Schöpp, Toutenburg:

Das symmetrische konditionale Regressionsmodell - alternative Parametrisierung bei korrelierten binären Responsevariablen

Sonderforschungsbereich 386, Paper 37 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

Das symmetrische konditionale Regressionsmodell – alternative Parametrisierung bei korrelierten binären Responsevariablen

A. Schöpp * H. Toutenburg*

25. Juli 1996

Zusammenfassung

Das hier vorgestellte spezifische Modell bietet mit dem dazugehörigen Schätzverfahren eine neue alternative Vorgehensweise für die regressive Analyse binärer korrelierter Zielgrößen. Das Schätzverfahren für das an die Korrelation angepaßte Modell wird über den Modellvergleich mit loglinearen Modellen und einer auf Odds-Ratios basierenden Reparametrisierung hergeleitet. Dabei wird zwischen verschiedenen Spezialfällen in Abhängigkeit von der Art und Anzahl der Einflußgrößen unterschieden. Die neue Methode besitzt gegenüber anderen den Vorteil, neben der einfacheren Berechnung der Schätzungen zugleich die Adäquatheit des Modells zu prüfen. Der theoretischen Darstellung folgt die auführliche Beschreibung des Verfahrens an zwei Datensätzen.

Keywords: Korrelierte binäre Variablen, Konditionale Regressionsmodelle, Loglineare Modelle, Odds-Ratio

1 Einführung

Die Theorie der generalisierten linearen Modelle bildet das Kernstck der Analyse von Zusammenhngen und Strukturen zwischen verschiedenen Untersuchungsgrößen in den unterschiedlichsten Anwendungsgebieten der Statistik. Die grundlegende Voraussetzung dieser allgemeinen Modellform und deren Schätzung ist die Unabhängigkeit der Responsevariablen. Da diese Annahme jedoch nicht immer erfüllt wird, ist die Untersuchung der Folgen einer Mißachtung der Korrelation der Zielgrößen sowie die Entwicklung von an die Abhängigkeit der Re-

*Institut für Statistik, LMU München, Ludwigstr. 33, 80539 München, Deutschland

sponsevariablen angepaßten Modellen ein zentrales Thema in der modernen Regressionsanalyse. Lösungsansätze bieten marginale Modelle, random-effects Modelle, Modelle mit latenten Variablen sowie konditionale (bedingte) Modelle, wobei sowohl die Zielsetzung der Analyse als auch die Art und Zahl der in die Regression involvierten Variablen ausschlaggebend für die Wahl der Modellform sind.

Das hier vorgestellte konditionale symmetrische Regressionsmodell (SRM) ist ein spezielles konditionales Modell. Im Gegensatz zu den asymmetrischen konditionalen Regressionsmodellen, die von Bonney (1987), [3], Fahrmeir and Kaufmann (1987), [5] und Fahrmeir and Tutz (1994, Ch. 6), [6], untersucht wurden und die auf eine zwischen den Zielgrößen existierende Ordnung aufbauen, werden in symmetrischen konditionalen Modellen alle Responsevariablen als gleichwertig betrachtet.

Ein Problem bei den von Autoren wie Connolly and Liang (1988), [4], Prentice (1988), [10], Qu et al. (1987), [11] und Rosner, [12], [13], [14], [15], [16], für sich gegenseitig gleichmäßig beeinflussende Responsevariablen entwickelten Formen des SRM ist die aufwendige und zum Teil auf subjektiven Annahmen aufbauende Parameterschätzung. Ein weiterer Nachteil der regressiven Analyse solcher Daten besteht in der dafür notwendigen Aufteilung der Variablen in exogene und endogene Größen. Häufig ist eine eindeutige Zuordnung der Variablen aufgrund mangelnder Informationen über die Daten schwierig, mit Informationsverlust verbunden und in den meisten Fällen ist eine Überprüfung der gemachten Annahmen nicht möglich. Loglineare Modelle umgehen diese Schwierigkeit, da hier nicht zwischen Response- und Regressorvariablen unterschieden wird. Das SRM nähert sich dem loglinearen Ansatz insofern, daß es im Gegensatz zu anderen Regressionsmodellen flexibler in Bezug auf die Definition der Einflußgrößen ist und sowohl Responsevariablen als auch echte exogene Variablen als beeinflussend betrachtet werden. Gegenüber den loglinearen Modellen, bei denen die zwischen den Variablen vorliegenden Zusammenhangsstrukturen in feinsten Abstufungen parametrisierbar sind, besitzt das SRM den Vorteil der direkteren und einfacheren Interpretation, da es für jede Einflußgröße bzw. für jede Kombination von Einflußgrößen einen expliziten Parameter enthält. Das hier vorgestellte Verfahren beruht auf der Überlegung, daß eine Kombination beider Modelltypen deren Vorteile verbinden würde. Deshalb wurde durch die Einbettung des spezifischen SRM in loglineare Modelle und der damit verbundenen Reparametrisierung basierend auf Odds-Ratios ein neues Schätzverfahren für diese Modellform entwickelt. Die neue Vorgehensweise bietet den Vorteil, daß damit nicht nur die Parameterschätzung im Vergleich zu allen anderen Verfahren vereinfacht wird, sondern sich zusätzlich die Möglichkeit ergibt, die Adäquatheit des SRM zu testen. Dies ist bei allen anderen regressiven Modellen nicht durchführbar, d.h. mit ihnen ist nur die Parameterschätzung ohne einen zugehörigen Modellanpassungstest möglich.

