit bivariaten binären Daten auseinander. Wir zeigen auf, wie sich der im vorherigen Kapitel vorgestellte Ansatz zur univariaten Modellierung auf den bivariaten Fall ausweiten lässt. Anschließend wird das auf einer bivariaten Normalverteilung basierende bivariate Probit-Modell näher erläutert. Am Ende des Kapitels werden kurz alternative Modelle dargestellt.

4.1 Das Schwellenwertkonzept im bivariaten Fall

Da der im univariaten Fall eingeführte Schwellenwertansatz höchst anschaulich und gut interpretierbar ist, bietet es sich an, diesen auf den bivariaten Fall zu übertragen.

Man betrachte also im Folgenden n unabhängige Beobachtungen eines bivariaten binären Responses $Y_i := (Y_{i1}, Y_{i2})', \ i = 1, \ldots, n$. Wir nehmen an, dass jeder Beobachtung eine ebenfalls bivariate, latente, in beiden Argumenten stetige Zufallsvariable $Z_i := (Z_{i1}, Z_{i2})$, $i = 1, \ldots, n$ zugrundeliegt. Y_{i1} nehme den Wert 1 an, wenn Z_{i1} eine bestimmte Schwelle θ_1 unterschreitet; $Y_{i2} = 1$ trete genau dann ein, wenn $Z_{i2} \leq \theta_2$ gilt. Z_{ij} soll sich wieder über die Kovariablen $x_{i1}, \ldots, x_{ip}$ und eine Störgröße ϵ_{ij} , $j = 1, 2$ ausdrücken lassen. Dieses bivariate Schwellenwertmodell kann man folgendermaßen formulieren:

$$Y_{i1} = 1|\boldsymbol{x_i} \Leftrightarrow Z_{i1} = \alpha_{10} + \alpha_{11}x_{i1} + \cdots + \alpha_{1p}x_{ip} + \epsilon_{i1} = \boldsymbol{\alpha}_1'\boldsymbol{x_i} + \epsilon_{i1} \leq \theta_1, \ i = 1, \ldots, n \quad (4.1)$$

$$Y_{i2} = 1|\boldsymbol{x_i} \Leftrightarrow Z_{i2} = \alpha_{20} + \alpha_{21}x_{i1} + \cdots + \alpha_{2p}x_{ip} + \epsilon_{i2} = \boldsymbol{\alpha}_2'\boldsymbol{x_i} + \epsilon_{i2} \leq \theta_2, \ i = 1, \ldots, n \quad (4.2)$$

$\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ bezeichnet hierbei wieder den Kovariablenvektor zur i-ten Beobachtung Y_i . Zudem nimmt man an, dass die Störgrößen $\epsilon_i := (\epsilon_{i1}, \epsilon_{i2})$ unabhängig sind für alle $i = 1, \ldots, n$. Innerhalb einer Beobachtung i müssen ϵ_{i1} und ϵ_{i2} allerdings keinesfalls unabhängig sein.

Im bivariaten Schwellenwertmodell ergeben sich vier verschiedene Responsekombinationen. Ihre jeweiligen Wahrscheinlichkeiten sind

$$\begin{aligned} \pi_{11}(\boldsymbol{x_i}) :&= P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{x_i}) = P(Z_{i1} \leq \theta_1, Z_{i2} \leq \theta_2) \\ &= P(\epsilon_{i1} \leq \theta_1 - \boldsymbol{\alpha}_1'\boldsymbol{x_i}, \ \epsilon_{i2} \leq \theta_2 - \boldsymbol{\alpha}_2'\boldsymbol{x_i}) = P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i}, \ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) \end{aligned} \quad (4.3)$$

$$\begin{aligned} \pi_{10}(\boldsymbol{x_i}) :&= P(Y_{i1} = 1, Y_{i2} = 0|\boldsymbol{x_i}) = P(Z_{i1} \leq \theta_1, Z_{i2} > \theta_2) \\ &= P(\epsilon_{i1} \leq \theta_1 - \boldsymbol{\alpha}_1'\boldsymbol{x_i}, \ \epsilon_{i2} > \theta_2 - \boldsymbol{\alpha}_2'\boldsymbol{x_i}) = P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i}, \ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}) \end{aligned} \quad (4.4)$$

$$\begin{aligned} \pi_{01}(\boldsymbol{x_i}) :&= P(Y_{i1} = 0, Y_{i2} = 1|\boldsymbol{x_i}) = P(Z_{i1} > \theta_1, Z_{i2} \leq \theta_2) \\ &= P(\epsilon_{i1} > \theta_1 - \boldsymbol{\alpha}_1'\boldsymbol{x_i}, \ \epsilon_{i2} \leq \theta_2 - \boldsymbol{\alpha}_2'\boldsymbol{x_i}) = P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i}, \ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) \end{aligned} \quad (4.5)$$

$$\begin{aligned} \pi_{00}(\boldsymbol{x_i}) :&= P(Y_{i1} = 0, Y_{i2} = 0|\boldsymbol{x_i}) = P(Z_{i1} > \theta_1, Z_{i2} > \theta_2) \\ &= P(\epsilon_{i1} > \theta_1 - \boldsymbol{\alpha}_1'\boldsymbol{x_i}, \ \epsilon_{i2} > \theta_2 - \boldsymbol{\alpha}_2'\boldsymbol{x_i}) = P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i}, \ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}), \end{aligned} \quad (4.6)$$

wobei $\boldsymbol{\beta}_j := (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})' \in \mathbb{R}^{p+1}, \ j = 1, 2$ die interessierenden Parametervektoren bezeichnet mit

$$\beta_{j0} := \theta_j - \alpha_{j0}, \ \beta_{j1} := -\alpha_{j1} \ , \ldots, \ \beta_{jp} := -\alpha_{jp}, \ j = 1, 2.$$

Die Interpretation der vier auftretenden Wahrscheinlichkeiten ist im bivariaten Schwellenwert-Modell also relativ einfach: Die beiden Werte $\boldsymbol{\beta}_1'\boldsymbol{x_i}$ und $\boldsymbol{\beta}_2'\boldsymbol{x_i}$ unterteilen die zweidimensionale Ebene in vier Quadranten. $\pi_{11}(\boldsymbol{x_i}), \pi_{10}(\boldsymbol{x_i})$, $\pi_{01}(\boldsymbol{x_i})$ und $\pi_{00}(\boldsymbol{x_i})$ entsprechen den Wahrscheinlichkeiten jeweils eines dieser Quadranten. Die Kovariablen bestimmen dabei die Lage des Teilpunktes $(\boldsymbol{\beta}_1'\boldsymbol{x_i}\,, \boldsymbol{\beta}_2'\boldsymbol{x_i})$ und damit, wieviel Wahrscheinlichkeitsmasse auf die einzelnen Quadranten entfällt. In Abbildung 4.1 ist dieser Sachverhalt graphisch dargestellt.

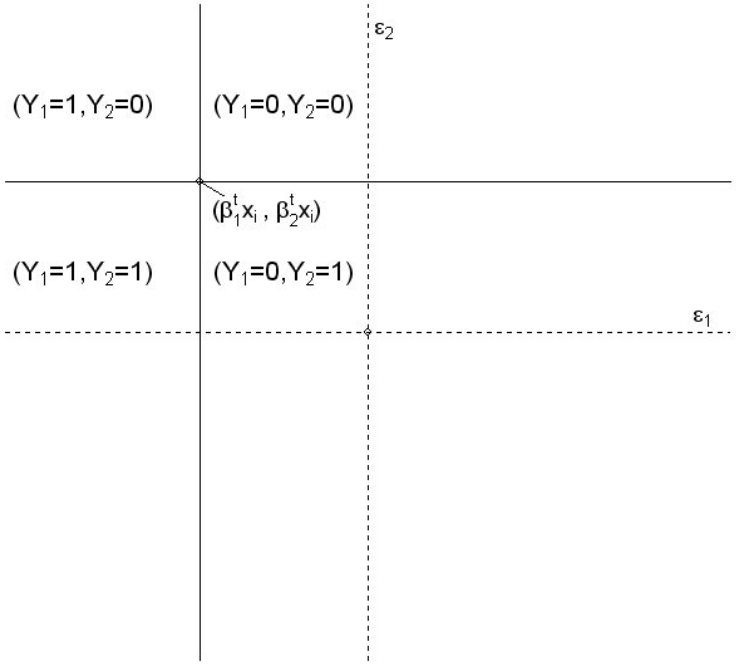


Abbildung 4.1: *Unterteilung der Ebene in vier Quadranten durch den Punkt* $(\boldsymbol{\beta}_1'\boldsymbol{x_i}\,,\,\boldsymbol{\beta}_2'\boldsymbol{x_i})$.

Sei F die gemeinsame Verteilungsfunktion von $(\epsilon_{i1}, \epsilon_{i2})$. Diese Funktion F wird wieder dazu benutzt, den Zusammenhang zwischen den auftretenden Wahrscheinlichkeiten und den Kovariablen zu beschreiben: Aus (4.3) ergibt sich analog zum univariaten Fall

$$\pi_{11}(\boldsymbol{x_i}) = F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})\ ,\ i = 1, \ldots, n. \tag{4.7}$$

Die übrigen Wahrscheinlichkeiten lassen sich mithilfe der marginalen Verteilungen F_j von ϵ_{ij}, $j = 1, 2$ bestimmen. Setze $\pi_{1+}(\boldsymbol{x_i}) := P(Y_{i1} = 1|\boldsymbol{x_i})$ bzw. $\pi_{+1}(\boldsymbol{x_i}) := P(Y_{i2} = 1|\boldsymbol{x_i})$ als marginale Erfolgswahrscheinlichkeiten. Diese ergeben sich zu

$$\pi_{1+}(\boldsymbol{x_i}) = P(Y_{i1} = 1|\boldsymbol{x_i}) = P(Z_{i1} \leq \theta_1) = P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i}) = F_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}), \tag{4.8}$$

$$\pi_{+1}(\boldsymbol{x_i}) = P(Y_{i2} = 1|\boldsymbol{x_i}) = P(Z_{i2} \leq \theta_2) = P(\epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) = F_2(\boldsymbol{\beta}_2'\boldsymbol{x_i}). \tag{4.9}$$

Offensichtlich gilt

$$P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i}) = P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) + P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}) \tag{4.10}$$

$$P(\epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) = P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) + P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) \tag{4.11}$$

und

$$\begin{aligned} &P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) + P(\epsilon_{i1} \leq \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}) \\ &+ P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} \leq \boldsymbol{\beta}_2'\boldsymbol{x_i}) + P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}) \ = \ 1. \end{aligned} \tag{4.12}$$

Unter Verwendung der Definitionen (4.3)–(4.6) und (4.8)–(4.9) lassen sich (4.10) und (4.11) umformulieren zu

$$\pi_{10}(\boldsymbol{x_i}) = \pi_{1+}(\boldsymbol{x_i}) - \pi_{11}(\boldsymbol{x_i}) = F_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}) \tag{4.13}$$

$$\pi_{01}(\boldsymbol{x_i}) = \pi_{+1}(\boldsymbol{x_i}) - \pi_{11}(\boldsymbol{x_i}) = F_2(\boldsymbol{\beta}_2'\boldsymbol{x_i}) - F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}) \tag{4.14}$$

und aus (4.12) ergibt sich

$$\begin{aligned}\pi_{00}(\boldsymbol{x_i}) =& 1 - \pi_{11}(\boldsymbol{x_i}) - \pi_{10}(\boldsymbol{x_i}) - \pi_{01}(\boldsymbol{x_i}) = 1 - \pi_{1+}(\boldsymbol{x_i}) - \pi_{+1}(\boldsymbol{x_i}) + \pi_{11}(\boldsymbol{x_i}) \\ =& 1 - F_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - F_2(\boldsymbol{\beta}_2'\boldsymbol{x_i}) + F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}).\end{aligned} \tag{4.15}$$

Wie aus (4.8) und (4.9) klar ersichtlich ist, besteht das bivariate Schwellenwert-Modell also marginal aus zwei univariaten Modellen mit der jeweiligen Linkfunktion F_1^{-1} bzw. F_2^{-1}. Es reicht im Allgemeinen jedoch keinesfalls, nur die beiden marginalen Modelle für die einzelnen Response-Komponenten anzupassen und

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}|\boldsymbol{x_i}) = P(Y_{i1} = y_{i1}|\boldsymbol{x_i}) \cdot P(Y_{i2} = y_{i2}|\boldsymbol{x_i}) \tag{4.16}$$

für $y_{i1}, y_{i2} = 0, 1$ zu setzen, da man dabei die Abhängigkeitsstruktur von Y_{i1} und Y_{i2} vollkommen vernachlässigt. Dieses Vorgehen ist deshalb nur im Falle zweier unabhängiger Zielvariablen innerhalb einer Beobachtung gerechtfertigt, denn nur in diesem Fall gilt

$$F(\epsilon_{i1}, \epsilon_{i2}) = F_1(\epsilon_{i1}) \cdot F_2(\epsilon_{i2})$$

für alle i und damit (4.16). Bei vielen Anwendungen lässt sich diese Annahme allerdings nicht bedenkenlos treffen, so z.B. bei Studien über Zwillingspaare. Auch bei Untersuchungen, die am selben Individuum vorgenommen werden, kann man im Allgemeinen nicht von unkorrelierten Responses ausgehen. Ein Beispiel hierfür ist das Auftreten bzw. Nichtauftreten einer Erkrankung des linken bzw. rechten Auges einer Testperson (vgl. Liang *et al.* (1992)).

Bei dem in dieser Arbeit untersuchten Datensatz muss man zunächst ebenfalls von einer Korreliertheit der beiden Ereignisse `storno` und `abschluss` ausgehen: Es ist nicht auszuschließen, dass sich ein bereits getätigtes Storno negativ auf die Entscheidung des Kunden auswirkt, einen neuen Vertrag abzuschließen. Ebenso wird man intuitiv vermuten, dass bei Kunden, die gerade einen weiteren Vertrag abgeschlossen und damit ihr Vertrauen in das Versicherungsunternehmen bekundet haben, eine Stornierung weniger wahrscheinlich ist. Daher muss durchaus ein (negativer) Zusammenhang zwischen beiden Variablen in Betracht gezogen werden.

Für die konkrete Modellierung stellt sich die Frage nach einer geeigneten bivariaten Verteilungsfunktion. Häufig verwendet man eine bivariate Standardnormalverteilung F. Das resultierende Modell bezeichnet man - in Anlehnung an den univariaten Fall - als bivariates Probit-Modell.

4.2 Das bivariate Probit-Modell

Beim bivariaten Probit-Modell spezifiziert man die gemeinsame Verteilung F der Störgrößen ϵ_{i1} und ϵ_{i2} als bivariate Standardnormalverteilung mit Korrelationsparameter ρ_i, also

$$\begin{pmatrix}\epsilon_{i1}\\ \epsilon_{i2}\end{pmatrix} \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_i := \begin{pmatrix}1 & \rho_i\\ \rho_i & 1\end{pmatrix}\right) \ , \quad i = 1, \ldots, n, \tag{4.17}$$

wobei Σ_i die Kovarianzmatrix zur i-ten Beobachtung bezeichnet.

Marginal gilt folglich

$$\epsilon_{ij} \sim N(0,1) \ , \quad j = 1, 2 \ . \tag{4.18}$$

Daraus folgt für $Z_{ij} = \boldsymbol{\alpha}_j' \boldsymbol{x_i} + \epsilon_{ij},\ j = 1, 2$ die gemeinsame Verteilung als

$$\begin{pmatrix}Z_{i1}\\ Z_{i2}\end{pmatrix} \sim N_2\left(\begin{pmatrix}\boldsymbol{\alpha}_1' \boldsymbol{x_i}\\ \boldsymbol{\alpha}_2' \boldsymbol{x_i}\end{pmatrix}, \begin{pmatrix}1 & \rho_i\\ \rho_i & 1\end{pmatrix}\right)$$

und die marginale Verteilung als

$$Z_{ij} \sim N(\boldsymbol{\alpha}_j' \boldsymbol{x_i}, 1) \ , \quad j = 1, 2 \ .$$

Während man innerhalb einer Beobachtung also von korrelierten Zielvariablen ausgeht, nimmt man bei verschiedenen Beobachtungen jedoch an, dass diese unabhängig sind, also dass

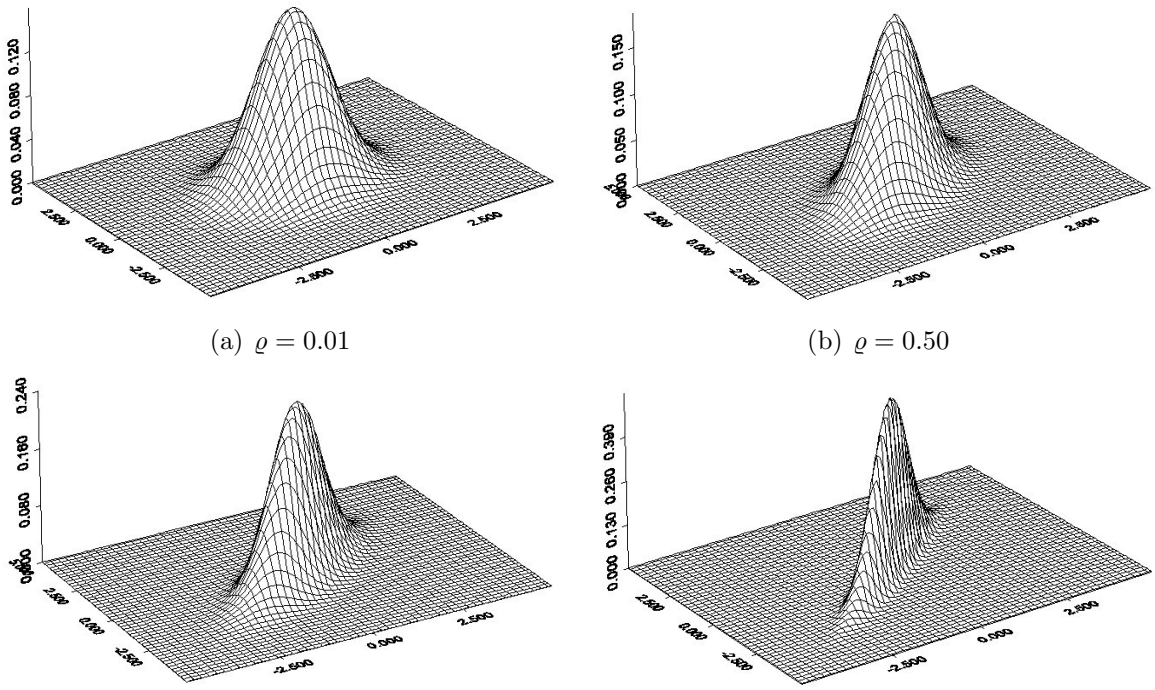
$$\operatorname{corr}(\epsilon_{ij}, \epsilon_{kl}) = \begin{cases} 1, & \text{falls } i = k,\ j = l \\ \rho_i, & \text{falls } i = k,\ j \neq l \\ 0, & \text{falls } i \neq k \end{cases}$$

für $i, k = 1, \ldots, n$ und $j, l = 1, 2$ gilt.

Die Wahrscheinlichkeit $\pi_{11}(\boldsymbol{x_i})$ berechnet sich mit (4.7) und (4.17) über die Dichte der bivariaten Normalverteilung als

$$\begin{aligned}
\pi_{11}(\boldsymbol{x_i}) &= \int_{-\infty}^{\boldsymbol{\beta}_1' \boldsymbol{x_i}} \int_{-\infty}^{\boldsymbol{\beta}_2' \boldsymbol{x_i}} \frac{1}{2\pi\ |\Sigma_i|^{\frac{1}{2}}} \ \exp\left\{-\frac{1}{2}\begin{pmatrix}\epsilon_{i1}\\ \epsilon_{i2}\end{pmatrix}' {\Sigma_i}^{-1} \begin{pmatrix}\epsilon_{i1}\\ \epsilon_{i2}\end{pmatrix}\right\} d\epsilon_{i2} d\epsilon_{i1} \\
&= \int_{-\infty}^{\boldsymbol{\beta}_1' \boldsymbol{x_i}} \int_{-\infty}^{\boldsymbol{\beta}_2' \boldsymbol{x_i}} \frac{1}{2\pi\sqrt{(1-{\rho_i}^2)}} \ \exp\left\{\frac{-({\epsilon_{i1}}^2 - 2\rho_i \epsilon_{i1}\epsilon_{i2} + {\epsilon_{i2}}^2)}{2(1-{\rho_i}^2)}\right\} d\epsilon_{i2} d\epsilon_{i1} \\
&= \Phi_2(\boldsymbol{\beta}_1' \boldsymbol{x_i}, \boldsymbol{\beta}_2' \boldsymbol{x_i}, \rho_i) \ ,
\end{aligned} \tag{4.19}$$

wobei $\Phi_2(w_1, w_2, \varrho)$ die Verteilungsfunktion der bivariaten Standardnormalverteilung an der Stelle (w_1, w_2) mit Korrelation ϱ bezeichnet.



(a) $\varrho = 0.01$ (b) $\varrho = 0.50$

(c) $\varrho = 0.75$ (d) $\varrho = 0.95$

Abbildung 4.2: *Dichten der bivariaten Standardnormalverteilung mit Korrelationsparametern* $\varrho = 0.01,\ 0.5,\ 0.75$ *und* 0.95

Abbildung 4.2 zeigt exemplarisch einige Dichtefunktionen der bivariaten Standardnormalverteilung mit Korrelationsparametern $\varrho = 0.01,\ 0.5,\ 0.75$ bzw. 0.95.

In Chib & Greenberg (1998) wird auf die Notwendigkeit einer Parameterisierung von Σ_i in Korrelationsform hingewiesen, da sonst Probleme mit der Identifizierbarkeit der Parameter auftreten, wie wir jetzt zeigen. Sei die Verteilung von $(\epsilon_{i1}, \epsilon_{i2})$ nicht standardisiert mit $\mathrm{Var}(\epsilon_{i1}) = {\sigma_{i1}}^2, \mathrm{Var}(\epsilon_{i2}) = {\sigma_{i2}}^2$ und $\mathrm{Cov}(\epsilon_{i1}, \epsilon_{i2}) = \rho_i \sigma_{i1} \sigma_{i2}$, also

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_i := \begin{pmatrix} {\sigma_{i1}}^2 & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & {\sigma_{i2}}^2 \end{pmatrix} \right).$$

Dann gilt mit der Substitution $\tilde{\epsilon}_{i1} := \frac{\epsilon_{i1}}{\sigma_{i1}},\ \tilde{\epsilon}_{i2} := \frac{\epsilon_{i2}}{\sigma_{i2}}$ und $\tilde{\boldsymbol{\beta}}_1 := \frac{\boldsymbol{\beta}_1}{\sigma_{i1}},\ \tilde{\boldsymbol{\beta}}_2 := \frac{\boldsymbol{\beta}_2}{\sigma_{i2}}$:

$$\pi_{11}(\boldsymbol{x_i}) = \int_{-\infty}^{\boldsymbol{\beta}_1' \boldsymbol{x_i}} \int_{-\infty}^{\boldsymbol{\beta}_2' \boldsymbol{x_i}} \frac{|\Omega_i|^{-\frac{1}{2}}}{2\pi} \exp\left\{ -\frac{1}{2} \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix}' \Omega_i^{-1} \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \right\} d\epsilon_{i2} d\epsilon_{i1}$$

$$= \int_{-\infty}^{\boldsymbol{\beta}_1' \boldsymbol{x_i}} \int_{-\infty}^{\boldsymbol{\beta}_2' \boldsymbol{x_i}} \frac{[\sigma_{i1}\sigma_{i2}]^{-1}}{2\pi\sqrt{(1-{\rho_i}^2)}} \exp\left\{ \frac{-1}{2(1-{\rho_i}^2)} \left[\frac{{\epsilon_{i1}}^2}{{\sigma_{i1}}^2} - \frac{2\rho_i \epsilon_{i1} \epsilon_{i2}}{\sigma_{i1}\sigma_{i2}} + \frac{{\epsilon_{i2}}^2}{{\sigma_{i2}}^2} \right] \right\} d\epsilon_{i2} d\epsilon_{i1}$$

$$= \int_{-\infty}^{\tilde{\boldsymbol{\beta}}_1' \boldsymbol{x_i}} \int_{-\infty}^{\tilde{\boldsymbol{\beta}}_2' \boldsymbol{x_i}} \frac{1}{2\pi\sqrt{(1-{\rho_i}^2)}} \exp\left\{ -\frac{\tilde{\epsilon}_{i1}^2 - 2\tilde{\epsilon}_{i2}\tilde{\epsilon}_{i1} + \tilde{\epsilon}_{i2}^2}{2(1-{\rho_i}^2)} \right\} d\tilde{\epsilon}_{i1} d\tilde{\epsilon}_{i2}\ .$$

Jedes nichtstandardisierte Modell mit Parametern $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ lässt sich also durch einfache Transformation auf ein dazu äquivalentes standardisiertes Modell mit neuen Parametern $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2$ zurückführen, so dass deren Eindeutigkeit nicht mehr gewährleistet ist. Eine Parameterisierung in Korrelationsform garantiert dagegen die eindeutige Identifizierbarkeit der Parameter.

Die übrigen Wahrscheinlichkeiten $\pi_{10}(\boldsymbol{x_i})$, $\pi_{01}(\boldsymbol{x_i})$ und $\pi_{00}(\boldsymbol{x_i})$ kann man nach (4.3)–(4.6) ebenfalls über die Dichte der bivariaten Standardnormalverteilung berechnen als

$$\pi_{10}(\boldsymbol{x_i}) = \int_{-\infty}^{\boldsymbol{\beta}_1'\boldsymbol{x_i}} \int_{\boldsymbol{\beta}_2'\boldsymbol{x_i}}^{\infty} \frac{1}{2\pi\sqrt{(1-{\rho_i}^2)}} \quad \exp\left\{\frac{-({\epsilon_{i1}}^2 - 2\rho_i\epsilon_{i1}\epsilon_{i2} + {\epsilon_{i2}}^2)}{2(1-{\rho_i}^2)}\right\} d\epsilon_{i2} d\epsilon_{i1}$$

$$\pi_{01}(\boldsymbol{x_i}) = \int_{\boldsymbol{\beta}_1'\boldsymbol{x_i}}^{\infty} \int_{-\infty}^{\boldsymbol{\beta}_2'\boldsymbol{x_i}} \frac{1}{2\pi\sqrt{(1-{\rho_i}^2)}} \quad \exp\left\{\frac{-({\epsilon_{i1}}^2 - 2\rho_i\epsilon_{i1}\epsilon_{i2} + {\epsilon_{i2}}^2)}{2(1-{\rho_i}^2)}\right\} d\epsilon_{i2} d\epsilon_{i1}$$

$$\pi_{00}(\boldsymbol{x_i}) = \int_{\boldsymbol{\beta}_1'\boldsymbol{x_i}}^{\infty} \int_{\boldsymbol{\beta}_2'\boldsymbol{x_i}}^{\infty} \frac{1}{2\pi\sqrt{(1-{\rho_i}^2)}} \quad \exp\left\{\frac{-({\epsilon_{i1}}^2 - 2\rho_i\epsilon_{i1}\epsilon_{i2} + {\epsilon_{i2}}^2)}{2(1-{\rho_i}^2)}\right\} d\epsilon_{i2} d\epsilon_{i1} \ .$$

Abbildung 4.3 zeigt die Niveaulinien einer bivariaten Standardnormalverteilung und die vier durch den Teilpunkt $(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})$ erzeugten Integrationsbereiche, die den auftretenden Wahrscheinlichkeiten zugrundeliegen.