Im folgenden wird zunächst das SRM kurz vorgestellt und dann das zugehörige zweistufige Schätzverfahren, das anwenderfreundlich sein soll und das alle verfügbare Information aus den Daten verwendet, theoretisch hergeleitet. Die gesamte Vorgehensweise wird anschließend anhand zweier Beispieldatensätze näher erläutert.

2 Definition des SRM

In der empirischen Datenerhebung tritt häufig das Problem korrelierter Variablen auf. Ursache hierfür können Erhebungen in Form von Clustern sein, bei denen die Clustervariablen voneinander abhängen, oder wiederholte Beobachtungen an einem einzelnen Untersuchungsobjekt. Diese Situationen liegen beispielsweise oft in medizinischen oder zahmedizinischen Studien vor, wenn an einem Patienten beide Ohren, Augen oder mehrere Implantate etc. untersucht werden oder wenn in bestimmten Zeitabständen wiederholt Untersuchungen an einem Objekt (linkes Auge, Frontzahnimplantat) durchgeführt werden. Die zwischen solchen Daten vorliegende Korrelationsstruktur wird in sehr unterschiedlicher Form bei der Konstruktion von Regressionsmodellen berücksichtigt.

Konditionale Modelle beruhen dabei auf der Überlegung, daß die Korrelation innerhalb der Cluster durch explizite gegenseitige Beeinflussung der Variablen entsteht. Im Gegensatz zu allen anderen Modellformen werden im konditionalen Modell die Einflußgrößen sowohl durch die echten Regressorvariablen X, als auch durch die eigentlichen Responsevariablen Y gebildet. Die konditionale Modellform wird bevorzugt dann angewendet, wenn ein Response durch das Auftreten einer oder mehrerer Responsevariablen bedingt ist und das Interesse der Studie den bedingten Erwartungswerten der Zielvariablen, gegeben die restlichen Clustervariablen und die Kovariablen, sowie der Abhängigkeitsstruktur zwischen den Zielvariablen gilt.

Im folgenden sei Y_{ij} die j-te ($j = 1, \ldots, n_i$) Responsevariable des i-ten Clusters ($i = 1, \ldots, M$), wobei im Fall bivariater Responsevariablen nur zwei Responsevariablen pro Cluster i vorliegen ($n_i = 2$). X_{ij} bezeichne zur Responsevariablen Y_{ij} gehörende Einflußgrößen. Hier ist ebenfalls eine wichtige Unterscheidung zu treffen, ob es sich dabei um clusterspezifische oder einheitsspezifische Einflußgrößen handelt. Clusterspezifische Regressoren sind für alle Variablen Y_{ij} des Clusters i identisch, wogegen sich einheitsspezifische Einflußgrößen jeweils auf eine einzelne Zielgröße innerhalb des Clusters beziehen und daher für verschiedene Zielgrößen unterschiedliche Ausprägungen haben können. Eine clusterspezifische Einflußgröße ist z. B. das Alter eines Patienten (Cluster), bei dem beide Augen (Clustervariablen Y_1, Y_2) untersucht werden. Eine einheitsspezifische Einflußgröße wäre in diesem Beispiel die Farbpigmentierung, der Pupillendurchmesser etc..

Die regressive Darstellung des konditionalen Modells baut auf den in Longitudinalstudien häufig verwendeten Markov-Modellen auf, in denen die bedingte Verteilung der Variablen Y_{ij} zum Zeitpunkt t_j, gegeben die gesamte Vergangenheit $Y_{i1}, \ldots, Y_{ij-1}$, als Funktion der erklärenden Variablen X_{ij} und der in der Vergangenheit beobachteten Zielgrößen modelliert wird. Konditionale Modelle als generalisierte lineare Modelle erweitern diese Beziehung in der Form

$$g(E(Y_{ij}|H_{ij})) = g(\mu_{ij}|H_{ij}) = X'_{ij}\beta + \sum_{k=1}^{c} f_k(H_{ij};\alpha)\,, \tag{1}$$

mit X_{ij} als Einflußgrößen, g als Linkfunktion, β als Regressionsparameter, α als Korrelationsparameter, f als jeweiliger Transformationsfunktion und H_{ij} als Menge von Responsevariablen. Die Zusammensetzung der Menge H_{ij} teilt das konditionale Modell in zwei Formen: das asymmetrische und das symmetrische konditionale Modell. Im asymmetrischen Fall besteht H_{ij} ausschließlich aus den (zeitlich) vor Y_{ij} beobachteten Variablen, im symmetrischen Fall aus allen Responsevariablen außer Y_{ij}.

Da die Anwendung der asymmetrischen Modellform nur bei einer gegebenen natürlichen Ordnung der Daten sinnvoll ist, diese aber oft nicht vorliegt und somit unter subjektiver Einschätzung definiert werden müßte, wird meist das symmetrische konditionale Modell bevorzugt.

Eine Spezifizierung dieser Modellform für den Fall bivariater korrelierter binärer Größen ist das SRM (Symmetrisches Regressions Modell). Das SRM ist für zwei binäre Responsevariablen und für binäre Regressorvariablen konzipiert, wobei es durch geeignete Kategorisierung metrischer Merkmale auch auf diese verallgemeinert werden kann. Als Linkfunktion wird die Logit-Funktion verwendet, obwohl auch ein Log-Log- oder ein Probit-Link Alternativen wären. Der lineare Prediktor

$$\eta_i = g(E(Y_{ij}|Y_{ik}, X_{ij})\,, \quad j \neq k\,, \tag{2}$$

unterscheidet sich vom Prediktor bei unabhängigen Variablen im bivariaten Fall, also von

$$\eta_i = g(E(Y_{ij}|X_{ij})) = X'_{ij}\beta = \alpha + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} \tag{3}$$

durch eine zusätzliche Funktion der anderen, als „Einflußgröße“ fungierenden Zielvariablen, d.h.

$$\eta_i = \text{logit}(E(Y_{ij}|Y_{ik}, X_{ij})) = \alpha(Y_{ik}) + X'_{ij}\beta\,.$$

Der ursprünglich konstante Intercept α des Modells für unabhängige Zielgrößen ist im SRM ein von der Korrelation und der jeweiligen Ausprägung der betrachteten Variablen abhängiger Ausdruck.
Unter diesem Aspekt kann das SRM als eine Verallgemeinerung des Modells

für unabhängige Responsevariablen betrachtet werden, da sich bei der Anwendung des SRM dann wieder ein konstanter Intercept der Form $\alpha(0) = \alpha(1) = \alpha$ ergibt.

Da oft nicht nur die marginalen Wahrscheinlichkeiten der Form

$$P(Y_{ij}|Y_{ik}, X_{ij}) = \frac{\exp\left(\left(\alpha(Y_{ik}) + X'_{ij}\beta\right) Y_{ij}\right)}{1 + \exp\left(\alpha(Y_{ik}) + X'_{ij}\beta\right)} \tag{4}$$

bzw.

$$P(Y_{ij} = 1|Y_{ik}, X_{ij}) = \frac{\exp\left(\alpha(Y_{ik}) + X'_{ij}\beta\right)}{1 + \exp\left(\alpha(Y_{ik}) + X'_{ij}\beta\right)} \tag{5}$$

bei der regressiven Analyse von Datenstrukturen interessieren, sondern häufig auch gemeinsame Wahrscheinlichkeiten der Responsevariablen unter Berücksichtigung der Einflußgrößen bestimmt werden sollen, kann das SRM auch zur Berechnung von $P(Y_{i1}, Y_{i2}|X)$ verwendet werden.