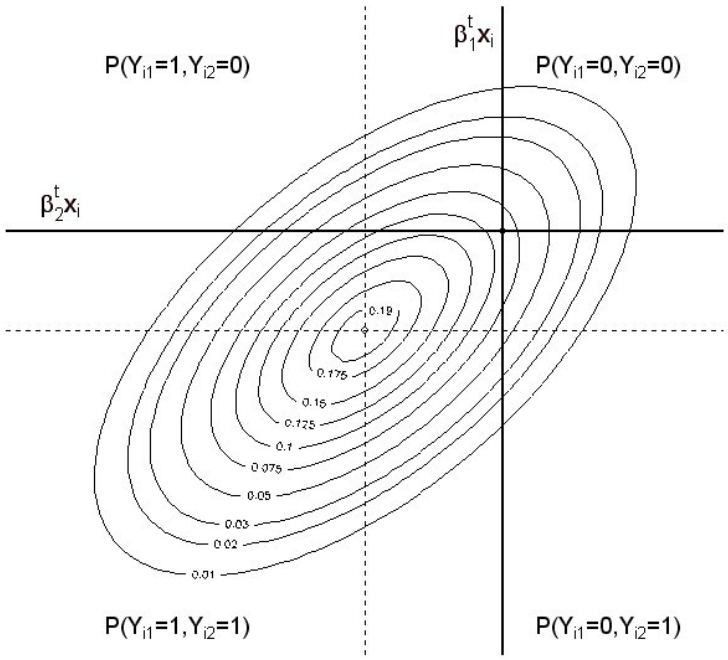


Abbildung 4.3: *Niveaulinienplot einer Standardnormalverteilung und die vier durch den Teilpunkt* $(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})$ *erzeugten Integrationsbereiche.*

Anstatt durch die – relativ aufwändige – Auswertung eines Doppelintegrals lassen sich die drei übrigen Wahrscheinlichkeiten aber auch auf einfachere Art und Weise bestimmen, sofern man $\pi_{11}(\boldsymbol{x_i}) = \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)$ schon kennt:

Da ϵ_{i1} und ϵ_{i2} nach (4.18) marginal standardnormalverteilt sind, ergeben sich aus (4.8) und (4.9) zunächst $\pi_{1+}(\boldsymbol{x_i})$ und $\pi_{+1}(\boldsymbol{x_i})$ zu

$$\pi_{1+}(\boldsymbol{x_i}) = P(Y_{i1} = 1|\boldsymbol{x_i}) = \Phi(\boldsymbol{\beta}_1'\boldsymbol{x_i}) \ , \tag{4.20}$$

$$\pi_{+1}(\boldsymbol{x_i}) = P(Y_{i2} = 1|\boldsymbol{x_i}) = \Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i}) \ , \tag{4.21}$$

wobei $\Phi(z)$ die Verteilungsfunktion der univariaten Standardnormalverteilung bezeichnet.

Daraus lassen sich dann mit (4.13)-(4.15) die Wahrscheinlichkeiten $\pi_{10}(\boldsymbol{x_i})$, $\pi_{01}(\boldsymbol{x_i})$ und $\pi_{00}(\boldsymbol{x_i})$ ganz einfach über

$$\pi_{10}(\boldsymbol{x_i}) = \Phi(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)$$

$$\pi_{01}(\boldsymbol{x_i}) = \Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i}) - \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)$$

$$\pi_{00}(\boldsymbol{x_i}) = 1 - \Phi(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - \Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i}) + \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)$$

berechnen.

Wie aus (4.20) und (4.21) klar ersichtlich ist, besteht das bivariate Probit-Modell also marginal aus zwei univariaten Probit-Modellen. Wie wir aber bereits im vorhergehenden Abschnitt erläuterten, ist es im Allgemeinen nicht ausreichend, nur diese beiden Modelle anzupassen, da dabei die Korrelation der Zielvariablen nicht ins Modell eingeht.

Der Korrelationsparameter ρ_i kann mithilfe eines weiteren Parametervektors $\boldsymbol{\alpha}_3$ ebenfalls in Abhängigkeit von den Kovariablen modelliert werden als

$$\rho_i = \rho(\boldsymbol{\alpha}_3'\boldsymbol{x_i}) \ , \quad i = 1, \ldots, n.$$

Dieses Modell wird von Lesaffre & Molenberghs (1991) im Kontext eines Probit-Modells für multivariate binäre Daten ausführlich behandelt.

Man kann aber auch annehmen, dass die Korrelation über alle Beobachtungen konstant ist, also dass $\rho_i = \rho, \ i = 1 \ldots n,$ so dass für die Störgrößen

$$(\epsilon_{i1}, \epsilon_{i2}) \overset{iid}{\sim} N_2(0, 0, \rho) \ , \quad i = 1, \ldots, n$$

gilt, wobei $N_2(\mu_1, \mu_2, \varrho)$ die bivariate Normalverteilung mit Erwartungswerten μ_1 und μ_2, auf 1 normierten Varianzen und Korrelationsparameter ϱ bezeichnet. Diese ursprüngliche Form des bivariaten Probit-Modells wurde erstmals von Ashford & Sowden (1970) vorgestellt.

Bisher wurde davon ausgegangen, dass die beiden Zielgrößen Y_{i1} und Y_{i2} von genau denselben Kovariablen beeinflusst werden, so dass nur ein einziger Kovariablenvektor $\boldsymbol{x_i}$ im Modell (4.1)–(4.2) auftritt. In der Praxis ist es aber sehr gut möglich, dass die Responses von unterschiedlichen Kovariablen abhängen, oder dass verschiedene Kategorisierungen einer Variablen verwendet werden. Für diesen speziellen Fall ist eine Modifikation des bisherigen Modells nötig, die im folgenden Abschnitt präsentiert wird.

4.3 Das bivariate „seemingly unrelated“ Probit-Modell

Abbildung 4.2 zeigt exemplarisch einige Dichtefunktionen der bivariaten Standardnormalverteilung mit Korrelationsparametern $\varrho = 0.01,\ 0.5,\ 0.75$ bzw. 0.95. Das bivariate Probit-Modell in (4.1)–(4.2) ist relativ unflexibel, da man dabei davon ausgeht, dass jede der p Kovariablen tatsächlich beide Zielvariablen beeinflusst. Für jede Kovariable werden somit automatisch jeweils zwei Parameter geschätzt. Tatsächlich aber kann die erste binäre Größen von ganz anderen (oder von unterschiedlich vielen) Kovariablen abhängen als die zweite.

Zudem kann man bei dieser Formulierung für jede Kovariable nur eine Form der Darstellung verwenden, während es in der Praxis für eine adäquate Modellierung durchaus vonnöten sein kann, unterschiedliche Transformationen für stetige Variablen oder zwei verschiedene Kategorisierungen zu verwenden.

Auch bei dem in dieser Arbeit untersuchten Datensatz stellt man fest, dass die beiden Zielvariablen von unterschiedlichen Kovariablen beeinflusst werden. Zudem kann man eine bessere Anpassung an die Daten erreichen, wenn man manche Kovariablen für die zwei Zielgrößen auf verschiedene Art modelliert.

Für diesen Spezialfall des bivariaten Probit-Modells existieren also zu jeder Beobachtung zwei unterschiedliche Kovariablenvektoren $\boldsymbol{x_{i1}}$ und $\boldsymbol{x_{i2}}$, und das Modell lautet:

$$Y_{i1} = 1|\boldsymbol{x_{i1}} \Longleftrightarrow Z_{i1} = \boldsymbol{\alpha}_1'\boldsymbol{x_{i1}} + \epsilon_{i1} \leq \theta_1\ , \quad i = 1, \ldots, n \tag{4.22}$$

$$Y_{i2} = 1|\boldsymbol{x_{i2}} \Longleftrightarrow Z_{i2} = \boldsymbol{\alpha}_2'\boldsymbol{x_{i2}} + \epsilon_{i2} \leq \theta_2\ , \quad i = 1, \ldots, n \tag{4.23}$$

Die Wahrscheinlichkeit $\pi_{11}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}) = P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}})$ ergibt sich entsprechend als

$$\pi_{11}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}) = \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_{i1}}, \boldsymbol{\beta}_2'\boldsymbol{x_{i2}}, \rho_i)\ ,$$

mit $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p_1})' \in \mathbb{R}^{p_1+1}, \quad \boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \ldots, \beta_{2p_2})' \in \mathbb{R}^{p_2+1}$.

Dieses Modell wird bei der von uns verwendeten Statistik-Software STATA® als bivariates „seemingly unrelated“ Probit-Modell bezeichnet.

Jedes solche Modell kann durch Reparameterisierung auf ein bivariates Probit-Modell zurückgeführt werden: Definiere zur i-ten Beobachtung den neuen gemeinsamen Kovaria-

blenvektor $\boldsymbol{x_i}$ als

$$\boldsymbol{x_i} := \boldsymbol{x_{i1}} \cup \boldsymbol{x_{i2}}\ , \quad i = 1, \ldots, n, \tag{4.24}$$

so dass jede Komponente von $\boldsymbol{x_{i1}}$ und $\boldsymbol{x_{i2}}$ nur einmal in $\boldsymbol{x_i}$ enthalten ist. Mit neu definierten Parametervektoren $\boldsymbol{\gamma}_j,\ j = 1, 2$ gilt dann

$$\boldsymbol{\beta}_j' \boldsymbol{x_{ij}} = \boldsymbol{\gamma}_j' \boldsymbol{x_i}\ , \quad j = 1, 2\,, \quad i = 1, \ldots, n, \tag{4.25}$$

wobei $\boldsymbol{\gamma}_j$ zum einen alle Elemente von $\boldsymbol{\beta}_j$ enthält, die nicht null sind. Für Kovariablen, die zwar im Kovariablenvektor $\boldsymbol{x_{i2}}$, nicht aber in $\boldsymbol{x_{i1}}$ vorkommen, setzt man ausserdem an die entsprechende Position in $\boldsymbol{\gamma}_1$ eine Null. Umgekehrt setzt man in $\boldsymbol{\gamma}_2$ entsprechend Nullen für diejenigen Kovariablen, die nicht in $\boldsymbol{x_{i2}}$ enthalten sind.

O.B.d.A gelte z.B. für $\boldsymbol{x_{i1}} = (1, x_{i11}, x_{i12}, \ldots, x_{i1p_1})'$ und $\boldsymbol{x_{i2}} = (1, x_{i21}, x_{i22}, \ldots, x_{i2p_2})'$, dass die ersten $s + 1$ $(s + 1 \leq \min(p_1, p_2))$ Einträge von $\boldsymbol{x_{i1}}$ zu Kovariablen gehören, die auch in $\boldsymbol{x_{i2}}$ vorkommen. (Der Interzept wird hierbei mitgezählt, da auch er in beiden Vektoren auftritt.) Damit gilt $x_{i1j} = x_{i2j}\ ,\ j = 1, \ldots, s$. Seien dagegen die Kovariablen, die den $p_1 - s$ übrigen Einträgen von $\boldsymbol{x_{i1}}$ zugrunde liegen, nicht in $\boldsymbol{x_{i2}}$ enthalten, während die Kovariablen zu den $p_2 - s$ übrigen Einträgen von $\boldsymbol{x_{i2}}$ nicht in $\boldsymbol{x_{i1}}$ vorkommen. Definiere die drei zugehörigen neuen Kovariablen- bzw. Parametervektoren $\boldsymbol{x_i}, \boldsymbol{\gamma}_1$ und $\boldsymbol{\gamma}_2 \in \mathbb{R}^{p_1+p_2-s+1}$ als

$$\boldsymbol{x_i} := (1,\ x_{i11},\ \ldots,\ x_{i1s},\ x_{i1(s+1)},\ \ldots,\ x_{i1p_1},\ x_{i2(s+1)},\ \ldots,\ x_{i2p_2})' = (\boldsymbol{x_{i1}},\ x_{i2(s+1)},\ \ldots,\ x_{i2p_2})'$$

$$\boldsymbol{\gamma}_1 := (\beta_{10},\ \beta_{11},\ \ldots,\ \beta_{1s},\ \beta_{1(s+1)},\ \ldots,\ \beta_{1p_1},\ \underbrace{0,\ \ldots,\ 0}_{p_2-s\ mal})' = (\boldsymbol{\beta_1},\ \mathbf{0})'$$

$$\boldsymbol{\gamma}_2 := (\beta_{20},\ \beta_{21},\ \ldots,\ \beta_{2s},\ \underbrace{0,\ \ldots,\ 0}_{p_1-s\ mal},\ \beta_{2(s+1)},\ \ldots,\ \beta_{2p_2})'.$$

Damit ist (4.25) erfüllt für alle $i\,,\ i = 1, \ldots, n\,,$ und es gilt

$$\pi_{11}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}) = \Phi_2(\boldsymbol{\beta}_1' \boldsymbol{x_{i1}}, \boldsymbol{\beta}_2' \boldsymbol{x_{i2}}, \rho_i) = \Phi_2(\boldsymbol{\gamma}_1' \boldsymbol{x_i}, \boldsymbol{\gamma}_2' \boldsymbol{x_i}, \rho_i) = \pi_{11}(\boldsymbol{x_i}).$$

In der neuen Parameterisierung handelt es sich daher nicht mehr um ein „seemingly unrelated“, sondern ein normales bivariates Probit-Modell der Form (4.1)–(4.2) .

4.4 Eigenschaften des bivariaten Probit-Modells

Wir betrachten im Folgenden ein Probit-Modell mit n unabhängigen Beobachtungen eines bivariaten binären Responses mit einem gemeinsamen Kovariablenvektor $\boldsymbol{x_i}$.

Wegen der nahen Verwandtschaft zum univariaten Probit-Modell lassen sich hinsichtlich der Eigenschaften einige Parallelen ziehen:

So gilt aufgrund der Symmetrieeigenschaften der bivariaten Normalverteilung (siehe Anhang A)

$$\begin{aligned}\pi_{00}(\boldsymbol{x_i}) &= P(\epsilon_{i1} > \boldsymbol{\beta}_1'\boldsymbol{x_i},\ \epsilon_{i2} > \boldsymbol{\beta}_2'\boldsymbol{x_i}) = P(-\epsilon_{i1} \leq -\boldsymbol{\beta}_1'\boldsymbol{x_i},\ -\epsilon_{i2} \leq -\boldsymbol{\beta}_2'\boldsymbol{x_i}) \\ &\overset{(A.14)}{=} \Phi_2(-\boldsymbol{\beta}_1'\boldsymbol{x_i}, -\boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)\ .\end{aligned}$$

Im bivariaten Fall wäre es also ebenfalls möglich, „Erfolg“ als das *Über*schreiten einer Schranke zu definieren, da sich dabei wieder nur die Parameterisierung, nicht aber das zugrunde liegende Modell ändert. O.B.d.A. könnte man deshalb auch hier $\theta_1 = 0 = \theta_2$ annehmen, da, wie schon in Abschnitt 2.1, Seite 5, festgestellt wurde, die Schranke θ_j und der Parameter $\alpha_{j0},\ j = 1, 2$ nicht beide gleichzeitig identifiziert werden können. Damit vereinfacht sich die Notation des Modells wieder etwas (verwendet z.B. in Czado (2000)). Insbesondere gilt dann $\boldsymbol{\alpha}_j = -\boldsymbol{\beta}_j,\ j = 1, 2$.

Auch beim bivariaten Probit-Modell ist es eine wichtige Frage, wie sich Veränderungen in $\boldsymbol{x_i}$ auf $\pi_{11}(\boldsymbol{x_i})$ auswirken. Dieser Einfluss auf die Erfolgswahrscheinlichkeit wird im Folgenden bestimmt. Für stetige Kovariablen ist dies allerdings weitaus aufwändiger als im univariaten Fall. Wie wir sehen werden, ist die Interpretation kategorialer Kovariablen auch im bivariaten Fall nur schwer möglich.

Für die bivariate Verteilungsfunktion $\Phi_2(z_1, z_2, \varrho)$ gilt (siehe (A.7), Anhang A)

$$\Phi_2(z_1, z_2, \varrho) = \int_{-\infty}^{z_1} \phi(w_1)\ \Phi\left(\frac{z_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right)\ dw_1. \tag{4.26}$$

Daraus ergibt sich sofort

$$\frac{\partial}{\partial z_1}\Phi_2(z_1, z_2, \varrho) = \phi(z_1) \cdot \Phi\left(\frac{z_2 - \varrho z_1}{\sqrt{1-\varrho^2}}\right) =: g_1(z_1, z_2, \varrho). \tag{4.27}$$

Analog bestimmt man über (A.8)

$$\frac{\partial}{\partial z_2}\Phi_2(z_1, z_2, \varrho) = \phi(z_2) \cdot \Phi\left(\frac{z_1 - \varrho z_2}{\sqrt{1-\varrho^2}}\right) =: g_2(z_1, z_2, \varrho). \tag{4.28}$$

Damit ergibt sich für das Probit-Modell nach der Kettenregel im Mehrdimensionalen für den Einfluss einer stetigen Kovariablen x_{ij}

$$\frac{\partial}{\partial x_{ij}}\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i) = g_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i) \cdot \beta_{1j} + g_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i) \cdot \beta_{2j}\ . \tag{4.29}$$

Im speziellen Fall unabhängiger Zielvariablen mit $\rho_i = 0,\ i = 1, \ldots, n$, vereinfachen sich g_1 und g_2 zu

$$g_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i) = \phi(\boldsymbol{\beta}_1'\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i}) \tag{4.30}$$

$$g_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i) = \phi(\boldsymbol{\beta}_2'\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{\beta}_1'\boldsymbol{x_i})\ . \tag{4.31}$$

Für kategoriale Kovariablen gestaltet sich die Interpretation, wie schon im univariaten Fall, schwierig. Betrachtet man wieder nur eine solche Kovariable in Dummy-Kodierung, so ergibt sich beim Wechsel von Referenzkategorie M in eine andere Kategorie I die Änderung der Erfolgswahrscheinlichkeiten als

$$\Phi_2(\beta_{10} + \beta_{1I}, \beta_{20} + \beta_{2I}, \rho_i) - \Phi_2(\beta_{10}, \beta_{20}, \rho_i).$$

Dieser Effekt ist aufgrund der Integraldarstellung von Φ_2 nur schwer zu quantifizieren.

Weiterhin lassen sich die Erwartungswerte $\mathbb{E}(Y_{ij}|\boldsymbol{x_i}),\ j = 1, 2$, bestimmen:

$$\mathbb{E}(Y_{ij}|\boldsymbol{x_i}) = 1 \cdot P(Y_{ij}|\boldsymbol{x_i} = 1) + 0 \cdot P(Y_{ij} = 0|\boldsymbol{x_i}) = P(Y_{ij} = 1|\boldsymbol{x_i}) = \Phi(\boldsymbol{\beta}_j'\boldsymbol{x_i}).$$

Der Erwartungswert $\mathbb{E}(Y_{ij}|\boldsymbol{x_i})$ entspricht also genau der marginalen Erfolgswahrscheinlichkeit.

Für den bedingten Erwartungswert $\mathbb{E}(Y_{i1}\,|Y_{i2} = 1, \boldsymbol{x_i})$ ergibt sich

$$\mathbb{E}(Y_{i1}\,|Y_{i2} = 1, \boldsymbol{x_i}) = P(Y_{i1} = 1|Y_{i2} = 1, \boldsymbol{x_i}) = \frac{P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{x_i})}{P(Y_{i2} = 1|\boldsymbol{x_i})} = \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)}{\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})}, \tag{4.32}$$

also die Erfolgswahrscheinlichkeit $\pi_{11}(\boldsymbol{x_i})$, die mit der marginalen Erfolgswahrscheinlichkeit $\pi_{+1}(\boldsymbol{x_i})$ gewichtet wird. Analog lässt sich $\mathbb{E}(Y_{i2}\,|Y_{i1} = 1, \boldsymbol{x_i})$ berechnen.

Da $P(Y_{i1} = 1, Y_{i2} = 0\,|\boldsymbol{x_i}) \overset{(A.12)}{=} \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, -\boldsymbol{\beta}_2'\boldsymbol{x_i}, -\rho_i)$ gilt (siehe Anhang A), ergibt sich

$$\begin{aligned}\mathbb{E}(Y_{i1}\,|Y_{i2} = 0\,, \boldsymbol{x_i}) &= P(Y_{i1} = 1|Y_{i2} = 0\,, \boldsymbol{x_i}) = \frac{P(Y_{i1} = 1, Y_{i2} = 0\,|\boldsymbol{x_i})}{P(Y_{i2} = 0\,|\boldsymbol{x_i})} \\ &= \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, -\boldsymbol{\beta}_2'\boldsymbol{x_i}, -\rho_i)}{1 - \Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})} = \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, -\boldsymbol{\beta}_2'\boldsymbol{x_i}, -\rho_i)}{\Phi(-\boldsymbol{\beta}_2'\boldsymbol{x_i})}\,.\end{aligned} \tag{4.33}$$

Die beiden bedingten Erwartungswerte (4.32) und (4.33) unterscheiden sich nur hinsichtlich der Vorzeichen in den Argumenten. Mithilfe des Faktors

$$2Y_{i2} - 1 = \begin{cases} 1, & \text{falls } Y_{i2} = 1 \\ -1, & \text{falls } Y_{i2} = 0 \end{cases} \tag{4.34}$$

der die notwendige Anpassung der Vorzeichen gewährleistet, kann man (4.32) und (4.33) zur Erwartungsfunktion $\mathbb{E}(Y_{i1}|Y_{i2}\,, \boldsymbol{x_i})$ zusammenführen:

$$\mathbb{E}(Y_{i1}|Y_{i2}\,, \boldsymbol{x_i}) = P(Y_{i1} = 1|Y_{i2}, \boldsymbol{x_i}) = \frac{\Phi_2\left[\boldsymbol{\beta}_1'\boldsymbol{x_i}, (2Y_{i2} - 1)\boldsymbol{\beta}_2'\boldsymbol{x_i}, (2Y_{i2} - 1)\rho_i\right]}{\Phi\left[(2Y_{i2} - 1)\boldsymbol{\beta}_2'\boldsymbol{x_i}\right]} \overset{(4.34)}{=}$$

$$= \begin{cases} \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho_i)}{\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})}\,, & \text{falls } Y_{i2} = 1 \\ \\ \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i}, -\boldsymbol{\beta}_2'\boldsymbol{x_i}, -\rho_i)}{\Phi(-\boldsymbol{\beta}_2'\boldsymbol{x_i})}\,, & \text{falls } Y_{i2} = 0\,. \end{cases}$$

Bei marginaler Betrachtung ergibt sich aus (4.32) für stetige Kovariablen

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{x_i}}\mathbb{E}(Y_{i1}|Y_{i2}=1,\boldsymbol{x_i}) = \\
&= \frac{1}{\Phi^2(\boldsymbol{\beta}_2'\boldsymbol{x_i})}\left[\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})\cdot\frac{\partial}{\partial \boldsymbol{x_i}}\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i) - \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\cdot\frac{\partial}{\partial \boldsymbol{x_i}}\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})\right] \\
&\overset{(4.29)}{=} \frac{[g_1(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\boldsymbol{\beta}_1 + g_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\boldsymbol{\beta}_2]}{\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})} - \frac{\Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\cdot\phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})\boldsymbol{\beta}_2}{\Phi^2(\boldsymbol{\beta}_2'\boldsymbol{x_i})} \\
&= \frac{g_1(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\boldsymbol{\beta}_1 + \left[g_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i) - \Phi_2(\boldsymbol{\beta}_1'\boldsymbol{x_i},\boldsymbol{\beta}_2'\boldsymbol{x_i},\rho_i)\cdot\frac{\phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})}{\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})}\right]\boldsymbol{\beta}_2}{\Phi(\boldsymbol{\beta}_2'\boldsymbol{x_i})}\,.
\end{aligned} \tag{4.35}$$

Die Ableitung $\frac{\partial}{\partial \boldsymbol{x_i}}\mathbb{E}(Y_{i1}|Y_{i2}=0\,,\boldsymbol{x_i})$ bildet man analog zu (4.35) unter Berücksichtigung der negativen Vorzeichen in den Argumenten.

4.5 Alternative Modelle

4.5.1 Das bivariate Probit-Modell mit logistischen Randverteilungen

Im Spezialfall $\theta_1 = 0 = \theta_2$ lassen sich auch andere marginale Verteilungen erzeugen, wenn man anstatt der linearen Formulierung

$$Z_{ij} = \boldsymbol{\alpha}_j'\boldsymbol{x_i} + \epsilon_{ij},\ j = 1,2$$

in (4.1), (4.2) einen nichtlinearen Ansatz verfolgt. So schlagen Le Cessie & van Houwelingen (1994) beispielsweise

$$Z_{ij} = -\Phi^{-1}\left(\frac{\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}{1+\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}\right) + \epsilon_{ij},\ j = 1,2 \tag{4.36}$$

vor. Damit erzeugt man logistische Randverteilungen, da

$$\begin{aligned}
P(Y_{ij}=1|\boldsymbol{x_i}) = P(Z_{ij}\le 0) = P\left(\epsilon_{ij} \le \Phi^{-1}\left(\frac{\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}{1+\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}\right)\right) &\overset{(4.18)}{=} \\
= \Phi\left(\Phi^{-1}\left(\frac{\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}{1+\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}\right)\right) = \frac{\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}{1+\exp(\boldsymbol{\alpha}_j'\boldsymbol{x_i})}\,, \quad j = 1,2\,.
\end{aligned} \tag{4.37}$$

Weitere Möglichkeiten zur Modellbildung ergeben sich, wenn man anstatt einer bivariaten Normalverteilung eine andere Verteilungsfunktion F zugrunde legt. Ein solches Modell wird im Folgenden kurz vorgestellt:

4.5.2 Das bivariate logistische Modell

Dieses relativ bekannte Modell resultiert aus einer zugrunde liegenden Plackett-Verteilung: Mit Hilfe einer von Plackett (1965) vorgeschlagenen Methode kann man zu zwei gegebenen univariaten Randverteilungen F und G der Zufallsvariablen X bzw. Y eine einparametrische Klasse von bivariaten Verteilungsfunktionen für (X, Y) konstruieren. Fréchet (1951) beweist, dass solch eine gemeinsame Verteilungsfunktion H

$$\max\{F(x) + G(y) - 1\,;\, 0\} \leq H(x, y) \leq \min\{F(x)\,;\, G(y)\} \tag{4.38}$$

erfüllen muss. Plackett fordert, dass die Funktion $H(x, y, \psi)$ der Gleichung

$$\psi = \frac{H(x, y, \psi)[1 - F(x) - G(y) + H(x, y, \psi)]}{[F(x) - H(x, y, \psi)][G(y) - H(x, y, \psi)]} \tag{4.39}$$

genügt. Betrachtet man die einzelnen Terme des Bruches (4.39), so ergibt sich

$$H(x, y, \psi) = P(X \leq x, Y \leq y)\ , \tag{4.40}$$

$$F(x) - H(x, y, \psi) = P(X \leq x) - P(X \leq x, Y \leq y) = P(X \leq x, Y > y)\ , \tag{4.41}$$

$$G(y) - H(x, y, \psi) = P(Y \leq y) - P(X \leq x, Y \leq y) = P(X > x, Y \leq y)\ , \tag{4.42}$$

$$\begin{aligned} &1 - F(x) - G(y) + H(x, y, \psi) = 1 - P(X \leq x) - P(Y \leq y) + P(X \leq x, Y \leq y) = \\ &= \underbrace{[1 - P(X \leq x)]}_{P(X>x)} - \underbrace{[P(Y \leq y) - P(X \leq x, Y \leq y)]}_{P(X>x, Y\leq y)} = P(X > x) - P(X > x, Y \leq y) \\ &= P(X > x, Y > y)\ . \end{aligned} \tag{4.43}$$

Der Parameter ψ stellt hierbei ein Maß für die Assoziation von X und Y dar, da mit (4.40)–(4.43)

$$\psi = \frac{P(X \leq x, Y \leq y)\, P(X > x, Y > y)}{P(X \leq x, Y > y)\, P(X > x, Y \leq y)} = \frac{P(X \leq x, Y \leq y)}{P(X \leq x, Y > y)} \Big/ \frac{P(X > x, Y \leq y)}{P(X > x, Y > y)}$$

gilt. Somit kann ψ interpretiert werden als Quotient der beiden Chancen

$$\frac{P(X \leq x, Y \leq y)}{P(X \leq x, Y > y)} \quad \text{und} \quad \frac{P(X > x, Y \leq y)}{P(X > x, Y > y)},$$

die das Verhältnis der Wahrscheinlichkeiten von $(Y \leq y)$ zu $(Y > y)$ jeweils für den Fall $X \leq x$ bzw. $X > x$ angeben. $\psi = 1$ signalisiert dabei die Unabhängigkeit der beiden Zufallsvariablen X und Y, wie wir noch zeigen werden.

Plackett beweist, dass (4.39) für festes $\psi \in (0, \infty)$ genau eine Lösung hat, die (4.38) erfüllt, und dass diese Lösung H eine zulässige gemeinsame Verteilungsfunktion für die Zufallsvariablen X und Y ist.

Da ψ für alle x, y konstant sein soll, entfernen wir zur Bestimmung der Lösung H zunächst die Argumente aus der Notation. Dann ergibt sich aus (4.39)

$$\begin{aligned} & \psi = \frac{H[1 - F - G + H]}{(F - H)(G - H)} \\ \iff \quad & \psi\,[(F - H)(G - H)] = H[1 - F - G + H] \\ \iff \quad & \psi\,[FG - FH - GH + H^2] = H - FH - GH + H^2 \\ \iff \quad & H^2(\psi - 1) - H[(F + G)(\psi - 1) + 1] + \psi FG = 0\,. \end{aligned} \tag{4.44}$$

Die gesuchte Funktion H ist also die positive Lösung der quadratischen Gleichung (4.44), die für $\psi \neq 1$ durch

$$\begin{aligned} & H(x, y, \psi) = \\ & = \frac{[F(x) + G(y)](\psi - 1) + 1 - \sqrt{\{[F(x) + G(y)](\psi - 1) + 1\}^2 - 4\psi(\psi - 1)F(x)G(y)}}{2(\psi - 1)} \\ & = \frac{1}{2}(\psi - 1)^{-1}\{[F(x) + G(y)](\psi - 1) + 1 - S(x, y, \psi)\} \end{aligned} \tag{4.45}$$

bestimmt ist, wobei $S(x, y, \psi) := \sqrt{\{[F(x) + G(y)](\psi - 1) + 1\}^2 + 4\psi(1 - \psi)F(x)G(y)}$.

Für $\psi = 1$, also im Fall der Unabhängigkeit von X und Y, erhält man die einfache Lösung

$$H(x, y, 1) = F(x)G(y)\,. \tag{4.46}$$

Dieses Vorgehen lässt sich problemlos auf das in Abschnitt 4.1 eingeführte bivariate Schwellenwertmodell (4.2)-(4.6) übertragen. Mit den marginalen Verteilungsfunktionen F_1, F_2 der Störgrößen ϵ_{i1} und ϵ_{i2} ergibt sich

$$\begin{aligned} \psi_i &= \frac{F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})\,[1 - F_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - F_2(\boldsymbol{\beta}_2'\boldsymbol{x_i}) + F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \psi_i)]}{[F_1(\boldsymbol{\beta}_1'\boldsymbol{x_i}) - F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \psi_i)]\,[F_2(\boldsymbol{\beta}_2'\boldsymbol{x_i}) - F(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \psi_i)]} \overset{(4.7),(4.13)-(4.15)}{=} \\ &= \frac{P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{x_i})\,P(Y_{i1} = 0, Y_{i2} = 0|\boldsymbol{x_i})}{P(Y_{i1} = 1, Y_{i2} = 0|\boldsymbol{x_i})\,P(Y_{i1} = 0, Y_{i2} = 1|\boldsymbol{x_i})} = \frac{\pi_{11}(\boldsymbol{x_i})}{\pi_{01}(\boldsymbol{x_i})} \Big/ \frac{\pi_{10}(\boldsymbol{x_i})}{\pi_{00}(\boldsymbol{x_i})} = \psi_i(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}). \end{aligned} \tag{4.47}$$

ψ_i gibt in diesem Falle also den sogenannten „Odds Ratio“ an, das Verhältnis der Chancen von Erfolg ($Y_{i1} = 1$) zu Misserfolg ($Y_{i1} = 0$) innerhalb der beiden Subpopulationen ($Y_{i2} = 1$) und ($Y_{i2} = 0$) unter der Bedingung $\boldsymbol{x_i}$. Falls $\psi_i = 1$, so sind die beiden binären Größen Y_{i1} und Y_{i2} nach (4.46) unabhängig. Für $\psi_i = 1$ sind die Chancen auf Erfolg bei

Y_{i1} innerhalb der beiden Subpopulationen gleich groß, die Ausprägung von Y_{i2} hat also offensichtlich keine Auswirkung auf diese Chancen.

Analog zum Korrelationsparameter ρ_i im bivariaten Probit-Modell kann der Odds Ratio ebenfalls in Abhängigkeit von den Kovariablen z.B. als

$$\log \psi_i = \boldsymbol{\beta}_3' \boldsymbol{x_i}$$

modelliert oder als konstant für alle n Beobachtungen angenommen werden, also

$$\psi_i = \psi\,,\ i = 1, \ldots, n.$$

Für festes ψ_i und bekannte Randverteilungen F_1, F_2 definiert Gleichung (4.47) die gemeinsame Verteilung F der beiden Störgrößen $(\epsilon_{i1}, \epsilon_{i2})$. Beim bivariaten logistischen Modell nimmt man an, dass die marginalen Wahrscheinlichkeiten π_{1+} und π_{+1} einer logistischen Verteilung folgen, d.h. dass gilt:

$$\pi_{1+}(\boldsymbol{x_i}) = P(Y_{i1} = 1|\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{\beta}_1' \boldsymbol{x_i})}{1 + \exp(\boldsymbol{\beta}_1' \boldsymbol{x_i})} \tag{4.48}$$

$$\pi_{+1}(\boldsymbol{x_i}) = P(Y_{i2} = 1|\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{\beta}_2' \boldsymbol{x_i})}{1 + \exp(\boldsymbol{\beta}_2' \boldsymbol{x_i})}\,. \tag{4.49}$$

Die Wahrscheinlichkeit $\pi_{11}(\boldsymbol{x_i})$ ergibt sich damit nach (4.45) zu

$$\pi_{11}(\boldsymbol{x_i}) = \begin{cases} \{1 + [\pi_{1+}(\boldsymbol{x_i}) + \pi_{+1}(\boldsymbol{x_i})](\psi_i - 1) - S(\pi_{1+}(\boldsymbol{x_i}), \pi_{+1}(\boldsymbol{x_i}), \psi_i)\}\cdot \\ \cdot \frac{1}{2}(\psi_i - 1)^{-1}, & \text{falls } \psi_i \neq 1 \\ \\ \pi_{1+}(\boldsymbol{x_i})\pi_{+1}(\boldsymbol{x_i}), & \text{falls } \psi_i = 1, \end{cases}$$

wobei

$$S(\pi_{1+}(\boldsymbol{x_i}), \pi_{+1}(\boldsymbol{x_i}), \psi) = \sqrt{[1 + [\pi_{1+}(\boldsymbol{x_i}) + \pi_{+1}(\boldsymbol{x_i})](\psi - 1)]^2 + 4\psi(1 - \psi)\pi_{1+}(\boldsymbol{x_i})\pi_{+1}(\boldsymbol{x_i})}.$$

Wie Molenberghs & Lesaffre (1994) konstatieren, handelt es sich bei diesem Modell um einen Spezialfall des von Dale (1986) vorgestellten „globalen Cross Ratio Modells“ , eines Regressionsmodells für bivariate ordinale Daten. Die Klasse der Cross Ratio Modelle von Dale beschränkt sich nicht auf logistische Randverteilungen. Der Vorteil dieser Modelle liegt in ihrer großen Flexibilität, da die marginalen Verteilungsfunktionen F_1 und F_2 völlig frei gewählt werden können. Lesaffre *et al.* (1994) betrachten beispielsweise ein bivariates Cross Ratio Modell mit normalverteilten Rändern. Da der Einfluss von Kovariablen aber gerade bei logistischen Rändern am einfachsten interpretiert werden kann (vgl. Abschnitt 2.3), zieht man am häufigsten diese Verteilung zur Modellierung heran.

Le Cessie & van Houwelingen (1994) vergleichen das Probit-Modell (4.36)–(4.37) mit dem bivariaten logistischen Modell (4.47)–(4.49) unter der Annahme, dass die Korrelation und der Odds Ratio für alle Beobachtungen konstant sind. Sie zeigen mittels Taylor-Approximationen erster Ordnung und einer graphischen Abschätzung, dass dann für den Korrelationsschätzer $\hat{\rho}$ und den logarithmierten Schätzer $\log \hat{\psi}$

$$\log \hat{\psi} \approx (1.7)^2 \hat{\rho}$$

gilt. Der Wert 1.7 entspricht dabei interessanterweise in etwa dem Faktor, um den sich auch die Schätzer des univariaten Probit- und Logit-Modells unterscheiden.

Eine praktische Anwendung erfährt das bivariate logistische Modell in der von T. Yee für R bzw. S-Plus entwickelten Library „VGAM“ . Darin findet sich die Implementierung einer auf Maximum-Likelihood-Schätzung basierenden Modellanpassung. Die Plattform S-Plus ist aber aufgrund langer Rechenzeiten generell eher ungeeignet für sehr umfangreiches Datenmaterial und damit auch für die uns vorliegenden Daten.

5 Parameterschätzung im bivariaten Probit-Modell

Kapitel 5 befasst sich ausführlich mit der Schätzung der auftretenden Parameter im bivariaten Probit-Modell. Wie schon im univariaten Fall geschieht dies mit der Maximum-Likelihood-Methode. Zunächst entwickeln wir die Log-Likelihoodfunktion und die Scoregleichungen ebenso wie die einzelnen Komponenten der Hessematrix. Dabei folgen wir im Wesentlichen der Notation von Greene (2003). Besonderes Augenmerk richten wir danach auf Kriterien zu Existenz und Eindeutigkeit der Maximum-Likelihood-Schätzer. Im Anschluss daran erläutern wir kurz alternative Möglichkeiten zur Modellanpassung.

5.1 Log-Likelihood und Scoregleichungen

Um die für die Maximum Likelihood-Schätzung nötigen Scoregleichungen zu erhalten, müssen wir eine Darstellung der Log-Likelihood im bivariaten Probit-Modell entwickeln.

Sei im Folgenden $\rho_i = \rho$ konstant für $i = 1, \ldots, n$. Wir betrachten also das ursprüngliche bivariate Probit-Modell nach Ashford & Sowden (1970). Um die Likelihood-Gleichung aufzustellen, benötigt man einige elementare Eigenschaften der bivariaten Standardnormalverteilung, die in Anhang A.4 zusammengefasst sind.

Zur Beobachtung $(Y_{i1} = y_{i1}, Y_{i2} = y_{i2})$ definiere $\delta_{i1} := 2y_{i1} - 1$ und $\delta_{i2} := 2y_{i2} - 1$, also

$$\delta_{ij} = \begin{cases} 1, & \text{falls } Y_{ij} = 1 \\ -1, & \text{falls } Y_{ij} = 0 \end{cases}, \quad j = 1, 2\,,$$

für $i = 1, \ldots, n$. Setze des Weiteren $z_{ij} := \delta_{ij}\, \boldsymbol{\beta}_1' \boldsymbol{x_i}\,,\ j = 1, 2\,,$ sowie $\varrho_i := \delta_{i1}\delta_{i2}\, \rho\,.$

Damit lässt sich $P(Y_{i1} = y_{i1}, Y_{i2} = y_{i1}\,|\boldsymbol{x_i})$ für $y_{i1}, y_{i2} = 0, 1$ unter Benutzung von (A.11) - (A.14) ausdrücken als

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}\,|\boldsymbol{x_i}) = \Phi_2(z_{i1}, z_{i2}, \varrho_i), \tag{5.1}$$

wobei die Faktoren $\delta_{ij},\ j = 1, 2$ die jeweils notwendige Anpassung der Vorzeichen gewährleisten. Die gesamte Likelihoodfunktion ergibt sich entsprechend als

$$L(\beta_1, \beta_2, \rho) = \prod_{i=1}^{n} P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}|\boldsymbol{x_i}) = \prod_{i=1}^{n} \Phi_2(z_{i1}, z_{i2}, \varrho_i) \tag{5.2}$$

und die Log-Likelihood als

$$l(\beta_1, \beta_2, \rho) = \ln L(\beta_1, \beta_2, \rho) = \sum_{i=1}^{n} \ln \Phi_2(z_{i1}, z_{i2}, \varrho_i) \ . \tag{5.3}$$

Unter Verwendung der Funktionen g_1 und g_2 aus (4.27), (4.28) lassen sich daraus die Scoregleichungen für das bivariate Probit-Modell entwickeln. Setze

$$\Phi_{2i} := \Phi_2(z_{i1}, z_{i2}, \varrho_i) \quad \text{und} \quad \phi_{2i} := \phi_2(z_{i1}, z_{i2}, \varrho_i) \ .$$

Dann gilt mit (A.7) und (A.8)

$$\frac{\partial l(\beta_1, \beta_2, \rho)}{\partial \boldsymbol{\beta}_j} = \frac{\partial l(\beta_1, \beta_2, \rho)}{\partial \Phi_{2i}} \cdot \frac{\partial \Phi_{2i}}{\partial z_{ij}} \cdot \frac{\partial z_{ij}}{\partial (\boldsymbol{\beta}_j' \boldsymbol{x_i})} \cdot \frac{\partial (\boldsymbol{\beta}_j' \boldsymbol{x_i})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} \frac{1}{\Phi_{2i}} \, g_{ij} \, \delta_{ij} \, \boldsymbol{x_i} \ , \quad j = 1, 2 \, , \tag{5.4}$$

mit

$$g_{i1} := g_1(z_{i1}, z_{i2}, \varrho_i) = \phi(z_{i1}) \, \Phi \left(\frac{z_{i2} - \varrho_i z_{i1}}{\sqrt{1 - \varrho_i^2}} \right)$$

$$g_{i2} := g_2(z_{i1}, z_{i2}, \varrho_i) = \phi(z_{i2}) \, \Phi \left(\frac{z_{i1} - \varrho_i z_{i2}}{\sqrt{1 - \varrho_i^2}} \right) .$$

Des Weiteren gilt

$$\frac{\partial l(\beta_1, \beta_2, \rho)}{\partial \rho} = \frac{\partial l(\beta_1, \beta_2, \rho)}{\partial \Phi_{2i}} \cdot \frac{\partial \Phi_{2i}}{\partial \varrho_i} \cdot \frac{\partial \varrho_i}{\partial \rho} = \sum_{i=1}^{n} \frac{1}{\Phi_{2i}} \, \phi_{2i} \, \delta_{i1} \delta_{i2} \ . \tag{5.5}$$

Dabei benutzt man in (5.5) die von Plackett (1954) bewiesene Tatsache, dass für die Dichte $\phi_n(x_1, \ldots, x_n \, ; \Sigma)$ jeder (nicht-singulären) n-dimensionalen Standardnormalverteilung

$$\frac{\partial \, \phi_n}{\partial \rho_{ij}} = \frac{\partial^2 \, \phi_n}{\partial x_i \partial x_j} \ , \quad i \neq j = 1, \ldots, n$$

gilt für jeden Eintrag ρ_{ij} der Korrelationsmatrix Σ. Da die Vertauschung von Differentiation und Integration hier zulässig ist, folgt daraus für unsere Situation sofort

$$\frac{\partial \Phi_{2i}}{\partial \varrho_i} = \frac{\partial \Phi_{2i}}{\partial z_{i1} \partial z_{i2}} = \phi_{2i} \, . \tag{5.6}$$

Die Maximum-Likelihood (ML)-Schätzer $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ und $\hat{\rho}$ erhält man, indem man die Scoregleichungen (5.4)–(5.5) gleich Null setzt. Dies ergibt ein nichtlineares Gleichungssystem, dessen Lösung man über einen iterativen Prozess bestimmt.

Mithilfe einiger Vereinfachungen lassen sich auch die zweiten Ableitungen und damit die Hesse-Matrix bestimmen: Setze zunächst

$$\lambda_i := \frac{1}{\sqrt{1 - \varrho_i^2}} \ , \ w_{i1} := \frac{z_{i2} - \varrho_i z_{i1}}{\sqrt{1 - \varrho_i^2}} = \lambda_i (z_{i2} - \varrho_i z_{i1}) \ , \ w_{i2} := \frac{z_{i1} - \varrho_i z_{i2}}{\sqrt{1 - \varrho_i^2}} = \lambda_i (z_{i1} - \varrho_i z_{i2}) \, . \tag{5.7}$$

Dann gilt

$$g_{i1} = \phi(z_{i1})\ \Phi(w_{i1})\ ,\ g_{i2} = \phi(z_{i2})\ \Phi(w_{i2})\ , \tag{5.8}$$

und aus (A.2)

$$\phi_{2i}(z_{i1}, z_{i2}, \varrho_i) = \lambda_i\ \phi(z_{i1})\ \phi(w_{i1}) = \lambda_i\ \phi(z_{i2})\ \phi(w_{i2})\ . \tag{5.9}$$

Mit

$$\phi'(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\bigl(-\frac{1}{2}\,z^2\bigr) \cdot (-z) = -z \cdot \phi(z)\ , \tag{5.10}$$

$$\frac{\partial \Phi_{2i}}{\partial \boldsymbol{\beta}_j} = g_{ij}\ \delta_{ij}\ \boldsymbol{x_i} \quad \text{und} \quad \frac{\partial \Phi_{2i}}{\partial \rho} = \frac{\partial \Phi_{2i}}{\partial \varrho_i} \cdot \frac{\partial \varrho_i}{\partial \rho} \overset{(5.6)}{=} \phi_{2i}\ \delta_{i1}\delta_{i2} \tag{5.11}$$

ergibt sich zunächst

$$\frac{\partial g_{i1}}{\partial \boldsymbol{\beta}_1} = \frac{\partial g_{i1}}{\partial z_{i1}} \cdot \frac{\partial z_{i1}}{\partial \boldsymbol{\beta}_1} = \left[\phi'(z_{i1})\ \Phi(w_{i1}) + \phi(z_{i1})\ \frac{\partial \Phi(w_{i1})}{\partial w_{i1}} \cdot \frac{\partial w_{i1}}{\partial z_{i1}}\right] \cdot \frac{\partial z_{i1}}{\partial \boldsymbol{\beta}_1}$$

$$\overset{(5.10)}{=} \bigl[-z_{i1}\ \underbrace{\phi(z_{i1})\Phi(w_{i1})}_{(5.8)} + \underbrace{\phi(z_{i1})\ \phi(w_{i1})\ (-\lambda_i}\ \varrho_i)\bigr]\delta_{i1}\ \boldsymbol{x_i} = \ [-z_{i1}\, g_{i1} - \varrho_i\, \phi_{2i}\,]\delta_{i1}\ \boldsymbol{x_i}\ , \tag{5.12}$$

$$\frac{\partial g_{i1}}{\partial \boldsymbol{\beta}_2} = \frac{\partial g_{i1}}{\partial z_{i2}} \cdot \frac{\partial z_{i2}}{\partial \boldsymbol{\beta}_2} = \left[\phi(z_{i1})\ \frac{\partial\, \Phi(w_{i1})}{\partial w_{i1}} \cdot \frac{\partial w_{i1}}{z_{i2}}\right] \cdot \frac{\partial z_{i2}}{\partial \boldsymbol{\beta}_2} = \underbrace{[\phi(z_{i1})\ \phi(w_{i1})\ \lambda_i]}_{(5.9)}\delta_{i2}\ \boldsymbol{x_i} = \phi_{2i}\ \delta_{i2}\ \boldsymbol{x_i}\ , \tag{5.13}$$

$$\frac{\partial g_{i1}}{\partial \rho} = \frac{\partial g_{i1}}{\partial \varrho_i} \cdot \frac{\partial \varrho_i}{\partial \rho} = \left[\phi(z_{i1}) \cdot \frac{\partial\, \Phi(w_{i1})}{\partial\, w_{i1}} \cdot \frac{\partial\, w_{i1}}{\partial \varrho_i}\right] \cdot \frac{\partial \varrho_i}{\partial \rho} =$$

$$= \left[\phi(z_{i1})\ \phi(w_{i1}) \cdot \frac{\sqrt{1-\varrho_i^2} \cdot (-z_{i1}) - (z_{i2} - \varrho_i z_{i1}) \cdot \frac{1}{2\sqrt{1-\varrho_i^2}}\ (-2\varrho_i)}{1-\varrho_i^2}\right] \delta_{i1}\delta_{i2} =$$

$$= \underbrace{\phi(z_{i1})\ \phi(w_{i1})\ \lambda_i}_{(5.9)}\ [-z_{i1} + \varrho_i w_{i1}\lambda_i]\ \delta_{i1}\delta_{i2} = \phi_{2i}\ [-z_{i1} + \varrho_i w_{i1}\lambda_i]\ \delta_{i1}\delta_{i2}\ . \tag{5.14}$$

Analog erhält man

$$\frac{\partial g_{i2}}{\partial \boldsymbol{\beta}_2} = [-z_{i2}\ g_{i2} - \varrho_i\phi_{2i}\,]\ \delta_{i2}\ \boldsymbol{x_i}\ , \quad \frac{\partial g_{i2}}{\partial \boldsymbol{\beta}_1} = \phi_{2i}\,\delta_{i1}\ \boldsymbol{x_i}\ , \quad \frac{\partial g_{i2}}{\partial \rho} = \phi_{2i}\ [-z_{i2} + \varrho_i w_{i2}\lambda_i]\ \delta_{i1}\delta_{i2}\ . \tag{5.15}$$

Unter Verwendung dieser Zwischenresultate ergibt sich

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} &= \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_{2i}\ \delta_{i1} \cdot \frac{\partial g_{i1}}{\partial \boldsymbol{\beta}_1'} - \delta_{i1}\ g_{i1} \cdot \frac{\partial \Phi_{2i}}{\partial \boldsymbol{\beta}_1'}}{\Phi_{2i}^2} \right] \overset{(5.11),(5.12)}{=} \\
& \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_2\ \delta_{i1}[-z_{i1}\ g_{i1} - \varrho_i\ \phi_{2i}]\ \delta_{i1}\ \boldsymbol{x_i'} - \delta_{i1}^2\ g_{i1}^2\ \boldsymbol{x_i'}}{\Phi_{2i}^2} \right] = \\
& \sum_{i=1}^{n} \left[\frac{-z_{i1}\ g_{i1} - \varrho_i\ \phi_{2i}}{\Phi_{2i}} - \frac{g_{i1}^2}{\Phi_{2i}^2} \right] \boldsymbol{x_i x_i'}\ ;
\end{aligned} \tag{5.16}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} = \frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1'} &= \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_{2i}\ \delta_{i1} \cdot \frac{\partial g_{i1}}{\partial \boldsymbol{\beta}_2'} - \delta_{i1}\ g_{i1} \cdot \frac{\partial \Phi_{2i}}{\partial \boldsymbol{\beta}_2'}}{\Phi_{2i}^2} \right] \overset{(5.11),(5.13)}{=} \\
& \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_{2i}\ \delta_{i1}\ \phi_{2i}\ \delta_{i2}\ \boldsymbol{x_i'} - \delta_{i1}\ g_{i1}\ g_{i2}\ \delta_{i2}\ \boldsymbol{x_i'}}{\Phi_{2i}^2} \right] = \\
& \sum_{i=1}^{n} \left[\frac{\phi_{2i}}{\Phi_{2i}} - \frac{g_{i1}\ g_{i2}}{\Phi_{2i}^2} \right] \delta_{i1}\delta_{i2}\ \boldsymbol{x_i x_i'}\ ;
\end{aligned} \tag{5.17}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \rho} &= \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_{2i}\ \delta_{i1} \cdot \frac{\partial g_{i1}}{\partial \rho} - \delta_{i1}\ g_{i1} \cdot \frac{\partial \Phi_{2i}}{\partial \rho}}{\Phi_{2i}^2} \right] \overset{(5.11),(5.14)}{=} \\
& \sum_{i=1}^{n} \boldsymbol{x_i} \left[\frac{\Phi_{2i}\ \delta_{i1}\ \phi_{2i}\ (-z_{i1} + \varrho_i\, w_{i1}\, \lambda_i)\ \delta_{i1}\delta_{i2} - \delta_{i1}\ g_{i1}\ \phi_{2i}\ \delta_{i1}\delta_{i2}}{\Phi_{2i}^2} \right] = \\
& \sum_{i=1}^{n} \left[(-z_{i1} + \varrho_i w_{i1} \lambda_i) - \frac{g_{i1}}{\Phi_{2i}} \right] \cdot \frac{\phi_{2i}}{\Phi_{2i}}\ \delta_{i2}\ \boldsymbol{x_i}\ ;
\end{aligned} \tag{5.18}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial^2 \rho} &= \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \left[\frac{\Phi_{2i} \cdot \frac{\partial \phi_{2i}}{\partial \varrho_i} - \phi_2 \cdot \frac{\partial \Phi_{2i}}{\partial \varrho_i}}{\Phi_{2i}^2} \right] \cdot \frac{\partial \varrho_i}{\partial \rho} \overset{(5.11)}{=} \\
& \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \left[\frac{\phi_{2i}\ [\varrho_i\, \lambda_i^2 - w_{i1}^2\ \varrho_i\, \lambda_i^2 + z_{i1}\, w_{i1}\, \lambda_i] - \phi_2^2}{\Phi_{2i}^2} \right]\ \delta_{i1}\delta_{i2} = \\
& \sum_{i=1}^{n} [\varrho_i \lambda_i^2 (1 - z_{i1}^2) - \varrho_i \lambda_i^4 z_{i2}^2 + 2\varrho_i^2 \lambda_i^4 z_{i1} z_{i2} - \varrho_i^3 \lambda_i^4 z_{i1}^2 + \lambda_i^2 z_{i1} z_{i2}] \cdot \frac{\phi_{2i}}{\Phi_{2i}^2} \\
& \quad - \frac{\phi_{2i}^2}{\Phi_{2i}^2}\ .
\end{aligned} \tag{5.19}$$

Analog berechnet man $\frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2'}$ und $\frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \rho}$. Die Resultate sind die gleichen wie in (5.16), (5.18), mit entsprechend vertauschten Indizes $_{i1}, _{i2}$.

Die Hessematrix setzt sich aus diesen einzelnen Komponenten zusammen als

$$\boldsymbol{H} := \begin{pmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}_1 \partial \rho} \\ \frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1'} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2'} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}_2 \partial \rho} \\ \frac{\partial^2 l}{\partial \rho \partial \boldsymbol{\beta}_1'} & \frac{\partial^2 l}{\partial \rho \partial \boldsymbol{\beta}_2'} & \frac{\partial^2 l}{\partial \rho^2} \end{pmatrix} \in \mathbb{R}^{(2p+3)\times(2p+3)} .$$

Die Maximierung der Log-Likelihood ist beispielsweise über das Newton-Raphson-Verfahren möglich, das z.B. Collett (1999) in allgemeiner Form erläutert. Auch die von uns benutzte Software STATA® verwendet eine modifizierte Version dieses Verfahrens (vgl. Gould & Sribney (1999)). Als Startwerte für $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ im Maximierungsprozess kann man die ML-Schätzer $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2$ der univariaten marginalen Modelle verwenden. Sie sind konsistente Schätzer für die wahren Werte, ausserdem hat sich gezeigt, dass die marginalen Schätzer in der Praxis oftmals nahe bei den gemeinsamen Schätzwerten liegen (Lesaffre & Kaufmann (1992)). Mögliche Startwerte für ρ sind ein von Kiefer (1982) vorgeschlagener Schätzer oder die von Pearson (1901) eingeführte tetrachorische Korrelation. Für das bivariate Probit-Modell mit zwei binären Zielvariablen Y_1 und Y_2 und insgesamt n Beobachtungen berechnet sich diese folgendermaßen:

Sei n_{ij} die Anzahl aller Beobachtungen im Datensatz mit $Y_1 = i$ und $Y_2 = j$ für $i, j = 0, 1$. Dann gilt $\sum_{i,j=0,1} n_{ij} = n$. Aus diesen Daten lässt sich zusammenfassend eine Kontingenztabelle erstellen:

	$Y_1 = 1$	$Y_1 = 0$
$Y_2 = 1$	n_{11}	n_{01}
$Y_2 = 0$	n_{10}	n_{00}

Die relative Häufigkeit von $(Y_1 = 1)$ ist $\frac{n_{10}+n_{11}}{n}$, die relative Häufigkeit von $(Y_2 = 1)$ berechnet sich zu $\frac{n_{01}+n_{11}}{n}$. Damit bestimmt man die Werte $\hat{\eta}_1$ und $\hat{\eta}_2$ so, dass die Gleichungen

$$\Phi(\hat{\eta}_1) = \frac{n_{10} + n_{11}}{n} =: \hat{\pi}_{1+}^{obs} \quad \text{und} \quad \Phi(\hat{\eta}_2) = \frac{n_{01} + n_{11}}{n} =: \hat{\pi}_{+1}^{obs}$$

erfüllt werden. Dabei nutzt man aus, dass dem bivariaten Probit-Modell marginal zwei Standardnormalverteilungen zugrunde liegen (vgl. Abschnitt 4.2). Damit gilt dort für die Ränder $\pi_{1+} := P(Y_1 = 1) = \Phi(\eta_1)$ und $\pi_{+1} := P(Y_2 = 1) = \Phi(\eta_2)$.

Mit $\frac{n_{00}}{n}$, der relativen Häufigkeit von $(Y_1 = 0, Y_2 = 0)$, schätzt man schließlich ρ als denjenigen Wert $\hat{\rho}$, so dass

$$\Phi_2(-\hat{\eta}_1, -\hat{\eta}_2, \hat{\rho}) = \frac{n_{00}}{n} =: \hat{\pi}_{00}^{obs}, \tag{5.20}$$

da im bivariaten Probit-Modell $\pi_{00} := P(Y_1 = 0, Y_2 = 0) \overset{(A.14)}{=} \Phi_2(-\eta_1, -\eta_2, \rho)$ gilt.

5.2 Existenz und Eindeutigkeit des ML-Schätzers

In diesem Abschnitt präsentieren wir einige von Lesaffre & Kaufmann (1992) entwickelte Resultate für die ML-Schätzung bei sogenannten korrelierten Vorhersagemodellen, denen auch das bivariate Probit-Modell zuzuordnen ist. Wir versuchen, die Ergebnisse anhand einfacher Beispiele zu veranschaulichen und skizzieren kurz die relevanten Beweisideen. Diese basieren zum Großteil auf Begriffen und Resultaten aus der Theorie konvexer Funktionen, die Rockafellar (1970) ausführlich behandelt. Die wichtigsten Definitionen finden sich in Anhang B.