Aufbauend auf den marginalen Wahrscheinlichkeiten (4) ist $P(Y_{i1}, Y_{i2}|X_i)$ wie folgt definiert:

$$P(Y_{i1}, Y_{i2}|X_{i1}, X_{i2}) = \tag{6}$$

$$= \frac{\exp\left[(\alpha(0) + X'_{i1}\beta)\, Y_{i1}\right] \exp\left[(\alpha(Y_{i1}) + X'_{i2}\beta)\, Y_{i2}\right]}{\exp\left(\alpha(0)+\alpha(1)+X'_{i1}\beta+X'_{i2}\beta\right)+\exp\left(\alpha(0)+X'_{i1}\beta\right)+\exp\left(\alpha(0)+X'_{i2}\beta\right)+1}$$

$$= s(X_{i1}X_{i2}) \left(\exp\left[(\alpha(0) + X'_{i1}\beta)\, Y_{i1}\right] \exp\left[\alpha(Y_{i1}) + X'_{i2}\beta)\, Y_{i2}\right]\right. . \tag{7}$$

Dabei ist $s(X_{i1}, X_{i2})$ de facto eine Kenngröße der Population im Sinne einer Skalierungseinheit, auf welche die eigentlich interessierenden Parameter $\alpha(0), \alpha(1)$ und β adjustiert werden.

In Zusammenhang mit der Korrelation zwischen binären Variablen gibt es eine weitere wichtige Kenngröße, den Odds-Ratio. Er ist ein Indikator für die Stärke und die Richtung der Korrelation und ist in indirekter Form das Verbindungsglied zwischen loglinearen Modellen und dem SRM, d.h. die im folgenden durchgeführte Reparametrisierung baut auf den Odds-Ratios des SRM und der loglinearen Modelle auf. Der Odds-Ratio besitzt im SRM eine bemerkenswerte Eigenschaft: er ist ein von den Einflußgrößen völlig unabhängiger und nur durch die Korrelationsparameter $\alpha(0), \alpha(1)$ bestimmter Ausdruck:

$$\begin{aligned} \text{OR} &= \frac{P(Y_{i1} = 1, Y_{i1} = 1|X_{i1}, X_{i2}) \cdot P(Y_{i1} = 0, Y_{i1} = |X_{i1}, X_{i2})}{P(Y_{i1} = 1, Y_{i1} = 0|X_{i1}, X_{i2}) \cdot P(Y_{i1} = 0, Y_{i1} = 1|X_{i1}, X_{i2})} \\ &= \exp\left[\alpha(1) - \alpha(0)\right] = \exp(k) . \end{aligned} \tag{8}$$

In der regressiven Darstellung des SRM kann der OR als Korrelationsparameter in der Logitgleichung verwendet werden, da gilt

$$\begin{aligned} \text{logit}P(Y_{ij} = 1|Y_{ik}, X_{ij}) &= \alpha(0) + \left[\alpha(1) - \alpha(0)\right] Y_{ik} + X'_{ij}\beta \\ &= \alpha(0) + kY_{ik} + X'_{ij}\beta . \end{aligned} \tag{9}$$

Die Darstellung (9) besitzt den Vorteil, daß die ursprüngliche Responsevariable Y_{ik} direkt mit einem eigenen Parameter als Einflußgröße auftritt und sich somit die Interpretation des Modells vereinfacht.

3 Loglineare Modelle und das SRM

3.1 Der Modellvergleich

Der Zusammenhang zwischen loglinearen Modellen und dem SRM wird durch eine auf Odds-Ratios basierende Reparametrisierung hergestellt. Der Odds-Ratio tritt in beiden Modellformen in umstrukturierter Form als Parameter auf: im SRM als $k = \ln \mathrm{OR} = \alpha(1) - \alpha(0)$ unabhängig von der Zahl der X-Variablen und beispielsweise im loglinearen Modell für zwei binäre Variablen als $\lambda_{11}^{Y_1 Y_2} = \frac{1}{4} \ln \mathrm{OR}$. Im loglinearen Modell mit drei binären Variablen gilt für den Dreifach-Wechselwirkungs-Parameter

$$\lambda_{111}^{Y_1 Y_2 X} = \frac{1}{8} \ln \frac{\theta_{1(1)1}}{\theta_{1(0)1}} = \frac{1}{8} \ln \frac{\theta_{(1)11}}{\theta_{(0)11}} = \frac{1}{8} \ln \frac{\theta_{11(1)}}{\theta_{11(0)}} ,$$