5.2.1 Die Klasse der korrelierten Vorhersagemodelle

Die Klasse der korrelierten Vorhersagemodelle wird von Lesaffre und Kaufmann folgendermaßen definiert:

Sei $P(\cdot|\boldsymbol{\rho})$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathbb{R}^k$, die vom Parametervektor $\boldsymbol{\rho}$ aus $C_q \subset \mathbb{R}^q$ abhängt. $\boldsymbol{\rho}$ kann beispielsweise ein Vektor von Korrelationen sein. Man nimmt an, dass es keine Atome in den eindimensionalen Rändern gibt. Eine Beobachtung bestehe aus k binären Größen, die jeweils die Werte $+1$ (Erfolg) und -1 (Misserfolg) annehmen. Bezeichne $\boldsymbol{s} = (s_1, \ldots, s_k)'$ den zugehörigen Responsevektor aus $S := \{-1, +1\}^k$. Definiere zu jeder möglichen Ergebniskonstellation $\boldsymbol{s} \in S$ den geschlossenen Orthanten $O_{\boldsymbol{s}}(\boldsymbol{\gamma})$ mit Ausrichtung $\boldsymbol{s}$ und Ausgangspunkt $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$ als die Menge aller Vektoren $\boldsymbol{z} = (z_1, \ldots, z_k)' \in \mathbb{R}^k$, für die $s_i z_i \leq s_i \gamma_i$ für alle $i = 1, \ldots, k$ gilt. Bezeichne $h_s(\boldsymbol{\gamma}, \boldsymbol{\rho}) := P\{O_{\boldsymbol{s}}(\boldsymbol{\gamma})|\boldsymbol{\rho}\}$ die Wahrscheinlichkeit des Orthanten $O_{\boldsymbol{s}}(\boldsymbol{\gamma})$ für alle $\boldsymbol{s} \in S, \boldsymbol{\gamma} \in \mathbb{R}^k, \boldsymbol{\rho} \in C_q$. Dann gilt $\sum_{\boldsymbol{s} \in S} h_s(\boldsymbol{\gamma}, \boldsymbol{\rho}) = 1$.

Im Zweidimensionalen, dem einfachsten Fall, gibt es vier mögliche Ergebniskonstellationen (1,1), (1,−1), (−1,1) und (−1, −1). Für den Ausgangspunkt $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ werden die zugehörigen Orthanten – in diesem Fall Quadranten – beschrieben durch die Mengen

$$O_{(1,1)}(\boldsymbol{\gamma}) = \{\boldsymbol{z} \in \mathbb{R}^2 : z_1 \leq \gamma_1, z_2 \leq \gamma_2\} \tag{5.21}$$

$$O_{(1,-1)}(\boldsymbol{\gamma}) = \{\boldsymbol{z} \in \mathbb{R}^2 : z_1 \leq \gamma_1, z_2 \geq \gamma_2\} \tag{5.22}$$

$$O_{(-1,1)}(\boldsymbol{\gamma}) = \{\boldsymbol{z} \in \mathbb{R}^2 : z_1 \geq \gamma_1, z_2 \leq \gamma_2\} \tag{5.23}$$

$$O_{(-1,-1)}(\boldsymbol{\gamma}) = \{\boldsymbol{z} \in \mathbb{R}^2 : z_1 \geq \gamma_1, z_2 \geq \gamma_2\} \tag{5.24}$$

In Abbildung 5.1 ist diese Situation graphisch dargestellt.

Beobachtet man zu jedem Responsevektor zusätzlich noch einen Vektor von Kovariablen $\boldsymbol{x}$, dann wird die Wahrscheinlichkeit von Ergebniskonstellation $\boldsymbol{s}$ gegeben $\boldsymbol{x}$ im korrelierten Vorhersagemodell beschrieben durch $\pi_{\boldsymbol{s}} = h_{\boldsymbol{s}}(\boldsymbol{X}'\boldsymbol{\beta}, \boldsymbol{\rho})$ für $\boldsymbol{s} \in S, \boldsymbol{\rho} \in C_q$, wobei $\boldsymbol{\beta}$ ein d-dimensionaler Parametervektor und die $d \times k$ Designmatrix $\boldsymbol{X}$ eine Funktion des Kovariablenvektors $\boldsymbol{x}$ ist. Die Kovariablen beeinflussen also die Lage des Punktes $\boldsymbol{\gamma}$, durch den die Orthanten erzeugt werden.

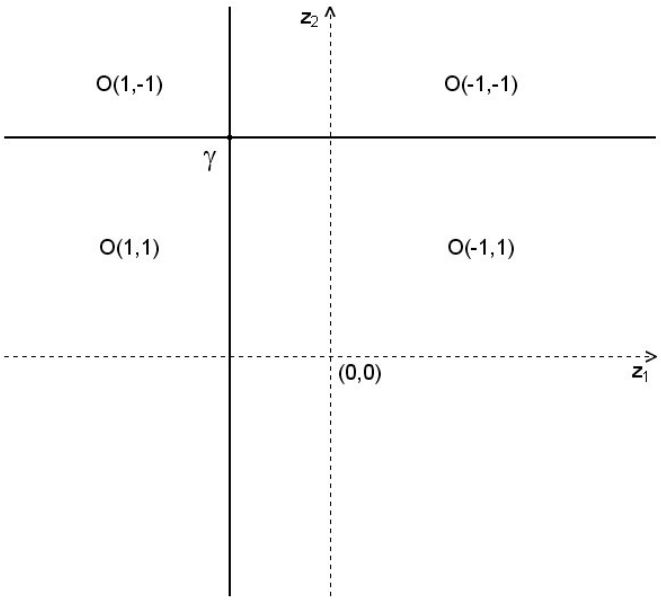


Abbildung 5.1: *Zweidimensionaler Fall: durch* $\boldsymbol{\gamma} = (-1, 2)'$ *erzeugte Orthanten*

Im Folgenden interpretieren wir $\boldsymbol{z}$ als k-dimensionalen latenten, stetigen Zufallsvektor. Insbesondere nimmt s_i genau dann den Wert $+1$ an, wenn die i-te Komponente z_i unterhalb des Schwellenwertes γ_i liegt. Falls dagegen $z_i \geq \gamma_i$ gilt, hat s_i die Ausprägung -1.

Weiterhin nehme man an, dass m unabhängige Gruppen vorliegen, wobei alle Beobachtungen der i-ten Gruppe den Kovariablenvektor $\boldsymbol{x_i}$ aufweisen für $i = 1, \ldots, m$. Bezeichne n_i die Gesamtanzahl von Beobachtungen in der i-ten Gruppe und $y_{i\boldsymbol{s}}$ die Anzahl von Beobachtungen in der i-ten Gruppe mit Response $\boldsymbol{s}$. Dann gilt $\sum_{\boldsymbol{s}\in S} y_{i\boldsymbol{s}} = n_i$, $i = 1, \ldots, m$. Damit ergibt sich für $(y_{i\boldsymbol{s}}, \boldsymbol{s} \in S)$ eine Multinomialverteilung mit jeweils n_i Versuchen und den zugehörigen Wahrscheinlichkeiten $(h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \boldsymbol{\rho}), \boldsymbol{s} \in S) =: \boldsymbol{\pi}_{i\boldsymbol{s}}$. Somit ist die Wahrscheinlichkeit, in der i-ten Gruppe das Ergebnis $(y_{i\boldsymbol{s}}, \boldsymbol{s} \in S)$ zu beobachten,

$$P(y_{i\boldsymbol{s}}, \boldsymbol{s} \in S) = \frac{n_i!}{\prod_{\boldsymbol{s}\in S} y_{i\boldsymbol{s}}!} \prod_{\boldsymbol{s}\in S} h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \boldsymbol{\rho})^{y_{i\boldsymbol{s}}} \ . \tag{5.25}$$

Gilt speziell

$$\boldsymbol{X_i} = I_k \otimes \boldsymbol{x_i}' = \begin{pmatrix} \boldsymbol{x_i}' & \boldsymbol{0} & \ldots & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x_i}' & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{x_i}' & \boldsymbol{0} \\ \boldsymbol{0} & \ldots & \ldots & \boldsymbol{0} & \boldsymbol{x_i}' \end{pmatrix} \quad \text{und} \quad \boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_k') \, , \tag{5.26}$$

dann ergibt sich für die Funktion $\boldsymbol{X_i}'\boldsymbol{\beta}$ der Kovariablen der Vektor

$$\boldsymbol{X_i}'\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1'\boldsymbol{x_i} \, , \, \boldsymbol{\beta}_2'\boldsymbol{x_i} \, , \, \ldots \, , \, \boldsymbol{\beta}_{k-1}'\boldsymbol{x_i} \, , \, \boldsymbol{\beta}_k'\boldsymbol{x_i}\right)' \ \in \mathbb{R}^k,$$

und damit ist $\boldsymbol{\pi}_{i\boldsymbol{s}} = (h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \boldsymbol{\rho}), \boldsymbol{s} \in S) = (h_{\boldsymbol{s}}(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \ldots, \boldsymbol{\beta}_k'\boldsymbol{x_i}, \boldsymbol{\rho}), \boldsymbol{s} \in S)$. Die marginalen Wahrscheinlichkeiten der j-ten binären Zielgröße werden in diesem Falle also nur vom Vektor $\boldsymbol{\beta}_j$ kontrolliert, andere Komponenten von $\boldsymbol{\beta}$ haben darauf keinen Einfluss.

Wir werden nun zeigen, dass das bivariate Probit-Modell mit konstantem Korrelationsparameter ρ zur Klasse der korrelierten Vorhersagemodelle gehört, allerdings nur in gruppierter Form:

Seien $Y_r = (Y_{r1}, Y_{r2})'$, $r = 1, \ldots, n$ die n Einzelbeobachtungen eines bivariaten binären Responses. Fasse wieder alle Beobachtungen mit gleichem Kovariablenvektor $\boldsymbol{x_i} \in \mathbb{R}^{p+1}$ zusammen. So entstehen $m \leq n$ Gruppen mit jeweils n_i , $i = 1, \ldots, m$ Beobachtungen, so dass $\sum_{i=1}^m n_i = n$ gilt.

Bezeichne Y_{i11}^* die Anzahl von Beobachtungen in der i-ten Gruppe mit Responsekombination $(1,1)$. Definiere analog Y_{i10}^* , Y_{i01}^* und Y_{i00}^*. Dann sind die $Y_i^* := (Y_{i11}^*, Y_{i10}^*, Y_{i01}^*, Y_{i00}^*)'$ für $i = 1, \ldots, m$ jeweils multinomial verteilt mit n_i Wiederholungen und Wahrscheinlichkeitsvektor $\boldsymbol{\pi}(\boldsymbol{x_i}) := \big(\pi_{11}(\boldsymbol{x_i}), \pi_{10}(\boldsymbol{x_i}), \pi_{01}(\boldsymbol{x_i}), \pi_{00}(\boldsymbol{x_i})\big)'$, also

$$Y_i^* \sim M(n_i, \boldsymbol{\pi}(\boldsymbol{x_i})),$$

wobei $\pi_{11}(\boldsymbol{x_i}), \ldots, \pi_{00}(\boldsymbol{x_i})$ wieder die Wahrscheinlichkeiten der vier durch den Teilpunkt $(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})$ erzeugten Quadranten unter der bivariaten Standardnormalverteilung mit Korrelationsparameter ρ bezeichnen. Die Wahrscheinlichkeit, in der i-ten Gruppe das Ergebnis $(Y_{ijk}^* = y_{ijk},\ j,k = 0,1)$ zu beobachten, beträgt

$$P(Y_{ijk}^* = y_{ijk},\ j,k = 0,1) = \frac{n_i!}{\prod_{j,k=0,1}\ y_{ijk}!} \prod_{j,k=0,1} \pi_{jk}(\boldsymbol{x_i})^{y_{ijk}} \ .$$

Im bivariaten Probit-Modell mit konstantem Korrelationsparameter ρ ist die Anzahl k der binären Zielvariablen also durch $k = 2$ bestimmt. Der Parametervektor $\boldsymbol{\rho}$ ist in diesem Fall eindimensional. Setze $q = 1$ und $C_q := [-1\,,\,1] \subset \mathbb{R}$. Ersetze des Weiteren in obiger Notation den Index 0 durch -1. Dann ergibt sich die Menge aller möglichen Responsekombinationen S im bivariaten Probit-Modell als $S = \{-1,+1\}^2$. Setze $\boldsymbol{\beta}' := (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2') \in \mathbb{R}^d = \mathbb{R}^{2p+2}$ und

$$\boldsymbol{X_i} := I_2 \otimes \boldsymbol{x_i}' = \begin{pmatrix} \boldsymbol{x_i}' & 0 \\ 0 & \boldsymbol{x_i}' \end{pmatrix}. \tag{5.27}$$

Definiert man nun für alle $\boldsymbol{s} \in S$ die Wahrscheinlichkeit $h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \rho)$ als

$$h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \rho) \overset{(5.27)}{=} h_{\boldsymbol{s}}(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho) := \Phi_2(s_1\boldsymbol{\beta}_1'\boldsymbol{x_i},\ s_2\boldsymbol{\beta}_2'\boldsymbol{x_i},\ s_1 s_2 \rho)\,, \tag{5.28}$$

so erhält man ein spezielles korreliertes Vorhersagemodell der Form (5.26). Nach Anhang A beschreiben die in (5.28) definierten $h_{\boldsymbol{s}}(\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i}, \rho)\,, \boldsymbol{s} \in S$ die Wahrscheinlichkeiten der vier Orthanten aus (5.21)– (5.24) zum Ausgangspunkt $\boldsymbol{\gamma}_i := (\boldsymbol{\beta}_1'\boldsymbol{x_i}, \boldsymbol{\beta}_2'\boldsymbol{x_i})'$ unter der bivariaten Standardnormalverteilung mit Korrelation ρ.

Analog kann man zeigen, dass alle Modelle der Form (5.26), bei denen $h_{\boldsymbol{s}}$ durch das Integrieren über eine zugrunde liegende k-dimensionale Standardnormalverteilungsdichte mit Korrelationsvektor $\boldsymbol{\rho} \in C_q$ erzeugt wird, zur Klasse der korrelierten Vorhersagemodelle gehören. Sie werden als multivariate Probit-Modelle bezeichnet.

5.2.2 Der Maximum-Likelihood-Schätzer $\hat{\beta}$ für festes $\boldsymbol{\rho}$

Zunächst wollen wir nur korrelierte Vorhersagemodelle mit festem $\boldsymbol{\rho}$ betrachten. Da in diesem Fall nur $\boldsymbol{\beta}$ geschätzt zu werden braucht, verschwindet in diesem Abschnitt $\boldsymbol{\rho}$ aus der Notation. Aus (5.25) ergibt sich die gemeinsame Likelihoodfunktion als

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} P(y_{i\boldsymbol{s}}, \boldsymbol{s} \in S) = \prod_{i=1}^{m} \frac{n_i!}{\prod_{\boldsymbol{s}\in S} y_{i\boldsymbol{s}}!} \prod_{\boldsymbol{s}\in S} h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta})^{y_{i\boldsymbol{s}}}.$$

Damit erhält man die Log-Likelihood des korrelierten Vorhersagemodells bei bekanntem $\boldsymbol{\rho}$ als

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{\boldsymbol{s}\in S} y_{i\boldsymbol{s}} \log h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}) + \sum_{i=1}^{m} \log \frac{n_i!}{\prod_{\boldsymbol{s}\in S} y_{i\boldsymbol{s}}!} \quad , \quad \boldsymbol{\beta} \in \mathbb{R}^d. \tag{5.29}$$

Der zweite Term kann bei der Bestimmung des Maximums von $l(\boldsymbol{\beta})$ vernachlässigt werden, da er nur von den gegebenen Daten, nicht aber von $\boldsymbol{\beta}$ abhängt. Um hinreichende und notwendige Kriterien für die Existenz und Eindeutigkeit eines ML-Schätzers zu finden, benötigt man einige Begriffe und Resultate aus der Theorie konvexer Funktionen. Diese werden in Anhang B zusammengefasst.

Der effektive Definitionsbereich dom l der Log-Likelihood ist gegeben durch

$$\operatorname{dom} l := \{\boldsymbol{\beta} : \boldsymbol{X_i}'\boldsymbol{\beta} \in \Gamma_{\boldsymbol{s}},\ i = 1, \ldots, m,\ \boldsymbol{s} \in S,\ y_{i\boldsymbol{s}} > 0\},$$

wobei $\Gamma_{\boldsymbol{s}} := \{\boldsymbol{\gamma} \in \mathbb{R}^k : h_{\boldsymbol{s}}(\boldsymbol{\gamma}) > 0\}$. Falls dieser Definitionsbereich nicht leer ist, so lässt sich der Abstiegskegel rec l von l bestimmen. Dieser enthält all jene Vektoren $\boldsymbol{\beta}$, die

$$y_{i\boldsymbol{s}}\, s_j\, \boldsymbol{x_{ij}}'\boldsymbol{\beta} \geq 0, \quad i = 1, \ldots, m\,, \quad j = 1, \ldots, k\,, \quad \boldsymbol{s} \in S \tag{5.30}$$

erfüllen (Beweis siehe Theorem 5.1), wobei $\boldsymbol{x_{ij}}$ die j-te Spalte von $\boldsymbol{X_i}$ bezeichnet.

Dies kann man mithilfe der marginalen Häufigkeiten umformulieren: Sei y_{ij+} (y_{ij-}) die Anzahl der Beobachtungen zum Kovariablenvektor $\boldsymbol{x_i}$, deren j-te Responsekomponente den Wert $+1$ (-1) hat, so dass $y_{ij+} + y_{ij-} = n_i$. Für festes i und j gilt nach (5.30) zunächst

$$\begin{aligned} y_{i\boldsymbol{s}}\, \boldsymbol{x_{ij}}'\boldsymbol{\beta} &\geq 0 \quad \forall \boldsymbol{s} \in S : s_j = +1, \\ y_{i\boldsymbol{s}}\, \boldsymbol{x_{ij}}'\boldsymbol{\beta} &\leq 0 \quad \forall \boldsymbol{s} \in S : s_j = -1. \end{aligned}$$

und damit auch

$$\sum_{\substack{\boldsymbol{s}\in S:\\ s_j=+1}} y_{i\boldsymbol{s}} \boldsymbol{x_{ij}}'\boldsymbol{\beta} = \boldsymbol{x_{ij}}'\boldsymbol{\beta} \cdot \overbrace{\sum_{\substack{\boldsymbol{s}\in S:\\ s_j=+1}} y_{i\boldsymbol{s}}}^{y_{ij+}} \geq 0 \quad , \quad \sum_{\substack{\boldsymbol{s}\in S:\\ s_j=-1}} y_{i\boldsymbol{s}} \boldsymbol{x_{ij}}'\boldsymbol{\beta} = \boldsymbol{x_{ij}}'\boldsymbol{\beta} \cdot \overbrace{\sum_{\substack{\boldsymbol{s}\in S:\\ s_j=-1}} y_{i\boldsymbol{s}}}^{y_{ij-}} \leq 0\,.$$

Also lässt sich (5.30) schreiben als

$$\begin{aligned} &y_{ij+}\, \boldsymbol{x_{ij}}'\boldsymbol{\beta} \geq 0\,, \\ &y_{ij-}\, \boldsymbol{x_{ij}}'\boldsymbol{\beta} \leq 0\,, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k. \end{aligned} \tag{5.31}$$

Im Spezialfall (5.26), bei dem $\boldsymbol{x_{ij}}'\boldsymbol{\beta} = \boldsymbol{\beta}_j'\boldsymbol{x_i}$ gilt, enthält der Abstiegskegel also alle Vektoren $\boldsymbol{\beta}' = (\beta_1', \ldots, \beta_k')$, die das Ungleichungssytem

$$\begin{aligned} &y_{ij+}\,\boldsymbol{\beta}_j'\boldsymbol{x_i} \geq 0\,, \\ &y_{ij-}\,\boldsymbol{\beta}_j'\boldsymbol{x_i} \leq 0\,, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k \end{aligned} \tag{5.32}$$

erfüllen. Im Hinblick auf die Existenz und Eindeutigkeit eines Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\beta}}$ im korrelierten Vorhersagemodell gilt folgendes Theorem:

Theorem 5.1.
(a) Falls der Erwartungswert $\mathbb{E}(\boldsymbol{z}'\boldsymbol{z})$ zum latenten stetigen Zufallsvektor $\boldsymbol{z}$ endlich oder $h_{\boldsymbol{s}}$ log-konkav ist, so existiert der ML-Schätzer $\hat{\boldsymbol{\beta}}$ genau dann, wenn kein $\boldsymbol{\beta} \neq \boldsymbol{0}$ (5.31) erfüllt.
Im Spezialfall $\boldsymbol{x_{ij}}'\boldsymbol{\beta} = \boldsymbol{\beta}_j'\boldsymbol{x_i}$ existiert der ML-Schätzer $\hat{\boldsymbol{\beta}}$, falls zu jedem $\boldsymbol{\beta} \neq \boldsymbol{0}$ ein Index i existiert, so dass für alle j gilt

$$y_{ij+}\,\boldsymbol{\beta}_j'\boldsymbol{x_i} < 0 \quad \textit{oder} \quad y_{ij-}\,\boldsymbol{\beta}_j'\boldsymbol{x_i} > 0.$$

(b) Falls dom $l \neq \emptyset$, und $h_{\boldsymbol{s}}$ für alle $\boldsymbol{s} \in S$ entweder strikt log-konkav auf $\mathbb{R}^k$ oder log-konkav auf $\mathbb{R}^k$ ist mit Definitheitseigenschaft, dann existiert ein eindeutiger Parameterschätzer $\hat{\boldsymbol{\beta}}$ genau dann, wenn kein $\boldsymbol{\beta} \neq \boldsymbol{0}$ (5.31) erfüllt.

Beweis. (Skizze)
(a) Chung (1974) zeigt, dass $h_{\boldsymbol{s}}(\boldsymbol{\gamma})$ stetig in $\boldsymbol{\gamma}$ ist. Zudem ist leicht zu sehen, dass für festes $\boldsymbol{s} \in S$

$$\operatorname{rec} h_{\boldsymbol{s}}(\boldsymbol{\gamma}) = -O_{\boldsymbol{s}}(\boldsymbol{0}) = O_{-\boldsymbol{s}}(\boldsymbol{0})$$

gilt: $h_{\boldsymbol{s}}(\boldsymbol{\gamma} + \lambda\boldsymbol{d})$ fällt als Funktion von $\lambda \in \mathbb{R}$ genau dann nicht, wenn sich der Orthant $O_{\boldsymbol{s}}(\boldsymbol{\gamma} + \lambda\boldsymbol{d})$ mit wachsendem λ nicht verkleinert, d.h. wenn für $\boldsymbol{d}$

$$s_j d_j \geq 0\,, \quad j = 1, \ldots, k$$

gilt. Somit ist $\operatorname{rec} h_{\boldsymbol{s}}(\boldsymbol{\gamma}) = \{\boldsymbol{d} : s_j d_j \geq 0\,,\ j = 1, \ldots, k\} = \{\boldsymbol{d} : -s_j d_j \leq 0\,,\ j = 1, \ldots, k\} = O_{-\boldsymbol{s}}(\boldsymbol{0})$. Mithilfe eines Beweises ähnlich dem von Lemma 1 in Kaufmann (1988b) kann man zeigen, dass der Abstiegskegel $\operatorname{rec} l$ der gesamten Likelihood dann durch (5.30) gegeben ist. Um Teil (a) zu vervollständigen, verwendet man den Beweis zu Theorem 3 in Kaufmann (1988c) mit Transformationsmatrizen $T_s := \operatorname{diag}(-s_1, \ldots, -s_k) \in \mathbb{R}^{k \times k}$, oder Theorem 27.1 in Rockafellar (1970).

(b) Log $h_{\boldsymbol{s}}$ ist auf $\Gamma_{\boldsymbol{s}}$ konkav mit Definitheitseigenschaft sowie stetig in γ auf $\mathbb{R}^k$. Da dom $l \neq \emptyset$, muss man lediglich zeigen, dass h injektiv ist, denn dann sind die Voraussetzungen der Theoreme 3 und 4 in Kaufmann (1988a) erfüllt. Man nehme also an, dass zwei Vektoren $\boldsymbol{\gamma} \neq \tilde{\boldsymbol{\gamma}} \in \mathbb{R}^k$ existieren mit $h_{\boldsymbol{s}}(\boldsymbol{\gamma}) = h_{\boldsymbol{s}}(\tilde{\boldsymbol{\gamma}}), \boldsymbol{s} \in S$. Betrachte nun $w_{\boldsymbol{s}}(t) := \log h_{\boldsymbol{s}}\left(\boldsymbol{\gamma} + t(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right), \boldsymbol{s} \in S, t \in \mathbb{R}$. Man kann davon ausgehen, dass $w_{\boldsymbol{s}}(t)$ nicht für alle $\boldsymbol{s}$ konstant ist. Da $\log h_{\boldsymbol{s}}$ konkav ist, müsste $w_{\boldsymbol{s}}(t)$ dann zumindest für ein $\boldsymbol{s}$ streng monoton wachsend sein im Intervall $[-\epsilon, +\epsilon]$, mit passend gewähltem ϵ. Dies ist aber nicht möglich, da für jedes $\boldsymbol{\gamma}$ stets $\sum_{\boldsymbol{s} \in S} h_{\boldsymbol{s}}(\boldsymbol{\gamma}) = 1$ gelten muss. □

Ein ML-Schätzer für das korrelierte Vorhersagemodell existiert also nach Theorem 5.1(a) unter gewissen Regularitätsbedingungen, wenn rec $l = \{\mathbf{0}\}$ gilt. Dies stellt ein Kriterium für die Datenkonstellation dar. Existiert zu einem Datensatz ein $\boldsymbol{\beta} \neq \mathbf{0}$, für das (5.31) gilt, so treten Divergenzprobleme auf. Für die Eindeutigkeit des ML-Schäzters müssen nach Teil (b) noch weitere Bedingungen erfüllt sein.

Falls $h_{\boldsymbol{s}}$ durch das Integrieren über einer zugrunde liegenden multivariaten Dichte berechnet wird, reicht es, die Eigenschaften jener Dichtefunktion zu untersuchen, da sich diese auf $h_{\boldsymbol{s}}$ übertragen. So zeigt z.B. Prékopa (1973), dass $h_{\boldsymbol{s}}$ strikt log-konkav ist, falls die Dichtefunktion strikt log-konkav auf $\mathbb{R}^k$ ist.

Die multivariate (und damit insbesondere die bivariate) Dichte der Standardnormalverteilung $N_k(\boldsymbol{\mu}, \Sigma)$ ist (im nicht-degenerierten Fall) strikt log-konkav, da die zu betrachtende Hessematrix in diesem Fall $-\Sigma^{-1}$, also die negative Inverse der Korrelationsmatrix ist, und Σ^{-1} ist stets positiv definit. Daher lässt sich Theorem 5.1 sofort auf die multivariaten Probit-Modelle anwenden. Anschaulich interpretiert existiert für diese speziellen Modelle Theorem 5.1 und dem Ungleichungssystem (5.32) zufolge kein ML-Schätzer, falls die Daten der Erfolge und Misserfolge jeder einzelnen Responsekomponente s_j jeweils durch eine Hyperebene $\boldsymbol{\beta}_j'\boldsymbol{x_i} = 0$ voneinander getrennt werden können.

Für das bivariate Probit-Modell bedeutet Theorem 1 insbesondere, dass es keinen ML-Schätzer gibt, falls ein Vektor $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ existiert, der

$$\begin{aligned} y_{i1+}\, \boldsymbol{\beta}_1'\boldsymbol{x_i} &\geq 0, \\ y_{i1-}\, \boldsymbol{\beta}_1'\boldsymbol{x_i} &\leq 0 \end{aligned} \tag{5.33}$$

und gleichzeitig

$$\begin{aligned} y_{i2+}\, \boldsymbol{\beta}_2'\boldsymbol{x_i} &\geq 0, \\ y_{i2-}\, \boldsymbol{\beta}_2'\boldsymbol{x_i} &\leq 0, \end{aligned} \tag{5.34}$$

für alle $i = 1, \ldots, m$, erfüllt.

In Abbildung 5.2 ist solch eine Datenkonstellation für ein Modell mit zwei Kovariablen graphisch dargestellt. Die Erfolge in der ersten Responsekomponente s_1 (Symbole $\triangle$ und $\times$) sind von den Misserfolgen (Symbole $\circ$ und $*$) durch die Gerade $\boldsymbol{\beta}_1'\boldsymbol{x} = 0$ separiert: die Daten zu $s_1 = +1$ liegen stets oberhalb davon im mit „+“ gekennzeichneten Bereich, diejenigen zu $s_1 = -1$ dagegen unterhalb von $\boldsymbol{\beta}_1'\boldsymbol{x} = 0$ (gekennzeichnet mit „−“). Die Erfolge (Symbole $*$ und $\triangle$) und die Misserfolge (Symbole $\times$ und $\circ$) der zweiten Komponente s_2 trennt die Gerade $\boldsymbol{\beta}_2'\boldsymbol{x} = 0$ voneinander. Die Erfolge liegen dabei wieder ober– und die Misserfolge unterhalb davon. Auch diese beiden Bereiche sind entsprechend mit „+“ bzw. „−“ kenntlich gemacht.

Für das bivariate Probit-Modell reduziert sich die Frage nach der Existenz des ML-Schätzers also auf das Existenzproblem der beiden marginalen univariaten Probit-Modelle,

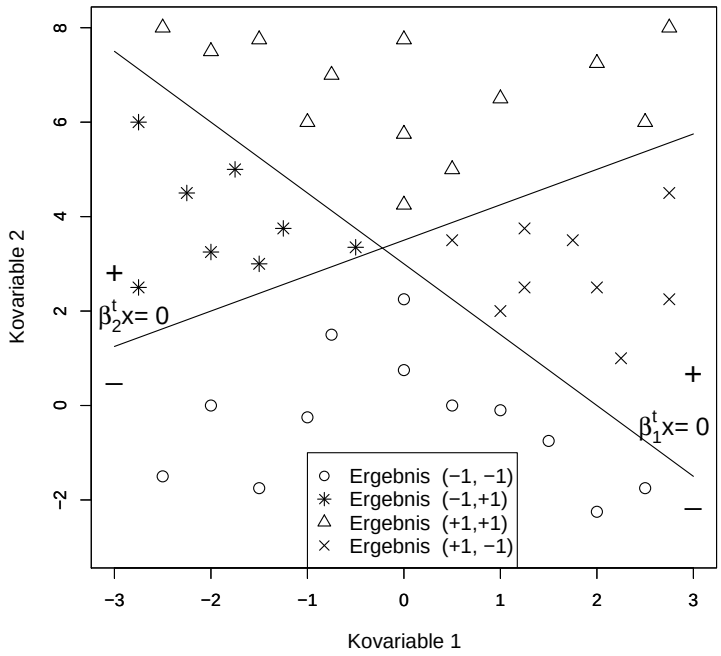


Abbildung 5.2: *Datenkonstellation im bivariaten Probit-Modell, bei der kein ML-Schätzer existiert: Die Daten zu Erfolgen und Misserfolgen in den einzelnen Komponenten sind vollständig separabel.*

das bereits in Abschnitt 3.3 angesprochen wurde: Die Bedingungen (5.33) und (5.34) für die marginalen Klassen entsprechen jeweils genau der Bedingung (3.8) im univariaten Fall.

Dies ist aber nicht nur beim bivariaten Probit-Modell der Fall. Theorem 2 gilt allgemein für alle Modelle der Form (5.26) und setzt Existenz und Eindeutigkeit des ML-Schätzers des gesamten Modells in Beziehung zur Existenz und Eindeutigkeit der ML-Schätzer für die marginalen Modelle. Dazu benötigt man zunächst folgende Definition:

Definition 5.1. *Eine multivariate Dichte auf $\mathbb{R}^k$ heißt* ***k-strikt (log)-konkav****, falls die gemeinsame Dichte sowie alle marginalen Dichten strikt (log)-konkav sind.*

Bezeichne im Folgenden $\hat{\boldsymbol{\beta}}$ den ML-Schätzer des gesamten Modells im Spezialfall (5.26), und sei $\hat{\boldsymbol{\beta}}_j$ die j-te Komponente dieses ML-Schätzers. Bezeichne weiterhin den ML-Schätzer des marginalen Modells zur j-ten Responsekomponente mit $\tilde{\boldsymbol{\beta}}_j$. Dann gilt

Theorem 5.2.
(a) Falls der Erwartungswert $\mathbb{E}(\boldsymbol{z}'\boldsymbol{z})$ endlich oder $h_{\boldsymbol{s}}$ log-konkav ist, so existiert $\hat{\boldsymbol{\beta}}_j$ genau dann, wenn $\tilde{\boldsymbol{\beta}}_j$ existiert.

(b) Falls die zugrunde liegende Dichtefunktion f k-strikt log-konkav auf $\mathbb{R}^k$ ist, dann existiert ein eindeutiger Schätzer $\hat{\boldsymbol{\beta}}_j$ genau dann, wenn $\tilde{\boldsymbol{\beta}}_j$ existiert und eindeutig ist.

Da – wie wir in Abschnitt 5.1 bereits erläuterten – für die Schätzung von $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')$ im bivariaten Probit-Modell oftmals die Schätzer $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2$ der univariaten marginalen Modelle als Startwerte verwendet werden, ergibt sich aus Theorem 5.2 eine einfache Möglichkeit, Divergenz frühzeitig zu erkennen und zu signalisieren. Ein gutes Programm sollte also zuerst die beiden marginalen Modelle anpassen. Falls schon hier Divergenzprobleme auftreten, kann man das Programm abbrechen lassen, noch bevor mit der gemeinsamen Schätzung begonnen wird.

Ein Beispiel für eine Datenkonstellation zu zwei Kovariablen, in der zwar $\hat{\boldsymbol{\beta}}_1$ (und $\tilde{\boldsymbol{\beta}}_1$) existieren, $\hat{\boldsymbol{\beta}}_2$ (und $\tilde{\boldsymbol{\beta}}_2$) dagegen nicht, ist in Abbildung 5.3 graphisch dargestellt.

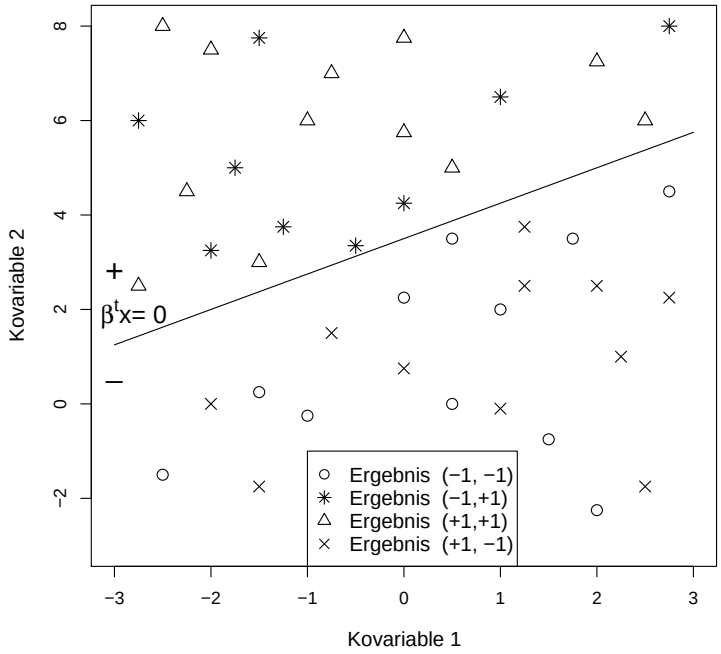


Abbildung 5.3: *Datenkonstellation im bivariaten Probit-Modell, bei der nur der marginale ML-Schätzer zur ersten Responsekomponente existiert.*

In dieser Graphik gibt es keine Gerade, mit der man die Daten der Erfolge und Misserfolge in der ersten Responsekomponente voneinander trennen kann, weshalb ein marginaler ML-Schätzer existiert. Für die zweite Komponente ist eine vollständige Trennung der Erfolge und Misserfolge allerdings möglich; somit gibt es für das entsprechende marginale Modell keinen ML-Schätzer. Auch das gemeinsame bivariate Modell ist deshalb nach Theorem 5.2 nicht schätzbar.

Bemerkenswert ist die Tatsache, dass für festes ρ der Schätzer $\hat{\boldsymbol{\beta}}$ bei entsprechend günstiger Datenkonstellation auch dann existieren kann, wenn eine oder mehrere mögliche Responsekombinationen im Datensatz gar nicht auftreten.

5.2.3 Der gemeinsame Maximum-Likelihood-Schätzer $(\hat{\beta}, \hat{\rho})$

Für eine gleichzeitige Schätzung von $\boldsymbol{\beta}$ und $\boldsymbol{\rho}$ stellt sich die Log-Likelihood analog zu (5.29) folgendermaßen dar:

$$l(\boldsymbol{\beta}, \boldsymbol{\rho}) = \sum_{i=1}^{m} \sum_{\boldsymbol{s} \in S} y_{i\boldsymbol{s}} \log h_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta}, \boldsymbol{\rho}) \quad , \quad \boldsymbol{\beta} \in \mathbb{R}^d, \ \boldsymbol{\rho} \in C_q, \tag{5.35}$$

wobei C_q kompakt und $q = \frac{1}{2}k(k-1)$ ist.

Da wir im vorherigen Abschnitt das Verhalten von $\hat{\boldsymbol{\beta}}$ bedingt auf $\boldsymbol{\rho}$ bestimmt haben, werden wir auch die Profil-Loglikelihood $pl(\boldsymbol{\rho}) := \sup_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \boldsymbol{\rho})$ betrachten.

Aussagen über die Existenz des gemeinsamen ML-Schätzers $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$ werden dadurch erschwert, dass $h_{\boldsymbol{s}}$ als Funktion von $\boldsymbol{\rho}$ nicht (strikt) log-konkav ist, wie wir jetzt anhand eines Beispiels zeigen werden. Mithilfe einiger elementarer Eigenschaften der bivariaten Standardnormalverteilung (siehe z.B. Sheppard (1899) oder Johnson & Kotz (1972)) kann man zeigen, dass im bivariaten Probit-Modell für $\rho \geq 0$

$$h_{\boldsymbol{s}}(\mathbf{0}, \rho) = \frac{1}{4} + \frac{1}{2\pi} sin^{-1}\rho$$

gilt. Die zweite Ableitung von $\log h_{\boldsymbol{s}}(\mathbf{0}, \rho)$ ist

$$\frac{d^2}{d\rho^2} h_{\boldsymbol{s}}(\mathbf{0}, \rho) = \frac{2\pi\rho\, h_{\boldsymbol{s}}(\mathbf{0}, \rho) - \sqrt{1-\rho^2}}{4\pi^2\sqrt{1-\rho^2}\,(1-\rho^2)\, h_{\boldsymbol{s}}^2(\mathbf{0}, \rho)},$$

und mit $\lim_{\rho \to 0} h_{\boldsymbol{s}}(\mathbf{0}, \rho) = \frac{1}{4}$ gilt

$$\lim_{\rho \to 0} \frac{d^2}{d\rho^2} h_{\boldsymbol{s}}(\mathbf{0}, \rho) = \frac{0-1}{4\pi^2 \cdot \left(\frac{1}{4}\right)^2} = -\frac{4}{\pi^2} < 0.$$

Wegen $\lim_{\rho \to 1} h_{\boldsymbol{s}}(\mathbf{0}, \rho) = \frac{1}{2}$ ist der Zähler der zweiten Ableitung für $\rho \to 1$ positiv und durch $2\pi \cdot 1 \cdot \frac{1}{2} = \pi$ beschränkt, während der Nenner von oben gegen Null konvergiert. Insgesamt ist $\frac{d^2}{d\rho^2} h_{\boldsymbol{s}}(\mathbf{0}, \rho)$ in der Nähe von 1 also positiv, und $\log h_{\boldsymbol{s}}(\mathbf{0}, \rho)$ ist dort nicht konkav.

Wir nehmen an, dass für jedes $\boldsymbol{\rho} \in C_q$ keine Atome in den eindimensionalen Rändern existieren. Bezeichne $P_{\boldsymbol{\rho}}$ das zu $\boldsymbol{\rho}$ gehörige Wahrscheinlichkeitsmaß und $F_{\boldsymbol{\rho}}$ die entsprechende Verteilungsfunktion. Weiterhin gelte, dass $\boldsymbol{\rho}_n$ genau dann gegen $\boldsymbol{\rho}$ konvergiert, wenn $F_{\boldsymbol{\rho}_n}$ an den Stetigkeitspunkten von $F_{\boldsymbol{\rho}}$ gegen $F_{\boldsymbol{\rho}}$ konvergiert. Dies garantiert die Stetigkeit von $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho})$ in beiden Argumenten. Dies ist auch dann richtig, falls $F_{\boldsymbol{\rho}}$ eine degenerierte Verteilungsfunktion ist. Klar ist weiterhin, dass $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho})$ log-konkav ist, falls $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho}_n)$ log-konkav ist für alle $\boldsymbol{\rho}_n$, und dass rec l (bezüglich $\boldsymbol{\beta}$) für alle $\boldsymbol{\rho}$ gleich ist. Die Definitheitseigenschaft überträgt sich jedoch im Allgemeinen nicht von den $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho}_n)$ auf $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho})$.

Unter diesen Annahmen gilt folgendes Resultat bezüglich der Existenz des ML-Schätzers $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$:

Theorem 5.3. *Falls $l(\boldsymbol{\beta}, \boldsymbol{\rho})$ in* (5.35) *nicht identisch $-\infty$ ist und $\mathbb{E}(\boldsymbol{z}'\boldsymbol{z}|\boldsymbol{\rho})$ als Funktion von $\boldsymbol{\rho}$ nach oben halbstetig, endlich und nicht identisch 0 ist, dann nehmen sowohl die Log-Likelihoodfunktion als auch die Profil-Loglikelihood ihr Supremum genau dann an, wenn kein $\boldsymbol{\beta} \neq \mathbf{0}$* (5.31) *erfüllt.*

Beweis. (Skizze)
Zunächst kann man mithilfe einiger standardmäßiger Stetigkeitsargumente folgendes Resultat beweisen: Falls $l(\boldsymbol{\beta}, \boldsymbol{\rho})$ eine (nach oben halb–)stetige Funktion auf $\mathbb{R}^p \times C$ ist, $l(\boldsymbol{\beta}, \boldsymbol{\rho})$ nicht identisch $-\infty$ ist und es eine Funktion $h(\|\boldsymbol{\beta}\|)$ gibt, so dass $l(\boldsymbol{\beta}, \boldsymbol{\rho}) \leq h(\|\boldsymbol{\beta}\|)$ für alle $\rho \in C$, C kompakt, und $h(\|\boldsymbol{\beta}\|) \to -\infty$ für $\|\boldsymbol{\beta}\| \to \infty$, dann ist die Profil-Likelihood ebenfalls (nach oben halb–)stetig, und sowohl $l(\boldsymbol{\beta}, \boldsymbol{\rho})$ als auch $pl(\boldsymbol{\rho})$ nehmen beide ihr Maximum an.

Solch eine Funktion $h(\|\boldsymbol{\beta}\|)$ kann man unter den Voraussetzungen des Theorems angeben: Sei $\Pi_{\boldsymbol{s}}(\boldsymbol{\gamma})$ die Projektion von $\boldsymbol{\gamma}$ auf den Orthanten $O_{\boldsymbol{s}}(\mathbf{0})$. Dann liefert die Ungleichung

$$\|\Pi_{\boldsymbol{s}}(\boldsymbol{\gamma})\|^2 \, h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho}) \leq \mathbb{E}(\boldsymbol{z}'\boldsymbol{z}|\boldsymbol{\rho})$$

eine dominierende Funktion für die Log-Likelihood: Aufgrund der Annahmen ist $r := \sup_{\boldsymbol{\rho} \in C} \mathbb{E}(\boldsymbol{z}'\boldsymbol{z}|\boldsymbol{\rho})$ endlich, positiv, und wird tatsächlich angenommen. Damit gilt

$$2 \log \|\Pi_{\boldsymbol{s}}(\boldsymbol{\gamma})\| + \log h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho}) \leq c := \log r \ , \quad -\infty < c < \infty.$$

Die Funktion

$$f_{\boldsymbol{s}}(\boldsymbol{\gamma}) := \min\{0 \, , \ c - 2 \log \|\Pi_{\boldsymbol{s}}(\boldsymbol{\gamma})\|\}$$

ist negativ, stetig, und beschränkt $\log h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho})$ nach oben. Zudem garantiert rec $l = \{\mathbf{0}\}$, dass $\|\Pi_{\boldsymbol{s}}(\boldsymbol{\gamma})\| \to \infty$ für $\|\boldsymbol{\beta}\| \to \infty$, und damit gilt $f_{\boldsymbol{s}}(\boldsymbol{\gamma}) \to -\infty$. Daraus folgt schließlich

$$l(\boldsymbol{\beta}, \boldsymbol{\rho}) \leq \sum_{y_{i\boldsymbol{s}} > 0} f_{\boldsymbol{s}}(\boldsymbol{X_i}'\boldsymbol{\beta})$$

und mit dem Resultat zu Anfang des Beweises ist Theorem 5.3 vollständig bewiesen. □

Auch dieses Theorem lässt sich auf das multivariate Probit-Modell übertragen. Hierbei ist C_q der Raum aller $k \times k$ Korrelationsmatrizen. Dieser ist kompakt und alle singulären Matrizen liegen auf dem Rand. Zudem gilt im multivariaten Probit-Modell

$$\mathbb{E}(\boldsymbol{z}'\boldsymbol{z}|\boldsymbol{\rho}) = \mathbb{E}\big(\sum_{i=1}^{k} z_i^2 \,|\, \boldsymbol{\rho}\big) = \sum_{i=1}^{k} \mathbb{E}(z_i^2 |\, \boldsymbol{\rho}) = \sum_{i=1}^{k} 1 = k \ ,$$

so dass die Voraussetzungen des Theorems offensichtlich erfüllt sind.

Über die Eindeutigkeit des gemeinsamen ML-Schätzers $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$ im korrelierten Vorhersagemodell gibt es keine allgemeingültigen Aussagen. Speziell für multivariate Probit-Modelle gilt aber zumindest folgendes Resultat:

Theorem 5.4. *Falls kein $\boldsymbol{\beta} \neq \mathbf{0}$ (5.31) erfüllt, dann existiert der ML-Schätzer $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$ für das multivariate Probit-Modell. Zudem ist für jede feste Korrelationsmatrix, die nicht auf dem Rand von C_q liegt, der ML-Schätzer von $\boldsymbol{\beta}$ eindeutig.*

Falls das Maximum im multivariaten Probit-Modell auf dem Rand von C_q angenommen wird, dann ist $h_{\boldsymbol{s}}(\boldsymbol{\gamma}, \boldsymbol{\rho})$ nicht mehr log-konkav in $\boldsymbol{\gamma}$, so dass in diesem Fall die Eindeutigkeit von $\hat{\boldsymbol{\beta}}$ verlorengeht. Im bivariaten Probit-Modell würde dieser Fall eintreten, wenn sich $\hat{\rho} = 1$ ergibt.

Für das multivariate Probit-Modell lassen sich dazu hinreichende Bedingungen für die Existenz eines Schätzers im Inneren des Parameterraums angeben: Bezeichne die Menge $\{\boldsymbol{s} \,|\, s_i = 1\,,\, s_j = -1\}$ als $S^{ij}_{1,-1}$. Definiere analog die Mengen $S^{ij}_{1,1}, S^{ij}_{-1,1}$ und $S^{ij}_{-1,-1}$. Definiere weiterhin $A^{ij}_{l,m}$ für $l, m = 1, -1$ als die Menge aller Kovariablenvektoren $\boldsymbol{x}_i$, zu denen eine Responsekombination $\boldsymbol{s}$ aus $S^{ij}_{l,m}$ beobachtet wird. Man kann zeigen, dass (für $i \neq j = 1, \ldots, k$) eine geschätzte Korrelation von $\hat{\rho}_{ij} = 1$ unmöglich ist, wenn die Daten $A^{ij}_{1,-1}$ und $A^{ij}_{-1,1}$ nicht vollständig separabel sind. $\hat{\rho}_{ij} = -1$ ist dagegen unmöglich, wenn die Daten $A^{ij}_{1,1}$ und $A^{ij}_{-1,-1}$ nicht vollständig getrennt werden können.

Betrachtet man speziell den bivariaten Fall, so müsste für eine vollständige Trennbarkeit der Daten $A^{12}_{1,-1}$ und $A^{12}_{-1,1}$ ein Vektor $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' \neq \mathbf{0}$ existieren, so dass für jedes $\boldsymbol{x_i}$, zu dem die Ergebniskonstellation (1,−1) vorliegt, die Ungleichungen

$$\begin{aligned} \boldsymbol{\beta}_1' \boldsymbol{x_i} &\geq 0 \\ \boldsymbol{\beta}_2' \boldsymbol{x_i} &\leq 0 \end{aligned} \tag{5.36}$$

gelten. Zudem müsste dieser Vektor $\boldsymbol{\beta}$ für jedes $\boldsymbol{x_i}$, zu dem man die Responsekombination (−1,1) beobachtet, das Ungleichungssystem

$$\begin{aligned} \boldsymbol{\beta}_1' \boldsymbol{x_i} &\leq 0 \\ \boldsymbol{\beta}_2' \boldsymbol{x_i} &\geq 0 \end{aligned} \tag{5.37}$$

erfüllen. Existieren also nun beispielsweise zu jedem Vektor $\boldsymbol{\beta}$ eine oder mehrere Beobachtungen mit der Responsekombination (1,−1), deren Kovariablenvektor $\boldsymbol{x_i}$ die Ungleichung $\boldsymbol{\beta}_2' \boldsymbol{x_i} > \boldsymbol{\beta}_1' \boldsymbol{x_i}$ erfüllt, und/oder gilt für Beobachtungen zur Kombination (−1,1) die umgekehrte Ungleichung $\boldsymbol{\beta}_2' \boldsymbol{x_i} < \boldsymbol{\beta}_1' \boldsymbol{x_i}$, so ist eine vollständige Trennung der Daten und damit eine geschätzte Korrelation von $\hat{\rho} = 1$ unmöglich: Denn für diese Beobachtungen gilt dann $\boldsymbol{\beta}_2' \boldsymbol{x_i} > \boldsymbol{\beta}_1' \boldsymbol{x_i} \geq 0$ bzw. $\boldsymbol{\beta}_2' \boldsymbol{x_i} < \boldsymbol{\beta}_1' \boldsymbol{x_i} \leq 0$, und somit ist Bedingung (5.36) und/oder (5.37) nicht erfüllt.

Sowohl beim bivariaten Probit- als auch beim logistischen Modell handelt es sich um sogenannte (voll-)parametrische Modelle: Hierbei wird nicht nur die marginale Abhängigkeit der einzelnen Response-Komponenten von den Kovariablen und die Korrelation modelliert. Darüber hinaus werden auch explizite Annahmen über die zugrunde liegende gemeinsame Verteilung getroffen, was eine Maximum-Likelihood-Schätzung überhaupt erst ermöglicht. Wie wir im nächsten Abschnitt sehen werden, ist es nicht immer nötig, diese Verteilungsfunktion vollständig festzulegen.

5.3 Die Generalized Estimating Equations-Methode

Einen solchen Ansatz zur Modellierung schlagen Liang & Zeger (1986) vor. Verwendet werden dabei sogenannte „Generalized Estimating Equations“ (GEE), die eine konkrete Spezifizierung der zugrunde liegenden bivariaten Verteilung für die Parameterschätzung überflüssig machen. Auch bei diesen Modellen werden die marginalen Kovariableneffekte und, in der Erweiterung des Modells von Prentice (1988), die zugrunde liegende Korrelationsstruktur geschätzt. Über den gesamten Einfluss der Kovariablen auf die gemeinsame bivariate Zielgröße wird dabei allerdings keine explizite Aussage gemacht.

Die GEE-Methode weist einige Vorteile gegenüber dem parametrischen Ansatz auf (siehe z.B. Lesaffre & Kaufmann (1992)): Die Schätzer sind konsistent und zudem auch oftmals effizienter. Des Weiteren ist der Rechenaufwand weitaus geringer. Für Anwendungen, in denen das Hauptaugenmerk auf den marginalen Effekten liegt und die Korrelation eher als Störgröße angesehen wird, ist der GEE-Ansatz daher oftmals vorzuziehen.

Jedoch ist es manchmal essentiell, individuelle Vorhersagen für die Wahrscheinlichkeiten jeder einzelnen Kombination der beiden univariaten Zielgrößen machen zu können. Als Beispiel lassen sich klinische Studien anführen, in denen man nicht nur daran interessiert ist, das Auftreten einzelner Krankheiten in Abhängigkeit von bestimmten Kovariablen zu beschreiben, sondern aufgrund von Patientendaten eine Vorhersage für die gemeinsame Zielgröße machen möchte. In solchen Fällen ist eine konkrete Annahme hinsichtlich der gemeinsamen Verteilung absolut unerlässlich.

Da man auch im vorliegenden Datenbeispiel nicht nur an den marginalen Effekten der Kovariablen interessiert ist, sondern eine explizite Aussage über das jeweilige Auftreten der vier verschiedenen Responsekategorien machen möchte, kommt hier nur ein vollparametrischer Ansatz in Betracht.

6 Zusammenhang zwischen Korrelation und Odds Ratio

Den sogenannten Odds Ratio haben wir in allgemeiner Form bereits in Abschnitt 4.5 eingeführt. Nun wollen wir diesen speziell für die bivariate Standardnormalverteilung näher betrachten und dabei insbesondere versuchen, die Zusammenhänge zwischen dem Korrelationsparameter ρ und dem Odds Ratio zu ergründen.

6.1 Der Odds Ratio für die bivariate Standardnormalverteilung

Wie in Abschnitt 4.5.2 definieren wir den Odds Ratio zu zwei Zufallsvariablen X und Y an der Stelle (x, y) als

$$\psi(x,y) := \frac{P(X \le x, Y \le y)\, P(X > x, Y > y)}{P(X \le x, Y > y)\, P(X > x, Y \le y)} = \frac{P(X \le x, Y \le y)}{P(X \le x, Y > y)} \Big/ \frac{P(X > x, Y \le y)}{P(X > x, Y > y)} \tag{6.1}$$

Falls die gemeinsame Verteilung der beiden Zufallsvariablen eine bivariate Standarnormalverteilung ist, so kann man dies auch formulieren als

$$\psi(x,y) = \frac{\Phi_2(x,y,\rho)\,[1 - \Phi(x) - \Phi(y) + \Phi_2(x,y,\rho)]}{[\Phi(x) - \Phi_2(x,y,\rho)]\,[\Phi(y) - \Phi_2(x,y,\rho)]} = \psi(x,y,\rho), \quad x, y \in \mathbb{R},\ \rho \in [-1,1] \tag{6.2}$$

wobei Φ_2 und Φ wieder die Verteilungsfunktion der bivariaten bzw. univariaten Standardnormalverteilung bezeichnen. Der Odds Ratio hängt hier also sowohl von der Lage des Punktes (x, y) als auch vom Korrelationsparameter ρ ab. Unter Verwendung einiger grundlegender Eigenschaften der bivariaten Standardnormalverteilung (siehe Anhang A) ergibt sich daraus die Darstellung

$$\psi(x,y,\rho) = \frac{\Phi_2(x,y,\rho)\ \cdot\ \Phi_2(-x,-y,\rho)}{\Phi_2(x,-y,-\rho)\ \cdot\ \Phi_2(-x,y,-\rho)}\ . \tag{6.3}$$

Offensichtlich gilt

$$\psi(-x,y,-\rho) = \frac{\Phi_2(-x,y,-\rho)\ \cdot\ \Phi_2(x,-y,-\rho)}{\Phi_2(-x,-y,\rho)\ \cdot\ \Phi_2(x,y,\rho)} = \frac{1}{\psi(x,y,\rho)}\ . \tag{6.4}$$

Weitere wichtige Beziehungen sind

$$\psi(x, y, \rho) = \psi(y, x, \rho) \tag{6.5}$$

$$\psi(-x, -y, \rho) = \psi(x, y, \rho) \tag{6.6}$$

$$\psi(x, -y, -\rho) = \frac{1}{\psi(x, y, \rho)} \tag{6.7}$$

$$\psi(x, y, -\rho) = \frac{1}{\psi(-x, y, \rho)} = \frac{1}{\psi(x, -y, \rho)} \tag{6.8}$$

Für fest gewählte Werte x und y kann man ψ als Funktion von ρ untersuchen. Abbildung 6.1 zeigt den Kurvenverlauf von $\psi(\rho)$ im Intervall $[-0.6\,, 0.6]$ für verschiedene Werte x, y. Graphik 6.2 zeigt den Ausschnitt von $[0\,, 0.4]$ in vergrößerter Form.