wobei $\theta_{1(1)1}, \theta_{(1)11}, \theta_{11(1)}$ bedingte Odds-Ratios sind, d.h. im loglinearen Modell sind die Assoziationsparameter proportional zu den logarithmierten Odds-Ratios. Die Parameter loglinearer Modelle und des SRM können über den Odds-Ratio direkt in Verbindung gebracht und unter bestimmten Voraussetzungen miteinander verbunden werden. Ein Modellvergleich aufbauend auf einer von Liang, Zeger and Qaqish (1992, S. 6), [8], mittels Odds-Ratios definierten Parametrisierung des loglinearen Modells mit dem SRM wird zeigen, daß die Kombination der Modelle nicht nur eine neue Möglichkeit bietet, die Parameter des SRM zu schätzen, sondern zusätzlich zu Restriktionen in den Parametern führt, deren Einhaltung die gerechtfertigte Anwendung des SRM garantiert. Dies ist notwendig, da für dessen Gültigkeit sowohl die symmetrische Betrachtungsweise der Responsevariablen, als auch die Definition der Einflußgröße als cluster- oder einheitsspezifisch zutreffend sein müssen.

Der Modellvergleich wird im folgenden exemplarisch für den Fall zweier binärer Responsevariablen und eines binären Regressors näher erläutert. Das von Liang, Zeger and Qaqish (1992), [8], für die allgemeine Darstellung der gemeinsamen Wahrscheinlichkeit der n Variablen gewählte loglineare Modell besitzt die Form

$$P(Y) = \exp\left(u_0 + \sum_{j=1}^{n} u_j y_j + \sum_{j<k} u_{jk} y_j y_k + \cdots + u_{12\cdots n} y_1 \cdots y_n \right) , \qquad (10)$$

wobei gilt

$$u_j \quad = \quad \mathrm{logit}[P(y_j = 1 | y_k = 0, k \neq j)] \quad j = 1, \ldots, n \qquad (11)$$

$$
\begin{aligned}
u_{jk} &= \log \mathrm{OR}(y_j, y_k | y_l = 0, l \neq j, k) \quad j < k = 1, \ldots, n \qquad (12)\\
u_{123} &= \log \mathrm{OR}(y_1, y_2 | y_3 = 1, y_l = 0, l > 3) \\
&\quad - \log \mathrm{OR}(y_1, y_2 | y_3 = 0, l > 3)\,. \qquad (13)
\end{aligned}
$$

Für die folgenden Überlegungen seien nun Y_1, Y_2 die interessierenden Responsevariablen und Y_3 die exogene Größe X. Dann gilt

$$
\begin{aligned}
P(Y, X) &= P(Y_1, Y_2, X) \\
&= \exp(u_0 + u_1 Y_1 + u_2 Y_2 + u_3 X + u_{12} Y_1 Y_2 + u_{13} Y_1 X + u_{23} Y_2 X + u_{123} Y_1 Y_2 X) \qquad (14)
\end{aligned}
$$

mit

$$
\begin{aligned}
u_1 &= \mathrm{logit} P(Y_1 = 1 | Y_2 = 0, X = 0) \\
u_2 &= \mathrm{logit} P(Y_2 = 1 | Y_1 = 0, X = 0) \\
u_3 &= \mathrm{logit} P(X = 1 | Y_1 = 0, Y_2 = 0) \\
u_{12} &= \ln \mathrm{OR}(Y_1, Y_2 | X = 0) = \ln \theta_{11(0)} \\
u_{13} &= \ln \mathrm{OR}(Y_1, X | Y_2 = 0) = \ln \theta_{1(0)1} \\
u_{23} &= \ln \mathrm{OR}(Y_2, X | Y_1 = 0) = \ln \theta_{(0)11} \\
u_{123} &= \ln \theta_{11(1)} - \ln \theta_{11(0)}
\end{aligned}
$$

Im Gegensatz zum Modell (10) bzw. (14), bei dem gemeinsame Wahrscheinlichkeiten ohne Nebenbedingungen beschrieben und somit alle Variablen als gleichwertig betrachtet werden, modelliert das SRM bedingte Wahrscheinlichkeiten, wobei die Bedingung durch die Einflußgröße X verkörpert wird. Hier gilt

$$
\begin{aligned}
P(Y|X) &= P(Y_1, Y_2 | X) \\
&= s(X) cdot\, (\exp[(\alpha(0) + \beta X) Y_1] \exp[(\alpha(Y_1) + \beta X) Y_2])
\end{aligned}
$$

mit

$$
\alpha(0) = \mathrm{logit} P(Y_1 = 1 | Y_2 = 0, X = 0) = \mathrm{logit} P(Y_2 = 1 | Y_1 = 0, X = 0)\,. \qquad (15)
$$