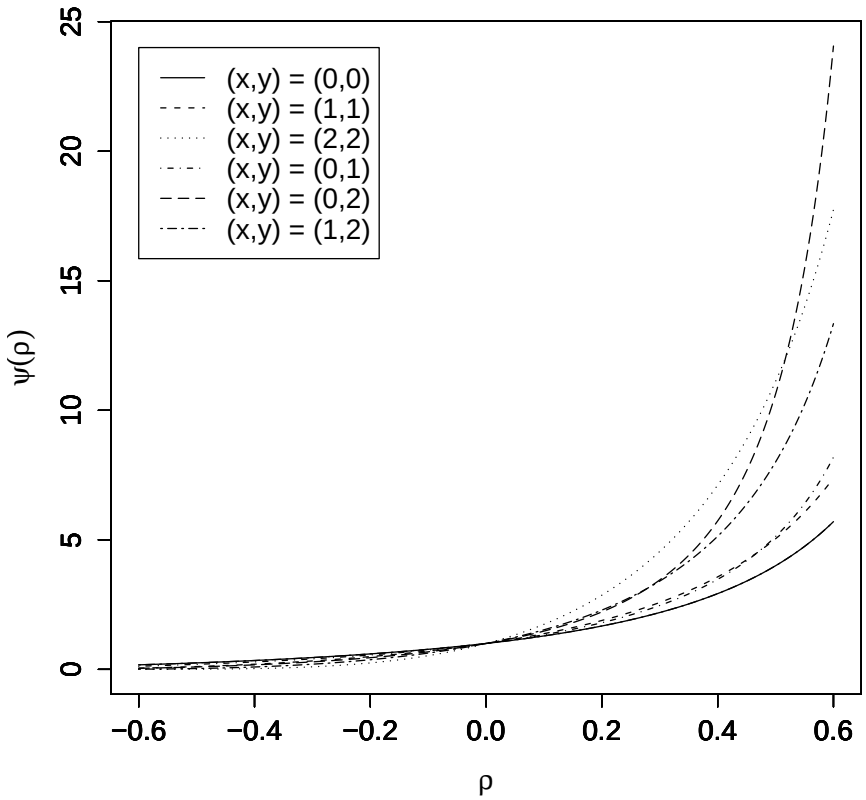


Abbildung 6.1: *Kurvenverlauf von $\psi(\rho)$ zu verschiedenen Wertepaaren (x, y) für $\rho \in [-0.6\,,\, 0.6]$*

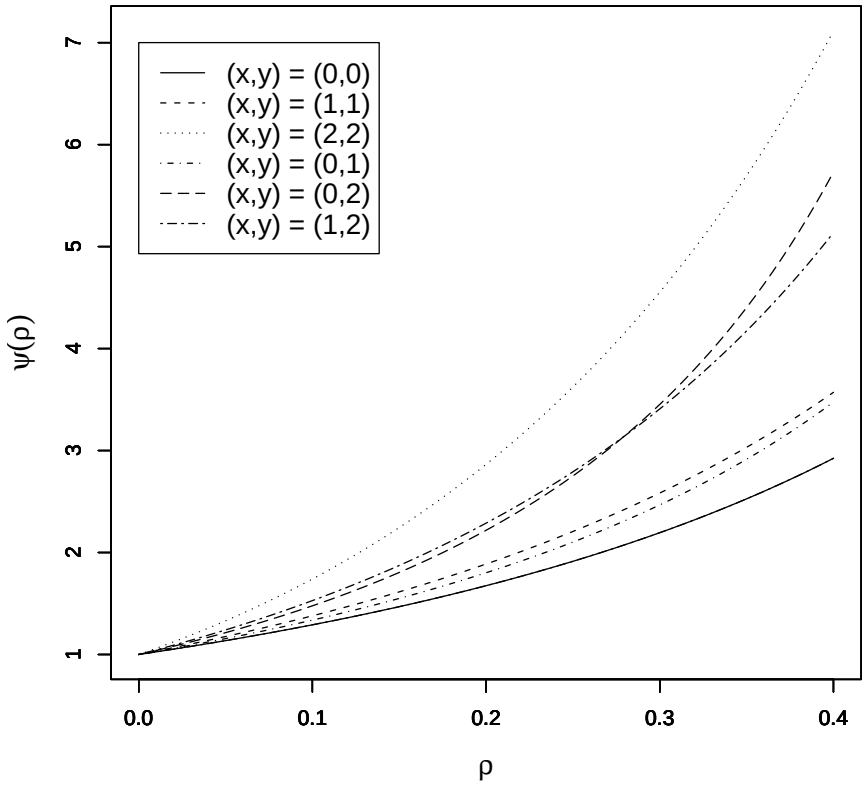


Abbildung 6.2: *Kurvenverlauf von $\psi(\rho)$ zu verschiedenen Wertepaaren (x, y) für $\rho \in [0\,,\,0.4]$*

(6.4) – (6.8) zufolge repräsentieren diese sechs Kurven eine Vielzahl von Wertepaaren. So entspricht der Plot zu $(x, y) = (1, 1)$ nach Eigenschaft (6.6) exakt dem zu $(x, y) = (-1, -1)$, die Kurve zu $(x, y) = (1, 2)$ ist beispielsweise mit denjenigen zu $(x, y) = (2, 1)$, $(x, y) = (-1, -2)$ und $(x, y) = (-2, -1)$ vollkommen identisch.

Der Punkt $(0, 1)$, an dem die beiden Zufallsvariablen X und Y stochastisch unabhängig sind, ist allen Kurven gemeinsam: Aus (6.3) ergibt sich bei $\rho = 0$ sofort

$$\psi(x, y, 0) = \frac{\Phi(x)\,\Phi(y)\,\Phi(-x)\,\Phi(-y)}{\Phi(x)\,\Phi(-y)\,\Phi(-x)\,\Phi(y)} = 1 \quad \forall\, x, y \in \mathbb{R}.$$

Für $\rho \in [0\,,\,1]$ sind die Unterschiede zwischen den verschiedenen Kurven sehr ausgeprägt. Für ρ im Bereich $[-1\,,\,0]$ nimmt $\psi(x, y, \rho)$ dagegen nur Werte zwischen 0 und 1 an. Eine Betrachtung von $\frac{1}{\psi(\rho)}$ ist daher in diesem Bereich sinnvoller, da Änderungen von $\psi(\rho)$ dabei deutlicher zutage treten.

Der Odds Ratio lässt sich hier – genau wie in Abschnitt 4.5 – wieder als Verhältnis zweier Chancen interpretieren: Darstellung (6.1) kann man auch mit bedingten Wahrscheinlichkeiten formulieren. Man erhält

$$\psi(x, y, \rho) = \frac{P(X \leq x \,|\, Y \leq y\,;\, \rho)}{P(X > x \,|\, Y \leq y\,;\, \rho)} \Big/ \frac{P(X \leq x \,|\, Y > y\,;\, \rho)}{P(X > x \,|\, Y > y\,;\, \rho)}\,,$$

also die Chance von $(X \leq x)$ zu $(X > x)$, falls $Y \leq y$ gilt, im Verhältnis zur Chance von $(X \leq x)$ zu $(X > x)$ unter der Bedingung $Y > y$.

Aufgrund dieses Zusammenhanges zwischen Odds Ratio und dem Korrelationsparameter ist es also möglich, Änderungen von ρ auch auf der Skala der Odds Ratios zu interpretieren, was eine etwas anschaulichere Auslegung erlaubt als der abstraktere Begriff der Korrelation. Das Chancenverhältnis kann sich dabei selbst bei verhältnismäßig kleinen Änderungen in ρ extrem verändern, wie man Abbildung 6.1 entnehmen kann.

6.2 Schätzung des Korrelationsparameters ρ mithilfe des Odds Ratios

Hat man Daten, bei denen n Beobachtungen der beiden Zufallsvariablen X und Y nur in dichotomer Form vorliegen, so ist eine Darstellung dieser Daten in Form einer 2×2 Kontingenztabelle möglich:

	$X \leq x$	$X > x$
$Y \leq y$	a	b
$Y > y$	c	d

Dabei gilt $a + b + c + d = n$. Aus diesen Daten lässt sich der beobachtete Odds Ratio $\hat{\psi}^{obs}$ berechnen als

$$\hat{\psi}^{obs} := \frac{ad}{bc}$$

und daraus kann man bei bekannten x, y die Korrelation von X und Y schätzen, wobei der geschätzte Wert $\hat{\rho}$ die Lösung der Gleichung

$$\psi(x, y, \rho) = \frac{ad}{bc} \tag{6.9}$$

ist, die man z.B. graphisch ermitteln kann. Datensimulationen zeigen, dass die solchermaßen geschätzte Korrelation $\hat{\rho}$ zumindest für x, y im Bereich von $[-1\,,\,1]$ relativ nahe am wahren Wert ρ liegt.

Falls die Schrankenwerte x und y jedoch unbekannt sind, kann man diese zunächst aus den Daten als

$$\hat{x} = \Phi^{-1}\left(\frac{a+c}{n}\right) \tag{6.10}$$

$$\hat{y} = \Phi^{-1}\left(\frac{a+b}{n}\right) \tag{6.11}$$

schätzen. Den geschätzten Korrelationsparameter $\hat{\rho}$ bestimmt man dann mithilfe der Schätzer aus (6.10) – (6.11) als Lösung der Gleichung

$$\psi(\hat{x}, \hat{y}, \rho) = \psi\left(\Phi^{-1}\left(\frac{a+c}{n}\right), \Phi^{-1}\left(\frac{a+b}{n}\right), \rho\right) = \frac{ad}{bc} \ . \tag{6.12}$$

Wie wir jetzt zeigen werden, ist diese Schätzung äquivalent zur Berechnung der von Pearson eingeführten tetrachorischen Korrelation, die wir bereits in Abschnitt 5.1 vorgestellt haben, und die wir im Folgenden mit $\hat{\rho}_{tetr}$ bezeichnen. Für diesen Schätzer bestimmt man zunächst die Schrankenwerte $\hat{x}$ und $\hat{y}$ genauso wie in (6.10)–(6.11). Dann berechnet man $\hat{\rho}_{tetr}$ als Lösung der Gleichung

$$\Phi_2(-\hat{x}, -\hat{y}, \hat{\rho}_{tetr}) = \frac{d}{n} \ . \tag{6.13}$$

Aus den Eigenschaften der bivariaten Standardnormalverteilung (siehe z.B. Johnson & Kotz (1972)) ergibt sich zunächst

$$\Phi(\hat{x}) + \Phi(\hat{y}) - \Phi_2(\hat{x}, \hat{y}, \hat{\rho}_{tetr}) + \Phi_2(-\hat{x}, -\hat{y}, \hat{\rho}_{tetr}) = 1 \ .$$

Daraus erhält man unter Verwendung von (6.10),(6.11) und (6.13)

$$\begin{aligned} \Phi_2(\hat{x}, \hat{y}, \hat{\rho}_{tetr}) &= \Phi(\hat{x}) + \Phi(\hat{y}) + \Phi_2(-\hat{x}, -\hat{y}, \hat{\rho}_{tetr}) - 1 \\ &= \frac{a+c}{n} + \frac{a+b}{n} + \frac{d}{n} - 1 = \frac{n+a}{n} - 1 = \frac{a}{n} \ . \end{aligned} \tag{6.14}$$

Damit kann man wiederum

$$\Phi_2(\hat{x}, -\hat{y}, \hat{\rho}_{tetr}) = \Phi(\hat{x}) - \Phi_2(\hat{x}, \hat{y}, \hat{\rho}_{tetr}) = \frac{a+c}{n} - \frac{a}{n} = \frac{c}{n}$$

und

$$\Phi_2(-\hat{x}, \hat{y}, \hat{\rho}_{tetr}) = \Phi(\hat{y}) - \Phi_2(\hat{x}, \hat{y}, \hat{\rho}_{tetr}) = \frac{a+b}{n} - \frac{a}{n} = \frac{b}{n}$$

herleiten, und somit gilt offensichtlich

$$\psi(\hat{x}, \hat{y}, \hat{\rho}_{tetr}) = \frac{(a/n)\ (d/n)}{(b/n)\ (c/n)} = \frac{ad}{bc} \ .$$

Für $\hat{\rho}_{tetr}$ gilt also nicht nur (6.13), sondern auch (6.12).

Erfüllt umgekehrt $\hat{\rho}$ (6.12), also

$$\psi(\hat{x},\hat{y},\hat{\rho}) \overset{(6.2)}{=} \frac{\Phi_2(\hat{x},\hat{y},\hat{\rho})\left[1-\Phi(\hat{x})-\Phi(\hat{y})+\Phi_2(\hat{x},\hat{y},\hat{\rho})\right]}{\left[\Phi(\hat{x})-\Phi_2(\hat{x},\hat{y},\hat{\rho})\right]\left[\Phi(\hat{y})-\Phi_2(\hat{x},\hat{y},\hat{\rho})\right]} = \frac{ad}{bc} \tag{6.15}$$

für $\hat{x} = \Phi^{-1}\left(\frac{a+c}{n}\right)$ und $\hat{y} = \Phi^{-1}\left(\frac{a+b}{n}\right)$, dann folgt aus einem Vergleich von

$$\frac{ad}{bc} = \frac{(a/n)\,(d/n)}{(b/n)\,(c/n)} = \frac{\frac{a}{n}\cdot\frac{n-(a+c)-(a+b)+a}{n}}{\frac{(a+b)-a}{n}\cdot\frac{(a+c)-a}{n}} = \frac{\frac{a}{n}\cdot\left[1-\Phi(\hat{x})-\Phi(\hat{y})+\frac{a}{n}\right]}{\left[\Phi(\hat{y})-\frac{a}{n}\right]\cdot\left[\Phi(\hat{x})-\frac{a}{n}\right]}$$

mit (6.15) sofort, dass $\Phi_2(\hat{x},\hat{y},\hat{\rho}) = \frac{a}{n}$. Damit kann man wie in (6.14) zeigen, dass

$$\Phi_2(-\hat{x},-\hat{y},\hat{\rho}) = 1-\Phi(\hat{x})-\Phi(\hat{y})+\Phi_2(\hat{x},\hat{y},\hat{\rho}) = 1-\frac{a+c}{n}-\frac{a+b}{n}+\frac{a}{n} = \frac{d}{n}.$$

Erfüllt $\hat{\rho}$ also (6.12), so ist auch (6.13) richtig.

Somit gilt $\hat{\rho} = \hat{\rho}_{tetr}$. Zudem zeigt Hamdan (1970), dass für Daten, die nur in obiger Form vorliegen, die tetrachorische Korrelation $\hat{\rho}_{tetr}$ der Maximum-Likelihood-Schätzer von ρ ist.

6.3 Der Odds Ratio im bivariaten Probit-Modell

Das hier beschriebene Konzept des Odds Ratios lässt sich auch auf das bivariate Probit-Modell übertragen, da man dabei ebenfalls von einer zugrunde liegenden bivariaten Standardnormalverteilung ausgeht.

Die Daten zu n unabhängigen Beobachtungen der beiden binären Zielvariablen Y_1 und Y_2 kann man in folgender Kontingenztabelle zusammenfassen:

	$Y_1 = 1$	$Y_1 = 0$
$Y_2 = 1$	n_{11}	n_{01}
$Y_2 = 0$	n_{10}	n_{00}

Dabei bezeichnet n_{ij} die Anzahl aller Beobachtungen im Datensatz mit $Y_1 = i$ und $Y_2 = j$ für $i, j = 0, 1$. Damit gilt $\sum_{i,j=0,1} n_{ij} = n$. Der beobachtete Odds Ratio $\hat{\psi}^{obs}$ ist hier

$$\hat{\psi}^{obs} = \frac{n_{11}\cdot n_{00}}{n_{01}\cdot n_{10}}\,.$$

Dieser Wert gibt, wie schon in Abschnitt 4.5 erläutert, das Verhältnis der Chancen von $(Y_1 = 1)$ zu $(Y_1 = 0)$ innerhalb der beiden Subpopulationen $(Y_2 = 1)$ und $(Y_2 = 0)$ an.

Problematisch ist hierbei allerdings, dass zwar jeder dieser n Einzelbeobachtungen $Y_i := (Y_{i1}, Y_{i2})'$, $i = 1, \ldots, n$, eine bivariate Standardnormalverteilung mit demselben Korrelationsparameter ρ zugrunde liegt, der Teilpunkt $(\boldsymbol{\beta}_1'\boldsymbol{x_i}\,,\ \boldsymbol{\beta}_2'\boldsymbol{x_i})$ hier jedoch nicht konstant ist, sondern vom jeweiligen Kovariablenvektor $\boldsymbol{x_i}$ abhängt und prinzipiell für jede der n Beobachtungen einen anderen Wert annehmen kann. Trotzdem kann man hier eine Schätzung der Schrankenwerte wie in (6.10) und (6.11) über die marginalen relativen Häufigkeiten vornehmen. Falls die linearen Prädiktoren $\boldsymbol{\beta}_1'\boldsymbol{x_i}$ und $\boldsymbol{\beta}_2'\boldsymbol{x_i}$ für alle Beobachtungen bereits bekannt sind, könnte man auch deren Mittelwerte als Schranken verwenden. Bei beiden Vorgehensweisen muss man aufgrund der Variation in den Teilpunkten aber mit einer größeren Ungenauigkeit des Schätzers $\hat{\rho}$ (bzw. $\hat{\rho}_{tetr}$) rechnen.

Möchte man für unseren Datensatz den Korrelationsparameter in bestimmten Subpopulationen abschätzen, so ist diese Herangehensweise daher eher ungeeignet. Da sich die wahren Korrelationen der Subpopulationen in einem sehr engen Intervall von $[-0.16\,, +0.15]$ zu bewegen scheinen, liefert der Schätzer $\hat{\rho}_{tetr}$ zu ungenaue Werte, um vernünftige Aussagen treffen zu können. Eine Schätzung des kompletten bivariaten Probit-Modells ist deshalb für die meisten Untergruppen unerlässlich.

7 Strategien zur Modellierung großer Datensätze

Die meisten standardmäßigen Software-Pakete verfügen heute über die Option, eine Modellanpassung über schrittweise Variablenselektion durchzuführen, so dass man theoretisch die Möglichkeit hätte, dem System einfach alle zur Verfügung stehenden Daten zu übergeben, und dem Programm die gesamte Arbeit zu überlassen.

Bei großen Datensätzen stößt man jedoch sehr schnell an die Grenzen der Leistungsfähigkeit solcher Programme, vor allem, wenn man auch Interaktionen in die Modellierung miteinbeziehen möchte: So müsste man beim vorliegenden Datensatz mit seinen 34 Kovariablen 595 Interaktionen betrachten! Da man für kategorielle Kovariablen im Allgemeinen eine Dummy-Kodierung verwendet, erhielte man im Datensatz tatsächlich ein Vielfaches dieser Zahl an möglichen Interaktionen. Eine zu hohe Variablenzahl führt allerdings zu großen Problemen, sie vergrößert den ohnehin schon sehr umfangreichen Datensatz immens und hat vor allem dramatische Auswirkung auf die Rechenzeiten und benötigten Speicher, besonders dann, wenn dazu noch eine große Anzahl von Beobachtungen vorliegt.

Deshalb ist speziell bei großen Datensätzen eine sehr sorgfältige Prüfung und Vorauswahl der Kovariablen und Interaktionen nötig. Dies geschieht über eine explorative Analyse der Daten, die auf einer Betrachtung der sogenannten empirischen Logits beruht. Im Folgenden stellen wir für univariate Probit-Modelle solche explorativen Verfahren zur Vorselektion der einzelnen Kovariablen und ihrer Interaktionen vor. Anschließend erläutern wir weitere Probleme, die aus der vorliegenden Datenkonstellation entstehen können, sowie Möglichkeiten zu deren Lösung.

7.1 Identifikation einflussreicher Kovariablen

Um die wichtigsten Kovariablen zu identifizieren, betrachtet man zunächst nur den Einfluss jeweils einer kategorialen Variablen (den sogenannten Haupteffekt) auf die Zielgrößen unter Vernachlässigung der anderen Kovariablen.

Sei Y eine binäre Zielvariable mit Ausprägungen 0 und 1 und sei x eine kategoriale Kovariable mit M Ausprägungen $\{1, \ldots, M\}$. Bezeichne n die Gesamtzahl an Beobachtungen im Datensatz, n_j die Anzahl von Beobachtungen in der j-ten Kategorie von x und n_{1j} die Anzahl von Beobachtungen in der j-ten Kategorie mit $Y = 1$. Definiere analog n_{0j}. Dann gilt $\sum_{j=1}^{M} n_j = n$ und $n_{1j} + n_{0j} = n_j$. Zusammenfassend kann man dann folgende

Kontingenztabelle erstellen:

x :	1	2	...	M
Y: 1	n_{11}	n_{12}	...	n_{1M}
0	n_{01}	n_{02}	...	n_{0M}
total:	n_1	n_2	...	n_M

Berechne für jede Kategorie den empirischen Logit

$$\widehat{\text{Logit}}\,(j) := \frac{n_{1j}}{n_{0j}} = \frac{n_{1j}}{n_j - n_{1j}} = \frac{n_{1j}/n_j}{(n_j - n_{1j})/n_j} = \frac{\hat{p}_{1j}^{obs}}{(1-\hat{p}_{1j}^{obs})}, \quad j = 1, \ldots, M,$$

wobei $\hat{p}_{1j}^{obs} := \frac{n_{1j}}{n_j}$ die aus den Daten geschätzte Wahrscheinlichkeit $P(Y = 1|x = j)$ bezeichnet. $\widehat{\text{Logit}}\,(j)$ ist demnach ein Schätzer für $\frac{P(Y=1|x=j)}{P(Y=0|x=j)}$, also für die Chance auf Erfolg in der j-ten Kategorie.

Diese Schätzer bilden die Grundlage für eine graphische Auswertung: Verändern sich die empirischen Logits, also die geschätzten Erfolgschancen, in einem Plot über die verschiedenen Kategorien hinweg nicht sehr, so hat die Kovariable x kaum Einfluss auf die Zielgröße Y. Treten dagegen starke Schwankungen in den empirischen Logits auf, so zeigt die Kovariable einen deutlichen Effekt auf Y.

Auf diese Weise kann man die Zahl der für die Modellierung verwendeten Kovariablen meist reduzieren. Im vorliegenden Datensatz wurden die Haupteffekte der Kovariablen für `storno` und `abschluss` jeweils gesondert untersucht. Einige Kovariablen konnten aufgrund der empirischen Logits von vornherein ausgeschlossen werden.

Mit den solchermaßen ausgewählten Kovariablen kann man bereits ein einfaches Haupteffektmodell erstellen. Allerdings werden hierbei Interaktionen zwischen den Kovariablen komplett vernachlässigt. Deshalb sollte zu Vergleichszwecken auch ein Interaktionsmodell erstellt werden. Auch dabei ist zunächst eine Vorselektion notwendig, um die enorme Zahl an möglichen Interaktionen so weit wie möglich zu reduzieren.

7.2 Identifikation wichtiger Interaktionen

Für die Hinzunahme von Interaktionen trifft man für eine sinnvolle Modellierung zunächst folgende Einschränkung:

Es werden grundsätzlich nur Interaktionen zwischen den Kovariablen betrachtet, die sich bei der Haupteffektanalyse nach Abschnitt 7.1 als relevant herausgestellt haben. Dies verringert nicht nur die Anzahl der zu untersuchenden Interaktionen immens, sondern bietet noch einen weiteren Vorteil: Ein Interaktionsmodell, das dieses Kriterium erfüllt, lässt sich gut mit dem Haupteffektmodell vergleichen, da die beiden genestet sind.

Den Einfluss einzelner Interaktionen auf eine binäre Zielgröße quantifiziert man ebenfalls mithilfe empirischer Logits: Sei Y wieder die binäre Zielvariable. Seien x_1 und x_2 kategoriale Kovariablen mit Ausprägungen $1, \ldots, K$ bzw. $1, \ldots, M$. Bezeichne n_{ij} die Anzahl von Beobachtungen mit $x_1 = i$ und $x_2 = j$ und bezeichne n_{1ij} (n_{0ij}) die Anzahl von Beobachtungen in Zelle (i, j) mit $Y = 1$ ($Y = 0$). Es gilt $n_{1ij} + n_{0ij} = n_{ij}$. Zusammenfassend ergeben sich folgende Kontingenztabellen:

	$x_1 : 1$	2	$\ldots$	K
$x_2 : 1$	n_{11}	n_{12}	$\ldots$	n_{1K}
2	n_{21}	n_{22}	$\ldots$	n_{2K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
M	n_{M1}	n_{M2}	$\ldots$	n_{MK}

$Y = 1$	$x_1 : 1$	2	$\ldots$	K
$x_2 : 1$	n_{111}	n_{112}	$\ldots$	n_{11K}
2	n_{121}	n_{122}	$\ldots$	n_{12K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
M	n_{1M1}	n_{1M2}	$\ldots$	n_{1MK}

$Y = 0$	$x_1 : 1$	2	$\ldots$	K
$x_2 : 1$	n_{011}	n_{012}	$\ldots$	n_{01K}
2	n_{021}	n_{022}	$\ldots$	n_{02K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
M	n_{0M1}	n_{0M2}	$\ldots$	n_{0MK}

Berechne für jede Zelle (i, j), $\quad i = 1, \ldots, K$, $j = 1, \ldots, M$, den empirischen Logit

$$\widehat{\text{Logit}}\,(i, j) := \frac{n_{1ij}}{n_{0ij}} = \frac{n_{1ij}}{n_{ij} - n_{1ij}} = \frac{n_{1ij}/n_{ij}}{(n_{ij} - n_{1ij})/n_{ij}} = \frac{\hat{p}_{1ij}^{obs}}{(1 - \hat{p}_{1ij}^{obs})},$$

wobei $\hat{p}_{1j}^{obs} := \frac{n_{1ij}}{n_{ij}}$ die geschätzte Wahrscheinlichkeit $P(Y = 1|x_1 = i, x_2 = j)$ bezeichnet. $\widehat{\text{Logit}}\,(i, j)$ ist demnach ein Schätzer für $\frac{P(Y=1|x_1=i,\, x_2=j)}{P(Y=0|x_1=i,\, x_2=j)}$, also für die Chance auf Erfolg in Zelle (i, j).

Für festes $x_1 = I$ lässt sich der Verlauf der empirischen Logits über die verschiedenen Kategorien von x_2 graphisch verfolgen. Ein Plot von $\widehat{\text{Logit}}\,(I, j)$ für $j = 1, \ldots, M$ zeigt die Änderung der Erfolgschancen in Abhängigkeit von x_2 innerhalb der Subpopulation ($x_1 = I$). Verlaufen diese Plots für die verschiedenen Subpopulationen parallel, so liegt keine Interaktion zwischen x_1 und x_2 vor. Sind die Kurven dagegen nicht parallel bzw. überschneiden sie sich sogar, so interagieren x_1 und x_2 miteinander.

Mithilfe dieser Entscheidungsregel lässt sich die Zahl der zu betrachtenden Interaktionen stark reduzieren. Im Datensatz verbleiben damit nur noch Kovariablen und Interaktionen, die tatsächlichen Einfluss auf die binären Zielgrößen zeigen.

Im Zusammenhang mit Interaktionen kann sich durch Kollinearitäten zwischen Kovariablen ein weiteres Problem ergeben. Falls durch direkte kausale Zusammenhänge zwischen zwei Kovariablen bestimmte Ausprägungskombinationen von vornherein ausgeschlossen sind, führt dies zu leeren Zellen in den zugehörigen Kontingenztabellen. Warum diese Datenkonstellation problematisch ist und wie man dies lösen kann, wird im folgenden Abschnitt erläutert.

7.3 Beseitigung von Kollinearitäten im Datensatz

Falls in der Kontingenztabelle zu zwei interagierenden Kovariablen leere Zellen auftreten, entstehen bei standardmäßiger Dummy-Kodierung linear abhängige Dummy-Variablen, so dass die Designmatrix $\boldsymbol{X} := (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T \in \mathbb{R}^{n\times(p+1)}$ nicht mehr vollen Rang hat. Dies kann zu Problemen mit dem Schätzalgorithmus führen, so dass man in dieser Situation entweder gar keine oder nur unbrauchbare Ergebnisse erhält.

Seien x_1 und x_2 kategoriale Kovariablen mit K bzw. M Ausprägungen $\{1, \ldots, K\}$ bzw. $\{1, \ldots, M\}$. Die Gesamtanzahl verwendeter Dummies für x_1 und x_2 setzt sich zusammen aus $K-1$ Haupteffektdummies für x_1, $M-1$ Haupteffektdummies für x_2 sowie $(K-1)\cdot(M-1) = K\cdot M - K - M + 1$ Interaktionsdummies von x_1 und x_2 und beträgt somit $K\cdot M - 1$. Im Falle einer vollbesetzen Kontingenztabelle, also bei $K\cdot M$ besetzten Zellen, ist $K\cdot M - 1$ gerade die größtmögliche Anzahl an Dummy-Variablen, die man verwenden kann, ohne dass lineare Abhängigkeiten im Datensatz entstehen. Im Falle einer nicht vollbesetzten Kontingenztabelle mit l leeren Zellen sind aber höchstens $K\cdot M - l - 1$ Dummies linear unabhängig. Da aufgrund unserer oben getroffenen Voraussetzung die den Interaktionen zugrunde liegenden Haupteffekte im Modell enthalten sein müssen, können zusätzlich nur noch $(K-1)\cdot(M-1)-l$ Interaktionsdummies in die Modellierung einfließen. Die Wahl der verwendeten Dummies ist zudem nicht beliebig, sondern muss stets so erfolgen, dass die ausgewählten Interakionsdummies einer Zeile bzw. einer Spalte linear unabhängig von der Dummyvariable des zugehörigen Haupteffektes sind.

Im betrachteten Datensatz tritt dieses Problem unter anderem zwischen den Kovariablen `hochwert`, die anzeigt, ob ein Kunde als hochwertig betrachtet wird oder nicht, und `anzvhv` auf, die angibt, wieviele Einschlüsse in der verbundenen Hausratsversicherung des Kunden existieren. Kunden mit 4 oder mehr Einschlüssen gelten hierbei automatisch als hochwertig (`hochwert`=1). Tabelle 7.1 zeigt die relativen Häufigkeiten der verschiedenen Ausprägungskombinationen der beiden Kovariablen.

`anzvhv`:	kein HR	0	1-2	3-4	≥5	total:
`hochwert`:			*(Angaben in %)*			
normal	35.75	5.30	6.58	4.28	–	51.90
hochwertig	21.09	0.81	3.09	10.92	12.19	48.10
total:	56.84	6.11	9.67	15.20	12.19	100.00

Tabelle 7.1: *Relative Häufigkeiten zu* `anzvhv` und `hochwert`

Verwendet man die vier Kategorien 0, 1-2, 3-4 und ≥5 als Haupteffektdummies für `anzvhv` und „hochwertig“ als Dummy für `hochwert`, so können nur noch drei Interaktionsdummies im Datensatz verbleiben. Es ist leicht zu sehen, dass die Dummyvariable zur Interaktion (hochwertig, ≥5) auf keinen Fall ausgewählt werden darf, da sie mit der Haupteffektdummy zu `anzvhv` ≥5 vollkommen identisch ist. In diesem Fall gibt es also nur die Möglichkeit, die Interaktionsdummies zu (hochwertig, 0),(hochwertig, 1-2) und (hochwertig, 3-4) im Datensatz zu belassen.

Eine besonders ungünstige Datenkonstellation ergibt sich, wenn zwei Kovariablen hinsichtlich einer Kategorie völlig übereinstimmen. Dann sind zunächst die beiden entsprechenden Haupteffektdummies identisch und damit kollinear. Falls eine weitere Kovariable x_3 mit zwei Kovariablen x_1 und x_2 interagiert, bei denen die Dummies zu $(x_1 = I)$ und $(x_2 = J)$ übereinstimmen, so setzt sich das Problem linearer Abhängigkeiten auch bei diesen Interaktionen fort: Die Interaktionsdummies zu $(x_1 = I, x_3 = k)$ und $(x_2 = J, x_3 = k)$ sind für alle Ausprägungen k von x_3 ebenfalls identisch, so dass nicht beide gleichzeitig im Datensatz verbleiben können. Bei einer großen Zahl von Interaktionen mit zwei sich auf diese Art überschneidenden Kovariablen wird die Kollinearitätsstruktur der Dummies schnell unübersichtlich und die Bereinigung des Datensatzes nicht mehr ganz einfach.

Eine Möglichkeit zur Umgehung dieses Problems besteht darin, die zusammenhängenden Variablen zu einer einzigen zu kombinieren. Dies bietet sich jedoch nur dann an, wenn die beiden Einflussgrößen inhaltlich eng verwandt sind und deshalb sinnvoll zusammengefasst werden können. Im vorliegenden Datensatz ist dies z.B. bei der kategorialen Version von `lastd` und `ldspart` der Fall, da beide den letzten Neuabschluss im Zeitraum von 1999 bis 2001 betrachten. Dabei bezeichnet `lastd` die Zeitspanne seit dem letzten Neuabschluss des Kunden und `ldspart` die Sparte dieses letzten Neuabschlusses. Eine völlige Überschneidung liegt bei der Kategorie „kein Neuabschluss von 99 bis 01“ vor, wie die Tabelle der relativen Häufigkeiten zeigt:

`ldspart`:	Kraft	Sach	Leben	kein Neuabschluss	total:
`lastd`:			*(Angaben in %)*		
≤1 Jahr	2.99	7.48	2.34	–	12.81
1-2 Jahre	2.70	6.54	1.06	–	10.30
2-3 Jahre	2.12	5.89	1.97	–	9.98
kein Neuab.	–	–	–	66.91	66.91
total:	7.89	19.92	5.37	66.91	100.00

Tabelle 7.2: *Relative Häufigkeiten zu* `ldspart` und `lastd`

Da viele Kovariablen sowohl mit `lastd` als auch mit `ldspart` interagieren, ist eine Zusammenfassung der beiden zu einer neuen Kovariablen äußerst ratsam. Dabei definiert man die Interaktionsdummies der beiden zu Haupteffektdummies der neuen Kovariablen um:

Für die Modellierung von `storno` definiert man die Kovariable `neukombstorno` über

`neukombstorno`	`ldspart`	`lastd`
0	kein Neuab.	kein Neuab.
1	Kraft	$\leq$1 Jahr
2	Kraft	1-2 Jahre
3	Kraft	2-3 Jahre
4	Leben	$\leq$1 Jahr
5	Leben	1-2 Jahre
6	Leben	2-3 Jahre
7	Sach	$\leq$1 Jahr
8	Sach	1-2 Jahre
9	Sach	2-3 Jahre

Dasselbe Problem tritt auch bei der für `abschluss` benutzten Kodierung von `ldspart` auf, bei der nur in „kein Neuabschluss“ und „Neuabschluss“ unterschieden wird. Analog zu `neukombstorno` wird `neukombabschluss` eingeführt mit

`neukombstorno`	`ldspart`	`lastd`
0	kein Neuab.	kein Neuab.
1	Neuab.	$\leq$1 Jahr
2	Neuab.	1-2 Jahre
3	Neuab.	2-3 Jahre

Die in dieser Arbeit verwendete Software STATA® erkennt zwar selbständig alle Kollinearitäten und entfernt automatisch genügend Dummies, um lineare Unabhängigkeit und damit eine Designmatrix vollen Ranges zu erzeugen. Wenn man jedoch die Entscheidung darüber, welche Dummies im Datensatz verbleiben und welche entfernt werden sollen, nicht einfach dem Programm überlassen möchte, ist eine eigene Analyse der Kontingenztabellen unerlässlich. Mithilfe der beiden hier beschriebenen Methoden kann man dann alle auftretenden linearen Abhängigkeiten im Datensatz selbst beseitigen.

7.4 Vorhersagbarkeit des Response in Kontingenztabellen

In Zusammenhang mit Kontingenztabellen sei auf ein weiteres mögliches Problem in der Datenkonstellation hingewiesen: Bei eher schwach besetzten Zellen einer Kontingenztabelle kann es auch vorkommen, dass nahezu alle Beobachtungen in einer solchen Zelle (i, j) denselben Response $Y = 0$ aufweisen, so dass dieser scheinbar perfekt vorhersagbar ist. Da die Schätzung der Wahrscheinlichkeit $P(Y = 0)$ dort nahe 1 und $P(Y = 1)$ fast 0

sein muss, müsste der geschätzte Logit

$$\widehat{\text{Logit}}\,(i,j) = \log\left(\frac{\hat{P}(Y=1|x_1=i,x_2=j)}{\hat{P}(Y=0|x_1=i,x_2=j)}\right)$$

den Wert $-\infty$ annehmen, was zu numerischen Problemen im Schätzalgorithmus führt. Falls fast alle Beobachtungen die Ausprägung $Y=1$ aufweisen, ist eine Schätzung ebenfalls schwierig, da der Logit dann gegen $+\infty$ gehen müsste. In beiden Fällen geht insbesondere auch die Standardabweichung gegen $+\infty$. In solch einer Situation entfernt STATA®die zugehörigen Beobachtungen aus dem Datensatz, da die Wahrscheinlichkeiten dort nicht mehr geschätzt zu werden brauchen. (vgl. STATA Base Reference Manual (2003)). Da es bei einer großen Anzahl von Kovariablen aber häufig vorkommt, dass Dummyvariablen nach dem Entfernen dieser Beobachtungen linear abhängig werden, kann dies wiederum zu weiteren Problemen führen.

8 Modellierung stetiger Kovariablen

Bei der bisherigen Vorgehensbeschreibung in Kapitel 7 wurde stets davon ausgegangen, dass alle Kovariablen in kategorialer Form vorliegen. Oftmals enthält ein Datensatz – auch der uns vorliegende – aber eine Mischung aus stetigen und kategorialen Kovariablen. Im Folgenden wollen wir erläutern, auf welche Arten man diese stetigen Einflussgrößen modellieren kann. Verschiedene Möglichkeiten zur Bewertung und zum Vergleich dieser unterschiedlichen Modelle schließen das Kapitel ab.

8.1 Modellierung in kategorialer Form

Für stetige Kovariablen gibt es grundsätzlich zwei Möglichkeiten der Modellierung. Zum einen kann man stets die stetige Kovariable x durch eine Unterteilung des Wertebereichs in Intervalle zu einer kategorialen Kovariablen $\tilde{x}$ umwandeln. Sei $\xi_0, \ldots, \xi_M$ eine Partition des Wertebereichs $[a, b]$ von x mit

$$a = \xi_0 < \xi_1 < \cdots < \xi_{M-1} < \xi_M = b.$$

Dann hat $\tilde{x}$ die Ausprägung $i \in \{1, \ldots, M\}$ genau dann, wenn $x \in [\xi_{i-1}, \xi_i)$ gilt.

Mit der so erzeugten kategorialen Kovariablen $\tilde{x}$ kann man im Weiteren genauso verfahren wie mit den anderen kategorialen Kovariablen. So lassen sich die Haupt- und Interaktionseffekte wie in Abschnitt 7.1 und 7.2 analysieren. Modelliert wird $\tilde{x}$ dann beispielsweise in Dummy-Kodierung mit Kategorie M als Referenzgruppe, so dass $\tilde{x}$ in der Form

$$\beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \cdots + \beta_{M-1} \tilde{x}_{M-1} \tag{8.1}$$

in den linearen Prädiktor $\boldsymbol{\beta}'\boldsymbol{x}$ des Modells eingeht, wobei $\tilde{x}_1, \ldots, \tilde{x}_{M-1}$ wieder als Dummyvariablen mit

$$\tilde{x}_i = \begin{cases} 1 & \text{falls } \tilde{x} = i \\ 0 & \text{sonst} \end{cases}$$

für $i = 1, \ldots, M$ definiert werden.

Für eine erste Beurteilung des Einflusses einer stetigen Kovariablen bietet es sich in jedem Fall an, zunächst solch eine Kategorisierung vorzunehmen. Allerdings hängt die Analyse der empirischen Logits dabei wesentlich von der Anzahl der Kategorien und der Wahl der Punkte $\xi_0, \ldots, \xi_M$ ab und spiegelt den wahren Effekt einer stetigen Kovariablen nur annäherungsweise wider. Je gröber die Unterteilung des Wertebereichs ist, desto ungenauer wird diese kategoriale Analyse.

Ein Nachteil der Darstellung (8.1) liegt in der großen Zahl von Parametern, die zur Modellierung benötigt werden. Deshalb versucht man häufig, nach Möglichkeit die stetige Form der Kovariablen in der Modellierung beizubehalten.

8.2 Modellierung in stetiger Form

Das Ziel stetiger Modellierung ist es, für den Effekt einer stetigen Kovariablen x eine Darstellung der Form

$$\beta_x \cdot f(x)$$

zu finden, wobei $f(x)$ eine stetige Transformation von x bezeichnet. Diese Darstellung kommt mit nur einem Parameter β_x aus und garantiert damit ein kleineres Modell und weniger Rechenaufwand bei der Parameterschätzung. Bevorzugt werden dabei möglichst einfache Transformationen, um eine gute Interpretierbarkeit des Effektes von x zu gewährleisten.

Um eine geeignete Transformation f für die stetige Kovariable x zu finden, ist eine eingehende Untersuchung ihres Effektes auf den Response nötig. Eine Kategorisierung von x und anschließende Untersuchung der empirischen Logits liefert nur ein grobes Bild und ist somit meist unzureichend. Für den in dieser Arbeit betrachteten Datensatz erstellten wir deshalb zur graphischen Analyse des Haupteffektes stetiger Kovariablen mithilfe des Programms *BayesX* (Brezger *et al.* (2003)) sogenannte P-Splines mit Random Walk-Bestrafung zweiter Ordnung, die z.B. in Lang & Brezger (2004) oder Brezger & Lang (2003) beschrieben werden.

Zeigt sich in der graphischen Auswertung ein linearer Einfluss der Kovariable x, so ist keine Transformation nötig und man kann x mit $f(x) = x$ direkt ins Modell einfließenen lassen. Zur Modellierung eines quadratischen Einflusses kann man $f(x) = x^2$ wählen. Eine Möglichkeit zur Modellierung weiterer nichtlinearer Effekte sind beispielsweise Box-Cox-Transformationen der Form $f(x) = \frac{x^k-1}{k}$ für ein $k \in \mathbb{N}$. Auch $f(x) = \log x$ und $f(x) = \exp(x)$ werden häufig verwendet. Manchmal kann es zudem sinnvoll sein, x in standardisierter Form zu transformieren. In diesem Fall verwendet man $f(\frac{x-\bar{x}}{\hat{\sigma}})$ anstatt von $f(x)$ zur Modellierung, wobei $\bar{x}$ den empirischen Mittelwert und $\hat{\sigma}^2$ die empirische Varianz von x bezeichnet.

Falls jedoch mithilfe dieser einfachen Transformationen keine adäquate Modellierung möglich ist, so muss man komplexere Funktionen heranziehen. So käme z.B. eine Modellierung als Polynom l-ten Grades in Betracht. Da Ausreisser hierbei aber einen sehr großen Einfluss auf den Kurvenverlauf haben können, und sich bei einem hohen Grad des Polynoms teilweise stark schwankende Kurven ergeben (Fahrmeir & Tutz (1994)), ist es vorteilhaft, anstatt eines Polynoms einen sogenannten *polynomialen Spline* zur Modellierung zu verwenden. Diese sind folgendermaßen definiert:

Definition 8.1. *Sei $a = \xi_1 < \xi_2 < \cdots < \xi_{m-1} < \xi_m = b$ eine Unterteilung des Intervalls $[a, b]$. Eine Funktion $s : [a, b] \to \mathbb{R}$ heißt Polynomspline vom Grad l, wenn gilt*

(i) $s(x)$ ist ein Polynom vom Grad l für $x \in [\xi_j, \xi_j + 1)$ für $j = 1, \ldots, m-1$,

(ii) s ist $(l-1)$ mal stetig differenzierbar.

Für jedes Teilintervall $[\xi_j, \xi_j + 1)$ lässt sich somit ein eigenes Polynom konstruieren, wodurch die Modellierung flexibler und die Anpassung erleichtert wird. In Forderung (ii) wird verlangt, dass die Übergänge zwischen den einzelnen Intervallen möglichst glatt sein sollen. Dies garantiert, dass die resultierende Gesamtfunktion ebenfalls glatt ist. Splines vom Grad 0 sind nach Definition 8.1 stückweise konstante Funktionen, während Splines vom Grad 1 Polygonzüge durch die Punkte $(\xi_j, s(\xi_j))$ beschreiben. Diese sind zwar stetig, aber an den Intervallgrenzen nicht differenzierbar. Ab Grad 2 sind Polynomsplines dann mindestens einmal stetig differenzierbar.

Man kann zeigen, dass der als $S_l(K_m)$ bezeichnete Raum der Polynomsplines vom Grad l zur Knotenmenge $K_m := \{\xi_1, \ldots, \xi_m\}$ ein Vektorraum der Dimension $l + m - 1$ ist (siehe z.B. Hämmerlin & Hoffman (1990)) . Zudem stellt $S_l(K_m)$ einen Unterraum des Vektorraums aller $(l-1)$ mal stetig differenzierbaren Funktionen auf dem Intervall $[a, b]$ dar. Folglich muss es $l + m - 1$ Basisfunktionen B_i und Koeffizienten γ_i geben, so dass jeder Spline $s \in S_l(K_m)$ eine (eindeutige) Darstellung folgender Form besitzt:

$$s(x) = \gamma_0 B_0(x) + \gamma_1 B_1(x) + \cdots + \gamma_{l+m-2} B_{l+m-2}(x) \ . \tag{8.2}$$

Es gibt verschiedene Basen, die in (8.2) verwendet werden können. Am häufigsten werden die sogenannte „Truncated Power Series“-Basis oder die B-Spline Basis benutzt, die beide von Fahrmeir & Tutz (1994) beschrieben werden. Da wir uns in unserem Datensatz auf Truncated Power Series beschränkt haben, werden wir nur diese im Folgenden kurz erläutern.

Die Basisfunktionen der Truncated Power Series Basis zum Vektorraum $S_l(K_m)$ sind gegeben durch

$$\begin{aligned}
B_0(x) &= 1 \\
B_1(x) &= x \\
B_2(x) &= x^2 \\
&\vdots \\
B_l(z) &= x^l \\
B_{l+1}(x) &= (x - \xi_2)_+^l \\
B_{l+2}(x) &= (x - \xi_3)_+^l \\
&\vdots \\
B_{l+m-2}(x) &= (x - \xi_{m-1})_+^l \ ,
\end{aligned}$$

wobei $(x-\xi_j)_+^l$ für $j = 2, \ldots, m-1$ definiert ist als

$$(x-\xi_j)_+^l = \begin{cases} (x-\xi_j)^l & \text{falls } x \geq \xi_j \\ 0 & \text{sonst.} \end{cases}$$

Die Truncated Power Series Basis besteht also aus Potenzen bzw. abgeschnittenen Potenzen. Jeder Spline s aus $S_l(K_m)$ besitzt damit eine Darstellung der Form

$$s(x) = \sum_{i=0}^{l} \gamma_i \, x^i + \sum_{j=2}^{m-1} \gamma_{i+j-1} (x-\xi_j)^l. \tag{8.3}$$

Abbildung 8.1 zeigt die Basisfunktionen B_0 bis B_5 für einen Polynomspline vom Grad 2 über dem Intervall $[0,1]$ zur Knotenmenge $\{0, 0.25, 0.5, 0.75, 1\}$.

Die Wahl der Knoten stellt einen eigenen Problemkreis dar, da Anzahl und Lage der Knoten die Modellierung maßgeblich beeinflussen. Möglich ist beispielsweise die Verwendung äquidistanter Knoten. Solchermaßen erzeugte Splines nennt man Kardinalsplines. Auch Quantile der Kovariablen können als Knotenpunkte herangezogen werden. Wir verwendeten für die uns vorliegenden Daten die graphischen Auswertungen von *BayesX*, um geeignete Knoten zu bestimmen. Häufig wählten wir z.B. Minima, Maxima oder Wendepunkte des jeweiligen P-Splines. Eine umfassende Diskussion der Knotenwahl findet sich z.B. in Friedman & Silverman (1989).

Möchte man den Effekt einer stetigen Kovariablen x in Form eines Polynomsplines ins Modell aufnehmen, so muss man (8.3) zufolge $l+m-1$ Parameter schätzen, so dass sich der Rechenaufwand bei stetiger im Vergleich zu kategorialer Modellierung nicht unbedingt reduzieren muss. Zudem ist ein Polynomspline oftmals schwieriger zu interpretieren als der Effekt der einzelnen Gruppen, die bei Kategorisierung einer stetigen Kovariablen entstehen.

Die Frage, welche Art der Modellierung im konkreten Datensatz vorteilhafter ist, ist nicht leicht zu entscheiden. Falls die Daten den gesamtem Wertebereich von x relativ gleichmäßig abdecken, wie es in unserem Datensatz bei der Altersangabe `pobage` der Fall ist, so kann eine stetige Darstellung eine bessere Modellanpassung bewirken. Konzentrieren sich andererseits die Daten hauptsächlich in einem Teilintervall des Wertebereichs oder liegen sogar in manchen Teilbereichen gar keine Daten vor, so kann eine Kategorisierung, die diesen Umständen Rechnung trägt, sinnvoller sein. Im vorliegenden Datensatz ist dies z.B. bei der Prämienangabe `jbtgad` der Fall.

Die Entscheidung zwischen kategorialer und stetiger Modellierung der Kovariablen wird durch ein geeignetes Kriterium zur Bewertung der jeweiligen Modellanpassung erleichtert. Zwei Möglichkeiten zum Vergleich dieser Modelle werden im Folgenden erläutert.

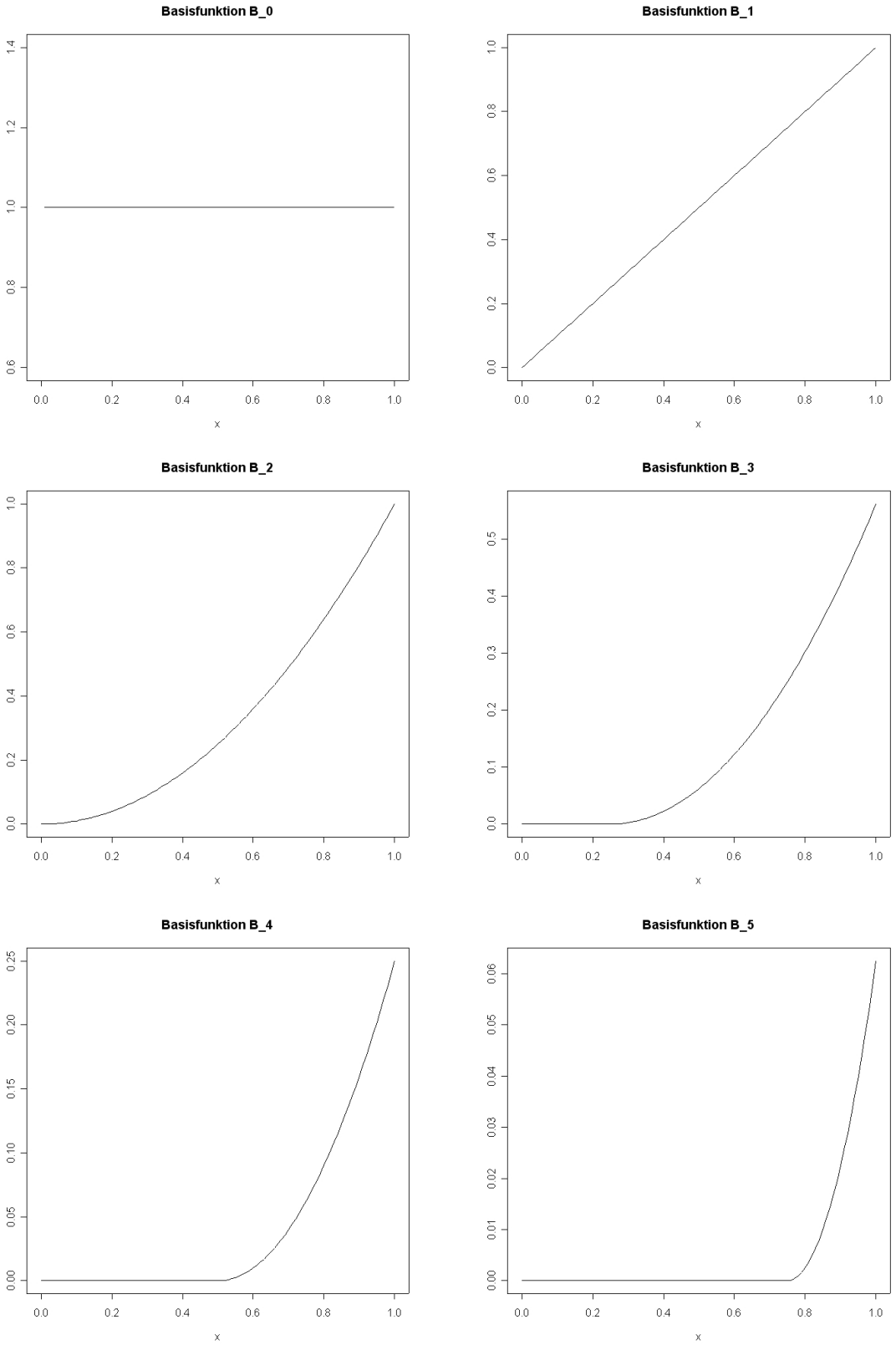


Abbildung 8.1: *Basisfunktionen* B_0 *bis* B_5 *für einen Polynomspline vom Grad 2 über dem Intervall* $[0, 1]$ *zur Knotenmenge* $\{0, 0.25, 0.5, 0.75, 1\}$

8.3 Vergleich von Modellen mit stetigen bzw. kategorialen Kovariablen

Während man bei den Modellen, die nur kategoriale Kovariablen enthalten, die Beobachtungen nach identischen Kovariablenvektoren gruppieren kann, ist dies bei Modellen mit stetigen Kovariablen aufgrund der hohen Anzahl verschiedener Ausprägungen im Allgemeinen nicht mehr möglich. Deshalb liegen den Modellen mit stetigen Kovariablen keine binomialverteilten Daten mehr zugrunde, sondern nur noch binäre Daten, also die Bernoulli-verteilten Einzelbeobachtungen, für die die Devianz als Anpassungsmaß keine wirkliche Aussagekraft hat (siehe z.B. Collett (1999)). Somit kommt bei Modellen, die Kovariablen in stetiger Form enthalten, die Devianz zur Modellbewertung nicht in Betracht. Zudem sind stetige und kategoriale Modelle nicht genestet, da keine echte Modellerweiterung vorliegt, sondern lediglich die Dummyvariablen der einzelnen Kategorien der Kovariablen aus dem Modell genommen und durch stetige Komponenten ersetzt werden. Eine Entscheidung zwischen zwei Modellen mittels eines Tests wie z.B. dem Partial Deviance Test ist damit nicht möglich.

Für nichtgenestete Modelle benötigt man somit andere Kriterien zum Modellvergleich. Hierzu zogen wir zunächst drei Möglichkeiten in Betracht:

- das von Akaike (1973) vorgestellte Informations-Kriterium (AIC),
- die Berechnung von Bayes-Faktoren (Kass & Raftery (1995) und Raftery (1996))
- das Bayes-Informationskriterium (BIC), eine abgewandelte Form des von Schwarz (1978) entwickelten Kriteriums.

Aufgrund des erhöhten Rechenaufwands für Bayes-Faktoren beschränken wir uns auf das AIC und das BIC, welches eine Approximation erster Ordnung an die Bayes-Faktoren darstellt. Die beiden Kriterien werden folgendermaßen berechnet:

$$AIC = -2 \cdot \text{Log-Likelihood} + 2 \cdot \text{df}$$
$$BIC = -2 \cdot \text{Log-Likelihood} + \log(n) \cdot \text{df}$$

Beide Kriterien ziehen zum Modellvergleich die Log-Likelihood heran, die um einen Term zur „Bestrafung“ zu großer Modelle ergänzt wird. In diesen Strafterm geht die Anzahl der Freiheitsgrade des Modells (df) ein, die mit einem zusätzlichen Faktor gewichtet werden. AIC und BIC unterscheiden sich nur hinsichtlich dieses Faktors: AIC benützt den Faktor 2 zur Gewichtung, während das BIC-Kriterium mit $\log(n)$ die logarithmierte Anzahl der zugrunde liegenden Beobachtungen verwendet. Das Modell mit dem kleinsten AIC- bzw. BIC-Wert wird dabei jeweils als das beste Modell angesehen. Da die Bestrafung beim BIC stärker ist, werden damit tendenziell kleinere Modelle als beim AIC ausgewählt. Die Vor- und Nachteile von AIC bzw. BIC diskutieren z.B. Shibata (1976), Katz (1981) und Findley (1991).

9 Zusammenfassung und Ausblick

Zum Abschluss dieser Arbeit soll noch einmal deutlich hervorgehoben werden, dass eine gemeinsame Modellierung binärer Zielvariablen in der Regel vorteilhafter ist als die Erstellung separater Modelle, da hierbei die den Zielgrößen zugrunde liegende Korrelationsstruktur zusätzlich berücksichtigt und explizit modelliert wird. Zwar sind im Allgemeinen bei relativ kleinen Korrelationen die marginalen Effekte im bivariaten Probit-Modell mit denen der marginalen Modelle nahezu identisch. Auch bei dem von uns untersuchten Datensatz, bei dem die geschätzte Korrelation sehr nahe bei Null liegt, unterscheiden sich die geschätzten Koeffizienten im bivariaten Modell kaum von denen der beiden univariaten Modelle. Das gemeinsame Modell bildet aber aufgrund des zusätzlich geschätzten Korrelationsparameters das Verhalten der Versicherungskunden tatsächlich genauer ab und erlaubt präzisere Aussagen über die Motivation von Kunden, Verträge zu stornieren und gleichzeitig Neuabschlüsse zu tätigen. Die gesonderte Untersuchung von einzelnen Kundengruppen vervollständigt und rundet dieses Bild noch ab. Mit diesen Ergebnissen könnten künftig für das Versicherungsunternehmen besonders interessante Zielgruppen identifiziert und Marketing-Kampagnen speziell auf diese Kunden zugeschnitten werden.

Der theoretische Hintergrund und die zugrunde liegenden Konzepte der univariaten und bivariaten Probit-Modelle wurden ausführlich erläutert. Wir haben gesehen, dass unter bestimmten Umständen kein Maximum-Likelihood-Schätzer existieren muss. Tatsächlich sind wir in der Praxis bei der Betrachtung einiger Untergruppen auf solche Fälle gestoßen. Auch beim Versuch, manche Kundengruppen noch feiner zu zergliedern, trat diese Situation häufig auf. Durch eine nähere Untersuchung der jeweiligen Datenkonstellation könnte man hier vielleicht tieferen Einblick in das vorliegende Datenmaterial gewinnen und und ein besseres Verständnis für die darin herrschenden Zusammenhänge entwickeln.

Bei unserem Datensatz war es nötig, für jede einzelne untersuchte Untergruppe ein eigenes vollständiges bivariates Probit-Modell aufzustellen, um Aussagen über die Korrelation innerhalb dieser Gruppe machen zu können. Da die Korrelationen sich in einem sehr engen Intervall bewegen, ist eine Abschätzung über Pearsons tetrachorische Korrelation in dieser Situation nicht ausreichend. Wenn dagegen Datenmaterial vorliegt, bei dem die Unterschiede zwischen den Korrelationen einzelner Subpopulationen ausgeprägter sind, so kann dieser relativ einfache Schätzer durchaus wertvolle Ergebnisse liefern. Dieser Ansatz sollte also nicht von vornherein verworfen werden, da damit die aufwändige Schätzung des vollen bivariaten Modells umgangen werden kann.

Eine interessante Möglichkeit zur Interpretation und Bewertung von verschiedenen Werten des Korrelationsparameters bietet die Betrachtung der geschätzten Korrelationen auf der Skala der Odds Ratios. Die Auslegung des Odds Ratios als Vergleich von Chancen innerhalb zweier Untergruppen des Datensatzes ist weitaus greifbarer und stellt einen unmittelbaren und klaren Bezug zum gegebenen Datenmaterial her. Deutlich tritt dabei zutage, welche Auswirkung eine Änderung der Korrelation auf das vorherrschende Chancengefüge innerhalb einer Population hat. Wie sich gezeigt hat, können selbst kleine Korrelationsänderungen den Odds Ratio massiv beeinflussen.

Es sei noch darauf hingewiesen, dass die Modellbewertung mittels AIC und BIC nicht klar darüber Aufschluss geben konnte, ob eine stetige Modellierung gewisser Kovariablen sinnvoll ist oder nicht. Ein Vergleich der Modelle mithilfe der Bayes-Faktoren wäre zwar relativ aufwändig, könnte hier aber Klarheit darüber verschaffen, welche Art der Modellierung tatsächlich vorzuziehen ist.

A Eigenschaften der bivariaten Standardnormalverteilung

In diesem Abschnitt geben wir einen kurzen Überblick über die Eigenschaften der bivariaten Standardnormalverteilung mit Korrelationsparameter ϱ, die wir als

$$N_2\left(\mathbf{0} := \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma := \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right)$$

bezeichnen. Dabei ist der Vektor $(0,0)'$ der Erwartungsvektor und Σ die Kovarianzmatrix der bivariaten Standardnormalverteilung. Im Folgenden gelte also für den Zufallsvektor $(W_1, W_2)'$

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N_2(\mathbf{0}, \Sigma).$$

Die bivariate Standardnormalverteilung ist ein Spezialfall der multivariaten Normalverteilung, die von Tong (1990) umfassend behandelt wird. Auch in Johnson & Kotz (1972) und Anderson (2003) finden sich einige wichtige Eigenschaften der multivariaten Normalverteilung.

A.1 Die Dichtefunktion von $N_2(\mathbf{0}, \Sigma)$

Die Dichte $\phi_2(w_1, w_2, \varrho)$ der bivariaten Standardnormalverteilung mit Korrelationsparameter ϱ ist gegeben durch

$$\begin{aligned}
\phi_2(w_1, w_2, \varrho) &= \frac{1}{2\pi\sqrt{(1-\varrho^2)}} \exp\left\{-\frac{1}{2}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}' \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right\} \\
&= \frac{1}{2\pi\,|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\varrho^2)}} \exp\left\{-\frac{({w_1}^2 - 2\varrho w_1 w_2 + {w_2}^2)}{2(1-\varrho^2)}\right\}, \qquad \text{(A.1)}
\end{aligned}$$

wobei $|\Sigma|$ die Determinante der Kovarianzmatrix Σ bezeichnet.

Mit der Dichte

$$\phi(w) = -\frac{1}{2\pi}\exp\left\{-\frac{w^2}{2}\right\}$$

der univariaten Standardnormalverteilung lässt sich $\phi_2(w_1, w_2, \varrho)$ alternativ auch schreiben als

$$\phi_2(w_1, w_2, \varrho) = \frac{1}{\sqrt{1-\varrho^2}}\, \phi(w_1) \cdot \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right), \tag{A.2}$$

da

$$\begin{aligned}
&\frac{1}{\sqrt{1-\varrho^2}}\, \phi(w_1) \cdot \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right) = \\
=&\frac{1}{\sqrt{1-\varrho^2}} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{w_1^2}{2}\right\} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right)^2\right\} \\
=& \frac{1}{2\pi\sqrt{1-\varrho^2}}\, \exp\left\{-\frac{[w_1^2(1-\varrho^2) - w_2^2 + 2\varrho w_1 w_2 - \varrho^2 w_1^2]}{2(1-\varrho^2)}\right\} \\
=& \frac{1}{2\pi\sqrt{1-\varrho^2}}\, \exp\left\{-\frac{(w_1^2 + w_2^2 - 2\varrho w_1 w_2)}{2(1-\varrho^2)}\right\} = \phi_2(w_1, w_2, \varrho).
\end{aligned}$$

Analog zeigt man, dass auch

$$\phi_2(w_1, w_2, \varrho) = \frac{1}{\sqrt{1-\varrho^2}}\, \phi(w_2) \cdot \phi\left(\frac{w_1 - \varrho w_2}{\sqrt{1-\varrho^2}}\right) \tag{A.3}$$

gilt. Eigenschaft (A.2) kann man umformulieren zu

$$\frac{1}{\sqrt{1-\varrho^2}}\, \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right) = \frac{\phi_2(w_1, w_2, \varrho)}{\phi(w_1)}\ . \tag{A.4}$$

Da $\phi(w_1)$ die marginale Dichte von W_1 ist (Beweis siehe Abschnitt A.3), stellt die rechte Seite den Quotienten aus gemeinsamer Dichtefunktion von $(W_1, W_2)'$ und der marginalen Dichtefunktion von W_1 dar. Damit ist

$$f_{W_2|W_1}(w_1, w_2) := \frac{1}{\sqrt{1-\varrho^2}}\, \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right)$$

also die bedingte Dichte von W_2 gegeben $W_1 = w_1$. Aus Eigenschaft (A.3) ergibt sich analog die bedingte Dichte von $W_1|W_2 = w_2$ als

$$f_{W_1|W_2}(w_1, w_2) := \frac{1}{\sqrt{1-\varrho^2}}\, \phi\left(\frac{w_1 - \varrho w_2}{\sqrt{1-\varrho^2}}\right).$$

A.2 Die Verteilungsfunktion von $N_2(\mathbf{0}, \Sigma)$

Die Verteilungsfunktion $\Phi_2(z_1, z_2, \varrho)$ der bivariaten Standardnormalverteilung mit Korrelationsparameter ϱ an der Stelle (z_1, z_2) berechnet sich aus (A.1) als

$$\Phi_2(z_1, z_2, \varrho) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \phi_2(w_1, w_2, \varrho) dw_2 \, dw_1$$

$$= \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \frac{1}{2\pi \, |\Sigma|^{\frac{1}{2}}} \, \exp\left\{ -\frac{1}{2} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right\} dw_2 \, dw_1. \tag{A.5}$$

Benutzt man für $\phi_2(w_1, w_2, \varrho)$ die Darstellung (A.2), so ergibt sich für die Verteilungsfunktion

$$\Phi_2(z_1, z_2, \varrho) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \phi_2(w_1, w_2, \varrho) \, dw_2 dw_1$$

$$= \int_{-\infty}^{z_1} \phi(w_1) \left[\int_{-\infty}^{z_2} \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}} \right) \cdot \frac{1}{\sqrt{1-\varrho^2}} \, dw_2 \right] dw_1.$$

Mittels Subsitution berechnet man zunächst für das innere Integral

$$\int_{-\infty}^{z_2} \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}} \right) \cdot \frac{1}{\sqrt{1-\varrho^2}} \, dw_2 \overset{u_2 := \frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}}{=}$$

$$\int_{-\infty}^{\frac{z_2 - \varrho w_1}{\sqrt{1-\varrho^2}}} \phi(u_2) \cdot \frac{1}{\sqrt{1-\varrho^2}} \, \sqrt{1-\varrho^2} \, du_2 \; = \; \Phi\left(\frac{z_2 - \varrho w_1}{\sqrt{1-\varrho^2}} \right), \tag{A.6}$$

da $\quad dw_2 = \sqrt{1-\varrho^2} \, du_2.$

Somit ergibt sich

$$\Phi_2(z_1, z_2, \varrho) = \int_{-\infty}^{z_1} \phi(w_1) \, \Phi\left(\frac{z_2 - \varrho w_1}{\sqrt{1-\varrho^2}} \right) \, dw_1. \tag{A.7}$$

Analog zeigt man

$$\Phi_2(z_1, z_2, \varrho) = \int_{-\infty}^{z_2} \phi(w_2) \, \Phi\left(\frac{z_1 - \varrho w_2}{\sqrt{1-\varrho^2}} \right) \, dw_2. \tag{A.8}$$

A.3 Die Marginaldichten von $N_2(\mathbf{0}, \Sigma)$

Insbesondere folgt aus (A.2)

$$\begin{aligned}\int_{-\infty}^{\infty} \phi_2(w_1, w_2, \varrho)\, dw_2 &= \phi(w_1) \int_{-\infty}^{\infty} \phi\left(\frac{w_2 - \varrho w_1}{\sqrt{1-\varrho^2}}\right) \cdot \frac{1}{\sqrt{1-\varrho^2}}\, dw_2 \\ &\overset{(A.6)}{=} \phi(w_1)\, \Phi(\infty) = \phi(w_1) \cdot 1 = \phi(w_1). \end{aligned} \tag{A.9}$$

Die marginale Dichte von W_1 ist demnach also $\phi(w_1)$. Analog ergibt sich aus der Darstellung (A.3) $\phi(w_2)$ als marginale Dichte von W_2. Es gilt also:

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N_2(\mathbf{0}, \Sigma) \Longrightarrow W_i \sim N(0,1), \quad i = 1, 2. \tag{A.10}$$

A.4 Weitere Eigenschaften von $N_2(\mathbf{0}, \Sigma)$

Allgemein gilt für multivariat normalverteilte Zufallsvektoren $W := (W_1, \ldots, W_n)'$ mit $W \sim N_n(\boldsymbol{\mu}, \Omega)$ mit Erwartungsvektor $\mu \in \mathbb{R}^n$ und Kovarianzmatrix $\Omega \in \mathbb{R}^n$ sowie eine beliebige Transformationsmatrix $T \in \mathbb{R}^{k \times n}$

$$TW \sim N_k(T\mu\,,\, T\Omega T').$$

Lineare Transformationen von normalverteilten Zufallsvektoren sind also ebenfalls wieder normalverteilt mit Erwartungsvektor $T\mu \in \mathbb{R}^k$ und Kovarianzmatrix $T\Omega T' \in \mathbb{R}^{k \times k}$. Im Spezialfall

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right)$$

kann man durch die Anwendung der Transformationen $T_1 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, T_2 := \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ bzw. mit $T_3 := \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ zeigen, dass

$$\begin{aligned} \begin{pmatrix} W_1 \\ -W_2 \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\varrho \\ -\varrho & 1 \end{pmatrix}\right), \\ \begin{pmatrix} -W_1 \\ W_2 \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\varrho \\ -\varrho & 1 \end{pmatrix}\right) \\ \begin{pmatrix} -W_1 \\ -W_2 \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right). \end{aligned}$$

Damit gilt

$$P(W_1 \leq w_1, W_2 \leq w_2) = \Phi_2(w_1, w_2, \varrho) \tag{A.11}$$
$$P(W_1 \leq w_1, W_2 > w_2) = P(W_1 \leq w_1, -W_2 < -w_2) = \Phi_2(w_1, -w_2, -\varrho) \tag{A.12}$$
$$P(W_1 > w_1, W_2 \leq w_2) = P(-W_1 < -w_1, W_2 \leq w_2) = \Phi_2(-w_1, w_2, -\varrho) \tag{A.13}$$
$$P(W_1 > w_1, W_2 > w_2) = P(-W_1 < -w_1, -W_2 < -w_2) = \Phi_2(-w_1, -w_2, \varrho). \tag{A.14}$$

B Begriffe und Resultate aus der konvexen Analysis

In diesem Abschnitt werden wir kurz die wichtigsten Definitionen aus der Theorie konvexer Funktionen wiedergeben, die wir zur Diskussion der Existenz und Eindeutigkeit des Maximum-Likelihood-Schätzers benötigen. Die Minimierung konvexer Funktionen (bzw. die Maximierung konkaver Funktionen) behandelt Rockafellar (1970) in umfassender Form. Auch in den für Kapitel 5.2 relevanten Artikeln von Lesaffre & Kaufmann (1992), Kaufmann (1988a), Kaufmann (1988b) und Kaufmann (1988c) finden sich in überblicksmäßiger Darstellung einige elementare Grundlagen der konvexen Analysis.

B.1 Grundlegende Definitionen

Definition B.1. *Eine Teilmenge C von $\mathbb{R}^n$ heißt **konvex**, falls $(1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y} \in C$ für alle $\boldsymbol{x}, \boldsymbol{y} \in C$ und $0 < \lambda < 1$.*

Definition B.2. *Sei f eine Funktion die reelle Werte oder $\pm\infty$ annimmt und deren Definitionsbereich eine Teilmenge S von $\mathbb{R}^n$ ist. f heißt **konvex** auf S, falls die Menge*

$$\{(\boldsymbol{x}, \mu) | \boldsymbol{x} \in S, \mu \in \mathbb{R}, \mu \geq f(\boldsymbol{x})\}$$

eine konvexe Teilmenge von $\mathbb{R}^{n+1}$ ist.

Die beiden folgenden Theoreme geben einfachere Kriterien zur Nachprüfung der Konvexität einer Funktion an:

Theorem B.1. *Sei f eine Funktion von C nach $(-\infty, +\infty]$, wobei C eine konvexe Menge ist. Dann ist f konvex auf C genau dann, wenn*

$$f((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y})) \leq (1-\lambda)f(\boldsymbol{x})+\lambda f(\boldsymbol{y}) \ , \quad 0 < \lambda < 1, \tag{B.1}$$

für alle $\boldsymbol{x}$ und $\boldsymbol{y}$ in C gilt. f ist strikt konvex, falls in (B.1) *stets „$<$“ gilt.*

Beweis. Theorem 4.1. in Rockafellar (1970). □

Theorem B.2. *Sei f eine zweimal stetig differenzierbare, reellwertige Funktion auf einer offenen konvexen Menge $C \in \mathbb{R}^n$. f ist konvex auf C genau dann, wenn die Hessematrix*

$$Q_x := (q_{ij}(\boldsymbol{x})), q_{ij}(\boldsymbol{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x_1, \ldots, x_n)$$

für jedes $\boldsymbol{x} \in C$ positiv semidefinit ist. f ist strikt konvex, falls Q_x für jedes $\boldsymbol{x} \in C$ positiv definit ist.

Beweis. Theorem 4.5. in Rockafellar (1970). □

Definition B.3. *Sei f eine Funktion von C nach $[-\infty, +\infty)$, wobei C eine konvexe Menge ist. f heißt (strikt)* ***konkav*** *auf C, falls $-f$ (strikt) konvex auf C ist.*

Definition B.4. *Sei f eine Funktion von C nach $[-\infty, +\infty)$, wobei C eine konvexe Menge ist. f heißt* ***log-konkav*** *auf C, falls $\log f$ konkav auf C ist.*

Definition B.5. *Sei f eine Funktion von C nach $(-\infty, +\infty]$, wobei C eine konvexe Menge ist. Dann heißt f* ***affin****, falls*

$$f((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y})) = (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \ , \quad 0 < \lambda < 1,$$

für alle $\boldsymbol{x}$ und $\boldsymbol{y}$ in C gilt.

Theorem B.3. *Jede affine Funktion auf $\mathbb{R}^n$ ist von der Form*

$$f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{a} \rangle + \alpha \ , \quad \textit{mit}\, \boldsymbol{a} \in \mathbb{R}^n, \alpha \in \mathbb{R}.$$

Beweis. Spezialfall von Theorem 1.5. in Rockafellar (1970). □

Definition B.6. *Eine (log-)konkave Funktion f besitzt die* ***Definitheitseigenschaft****, falls die Funktion entlang jeder Richtung $\boldsymbol{d} \in \operatorname{dir} f$ entweder strikt (log-)konkav oder affin ist (siehe Kaufmann (1988b)).*

Definition B.7. *Eine reellwertige Funktion f auf $S \subset \mathbb{R}^n$ heißt* ***nach oben halbstetig*** *im Punkt $\boldsymbol{x} \in S$, falls gilt*

$$\lim_{i \to \infty} f(\boldsymbol{x_i}) \leq f(\boldsymbol{x})$$

für alle Folgen $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots$ in S, bei denen $\boldsymbol{x_i}$ gegen $\boldsymbol{x}$ konvergiert, und der Limes von $f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots$ in $[-\infty, +\infty]$ existiert. f ist also nach oben halbstetig in $\boldsymbol{x}$, wenn

$$f(\boldsymbol{x}) = \limsup_{\boldsymbol{y} \to \boldsymbol{x}} = \lim_{\epsilon \searrow 0} \left(\sup \{ f(\boldsymbol{y}) : \ |\boldsymbol{y} - \boldsymbol{x}| \leq \epsilon \} \right).$$

B.2 Wichtige Teilmengen von $\mathbb{R}^n$ bezüglich konvexer Funktionen

Im Zusammenhang mit der Minimierung konvexer (bzw. der Maximierung konkaver Funktionen) sind mehrere Teilmengen von $\mathbb{R}^n$ von besonderer Bedeutung. Zunächst betrachtet man den effektiven Definitionsbereich einer konvexen Funktion. Dieser ist folgendermaßen definiert:

Definition B.8. *Der effektive Definitionsbereich einer konvexen Funktion f ist gegeben durch*

$$\operatorname{dom} f := \{\boldsymbol{x} : f(\boldsymbol{x}) < +\infty\}.$$

Offensichtlich ist $\operatorname{dom} f$ eine konvexe Teilmenge von $\mathbb{R}^n$. Für konkave Funktionen gilt entsprechend

$$\operatorname{dom} f := \{\boldsymbol{x} : f(\boldsymbol{x}) > -\infty\}.$$

Definition B.9. *Das Innere einer konvexen Menge C relativ zu ihrer affinen Hülle* $\operatorname{aff} C$*, im Folgenden bezeichnet mit* $\operatorname{ri} C$*, ist definiert als*

$$\operatorname{ri} C := \{\boldsymbol{x} \in \operatorname{aff} C : \exists\, \epsilon > 0, (\boldsymbol{x} + \epsilon B) \cap (\operatorname{aff} C) \subset C\},$$

wobei $B := \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq 1\}$ die Einheitskugel bezüglich der Euklidischen Norm bezeichnet.

$\operatorname{ri} C$ enthält also all jene Vektoren $\boldsymbol{x} \in \operatorname{aff} C$, für die ein $\epsilon > 0$ existiert, so dass $\boldsymbol{y} \in C$ gilt für alle $\boldsymbol{y} \in \operatorname{aff} C$ mit $\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \epsilon$. Bei konvexen Funktionen betrachten wir das Innere des effektiven Definitionsbereichs, also $\operatorname{ri}(\operatorname{dom} f)$.

Ausgehend von $\operatorname{ri}(\operatorname{dom} f)$ bezeichnet $\operatorname{dir} f \subset \operatorname{ri} C$ eine weitere wichtige Menge von Vektoren:

Definition B.10. $\operatorname{dir} f$ *bezeichnet die Menge aller Vektoren $\boldsymbol{d}$, für die $\boldsymbol{x} + \lambda \boldsymbol{d}$ in* $\operatorname{dom} f$ *bleibt für genügend kleine $\lambda > 0$ für alle $\boldsymbol{x} \in \operatorname{ri}(\operatorname{dom} f)$, also*

$$\operatorname{dir} f := \{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{x} + \lambda \boldsymbol{d} \in \operatorname{dom} f \text{ für genügend kleine } \lambda\ \forall \boldsymbol{x} \in \operatorname{ri}(\operatorname{dom} f)\}$$

Für die Minimierung einer konvexen Funktion ist es sinnvoll, nur solche Richtungen $\boldsymbol{d}$ aus $\operatorname{dir} f$ zu betrachten, für die $f(\boldsymbol{x} + \lambda \boldsymbol{d})$ als Funktion von $\lambda \in \mathbb{R}$ nichtwachsend ist für alle $\boldsymbol{x} \in \mathbb{R}^d$. Die Menge all dieser Vektoren wird als Abstiegskegel $\operatorname{rec} f$ bezeichnet. Für konkave Funktionen definiert man $\operatorname{rec} f$ folgendermaßen:

Definition B.11. $\operatorname{rec} f$ *bezeichnet die Menge aller Vektoren $\boldsymbol{d} \in \operatorname{dir} f$, für die $f(\boldsymbol{x} + \lambda \boldsymbol{d})$ als Funktion von $\lambda \in \mathbb{R}$ nichtfallend ist für alle $\boldsymbol{x} \in \mathbb{R}^d$.*

Offensichtlich ist für $\boldsymbol{d} \in \operatorname{rec} f$ der effektive Definitionsbereich $\operatorname{dom} f$ in Richtung $\boldsymbol{d}$ unbeschränkt, also $\boldsymbol{x} + \lambda \boldsymbol{d} \in \operatorname{dom} f$ für alle $\lambda > 0$.

Literaturverzeichnis

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. *Page 267 of:* Petrox, B.N., & Caski, F. (eds), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Albert, A., & Anderson, J.A. 1984. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**, No. 1, 1–10.

Anderson, T.W. 2003. *An Introduction to Multivariate Statistical Analysis.* 3rd edn. Series in Probability and Statistics. Wiley.

Ashford, J.R., & Sowden, R.R. 1970. Multi-Variate Probit Analysis. *Biometrics*, **26**, 535–546.

Bartholomew, D.J., & Knott, M. 1999. *Latent Variable Models and Factor Analysis.* 2nd edn. Kendall's Library of Statistics 7. Arnold.

Brezger, A., & Lang, S. 2003. Generalized Structured Additive Regression based on Bayesian P-splines. *Discussion Paper 321, SFB 386.*

Brezger, A., Kneib, T., & Lang, S. 2003. BayesX: Analysing Bayesian Structured Additive Regression Models. *Discussion Paper 332, SFB 386.*

Chambers, E.A., & Cox, D.R. 1967. Discrimination Between Alternative Binary Response Models. *Biometrika*, **54**, No. 3/4, 573–578.

Chib, S., & Greenberg, E. 1998. Analysis of Multivariate Probit Models. *Biometrika*, **85**, No. 2, 347–361.

Chung, K.L. 1974. *A Course in Probability Theory.* New York: Academic Press.

Collett, D. 1999. *Modelling Binary Data.* London: Chapman & Hall.

Czado, C. 2000. Multivariate Regression Analysis of Panel Data with Binary Outcomes applied to Unemployment Data. *Statistical Papers*, **41**, 281–304.

Dale, J.R. 1986. Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses. *Biometrics*, **42**, 909–917.

Dobson, A.J. 1990. *An Introduction to Generalized Linear Models.* London: Chapman & Hall.

Fahrmeir, L., & Tutz, G. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd edn. Series in Statistics. Springer.

Findley, D.F. 1991. Counterexamples to Parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, **43**, 505–514.

Finney, D.J. 1973. *Statistical Method in Biological Assay*. 2nd edn. London: Hafner.

Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon*, Section A, Serie 3, **14**, 53–77.

Friedman, J. H., & Silverman, B. W. 1989. Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, **31**, 3–39.

Glonek, G.F.V., & McCullagh, P. 1995. Multivariate Logistic Models. *Journal of the Royal Statistical Society*, **Ser. B, 57**, No. 3, 533–546.

Gould, W., & Sribney, W. 1999. *Maximum Likelihood Estimation with Stata*. Texas: Stata Press.

Greene, W.H. 2003. *Econometric Analysis*. 5th edn. Upper Saddle River, New Jersey: Prentice Hall.

Haberman, S.J. 1977. Maximum Likelihood Estimates in Exponential Response Models. *The Annals of Statistics*, **5**, 815–841.

Hamdan, M.A. 1970. The Equivalence of Tetrachoric and Maximum Likelihood Estimates of ρ in 2×2 Tables. *Biometrika*, **57**, 212–215.

Hämmerlin, G., & Hoffman, K.H. 1990. *Numerische Mathematik*. Berlin: Springer Verlag.

Johnson, N.L., & Kotz, S. 1972. *Distributions in Statistics 4: Continuous Multivariate Distributions*. Series in Probability and Mathematical Statistics - Applied. Wiley.

Kass, R.E., & Raftery, A.E. 1995. Bayes Factors. *Journal of the American Statistical Association*, **90**, No. 430, 773–795.

Katz, R.W. 1981. On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics*, **23**, 243–249.

Kaufmann, H. 1988a. Existence and Uniqueness of Maximum Likelihood Estimators in Quantal and Ordinal Response Models. *Metrika*, **35**, 291–313.

Kaufmann, H. 1988b. On Directions of Strictness, Affinity and Constancy and the Minimum Set of a Proper Convex Function. *Optimization*, **19**, 157–167.

Kaufmann, H. 1988c. On Existence and Uniqueness of a Vector Minimzing a Convex Function. *ZOR - Methods and Models of Operations Research*, **32**, 357–373.

Kiefer, N.M. 1982. Testing for Dependence in Multivariate Probit Models. *Biometrika*, **69**, No. 1, 161–166.

Lang, S., & Brezger, A. 2004. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Le Cessie, S., & van Houwelingen, J.C. 1994. Logistic Regression for Correlated Binary Data. *Applied Statistics*, **43**, No. 1, 95–108.

Lesaffre, E., & Kaufmann, H. 1992. Existence and Uniqueness of the Maximum Likelihood Estimator for a Multivariate Probit Model. *Journal of the American Statistical Association*, **87**, No. 419, 805–811.

Lesaffre, E., & Molenberghs, G. 1991. Multivariate Probit Analysis: A Neglected Procedure in Medical Statistics. *Statistics in Medicine*, **10**, 1391–1403.

Lesaffre, E., Verbeeke, G., & Molenberghs, G. 1994. A Sensitivity Analysis of Two Multivariate Response Models. *Computational Statistics and Data Anlysis*, **17**, 363–391.

Liang, K., & Zeger, S. 1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, No. 1, 13–22.

Liang, K.Y., Zeger, S.L., & Qaqish, B. 1992. Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, **Ser. B, 54**, 3–40.

McCullagh, P., & Nelder, J. A. 1989. *Generalized Linear Models*. 2nd edn. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Molenberghs, G., & Lesaffre, E. 1994. Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *Statistics in Medicine*, **89**, No. 426, 633–644.

Morimune, K. 1979. Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis. *Econometrica*, **47**, No. 4, 957–975.

Pearson, K. 1901. Mathematical Contributions to the Theory of Evolution – VII. On the Correlation of Characters not quantitatively measurable. *Philosophical Transactions of The Royal Society of London, Series A*, **200**, 1–66.

Pindyck, R.S., & Rubinfeld, D.L. 1998. *Econometric Models and Econometric Forecasts*. Economics Series. Boston, Mass.: Irwin McGraw-Hill.

Plackett, R. L. 1954. A Reduction Formula for Normal Multivariate Integrals. *Biometrika*, **41**, 351–360.

Plackett, R. L. 1965. A Class of Bivariate Distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Prentice, R.L. 1988. Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics*, **44**, 1033–1048.

Prékopa, A. 1973. On Logarithmic Concave Measures and Functions. *Acta Scientiarum Mathematicarum (Szeged)*, **34**, 335–343.

Raftery, A.E. 1996. Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Biometrika*, **83**, No. 2, 251–266.

Rockafellar, R.T. 1970. *Convex Analysis.* Princeton University Press.

Santner, T.J., & Duffy, D.E. 1986. A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **73**, No. 3, 755–758.

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, No. 6, 461–464.

Sheppard, W.F. 1899. On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation. *Philosophical Transactions of the Royal Society of London, Series A*, **192**, 101–167.

Shibata, R. 1976. Selection of the Order of an Autoregressive Model by Akaike's Information Criterion. *Biometrika*, **63**, 117–126.

Silvapulle, M.J. 1981. On the Existence of Maximum Likelihood Estimators for the Binomial Response Models. *Journal of the Royal Statistical Society*, **Ser. B, 43**, No. 3, 310–313.

STATA Base Reference Manual, Release 8. 2003. *Volume 3, N–R.* Texas: Stata Press.

Tong, Y.L. 1990. *The Multivariate Normal Distribution.* Series in Statistics. Springer Verlag.

Tutz, G. 2000. *Die Analyse kategorialer Daten.* R. Oldenbourg Verlag München Wien.

Wedderburn, R.W.M. 1976. On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. *Biometrika*, **63**, 27–32.