Das Problem des Vergleichs der bedingten mit den gemeinsamen Wahrscheinlichkeiten wird bei drei Variablen durch die Bildung und den Vergleich von Quotienten gelöst, denn nach Bayes gilt:

$$
\frac{P(a, b|k)}{P(c, d|k)} = \frac{\frac{P(a,b,k)}{P(k)}}{\frac{P(c,d,k)}{P(k)}} = \frac{P(a, b, k)}{P(c, d, k)}\,, \quad a, b, c, d, k = 0, 1 \qquad (16)
$$

Durch Einsetzen der entsprechend parametrisierten Form des SRM auf der linken Seite von (16) und des loglinearen Modells (14) auf der rechten Seite für alle möglichen Quotienten der jeweiligen Werte a, b, c, d, k von Y_1, Y_2 und X erhält man ein Gleichungssystem bestehend aus den Parametern $u_1, u_2, u_{12}, u_{13}, u_{23},$

u_{123} des loglinearen Modells und $\alpha(0), \alpha(1), \beta$ des SRM. Dessen Auflösung führt zu folgenden Bedingungen, deren Einhaltung die Adäquatheit des SRM gewährleistet:

$$\begin{aligned} u_1 &= u_2 \\ u_{13} &= u_{23} \\ u_{123} &= 0\,. \end{aligned} \tag{17}$$

Die Restriktionen sind unter dem Aspekt der symmetrischen Betrachtungsweise der Korrelation der Zielvariablen und durch die Clusterspezifität der Einflußgröße leicht zu erklären: Der gegenseitige Einfluß von Y_1 und Y_2 wird im loglinearen Modell durch die Logit-Parameter u_1 und u_2 beschrieben, d.h. nur wenn die bedingte Wahrscheinlichkeit von Y_1 gegeben Y_2 gleich der bedingten Wahrscheinlichkeit von Y_2 gegeben Y_1 ist ($u_1 = u_2$), kann von einer symmetrischen Korrelationsstruktur der vorliegenden abhängigen Größen ausgegangen werden und nur dann ist die Anwendung des SRM sinnvoll. Die Beziehung $u_{13} = u_{23}$ beschreibt die Annahme einer clusterspezifischen Einflußgröße X, die auf Y_1 und Y_2 dieselbe Wirkung hat. Nach Agresti (1990, S. 145), [1], ist die Restriktion $u_{123} = 0$ damit zu erklären, daß der Odds-Ratio im SRM unabhängig von der Ausprägung der Variablen X ist, d.h. $\theta_{11(1)} = \theta_{11(0)}$.

Da bei der späteren Kombination loglinearer Modelle mit dem SRM nicht die komplizierte From des loglinearen Modells von Liang, Zeger and Qaqish (1992), [8], verwendet wird, ist es sinnvoll, die Restriktionen durch die λ-Parameter des einfachen loglinearen Modells auszudrücken, d.h. durch entsprechende Umformung ergibt sich

$$\begin{aligned} \lambda_1 &= \lambda_2 \\ \lambda_{13} &= \lambda_{23} \\ \lambda_{123} &= 0\,. \end{aligned} \tag{18}$$

3.2 Spezialfälle

Die Restriktionen für den Fall zweier clusterspezifischer Einflußgrößen X_1, X_2 oder einer einheitsspezifischen Einflußgröße ergeben sich auf demselben Weg des Modellvergleichs und lassen sich genauso wie bereits im Fall einer clusterspezifischen Einflußgröße aus der Betrachtungsweise der Einflußgrößen und des Zusammenhangs der Responsevariablen begründen.

a) Eine einheitsspezifische Einflußgröße
   Falls X eine einheitsspezifische Einflußgröße ist, wie z.B. der Pupillendurchmesser eines Auges, so müssen für die Gültigkeit des SRM folgende

Bedingungen erfüllt sein:

$$\begin{aligned} \lambda_{111}^{Y_1Y_2X_1} = \lambda_{111}^{Y_1X_1X_2} = \lambda_{111}^{Y_1Y_2X_2} &= \lambda_{111}^{Y_2X_1X_2} = \lambda_{1111}^{Y_1Y_2X_1X_2} = 0 \\ \lambda_{11}^{Y_1X_2} &= \lambda_{11}^{Y_2X_1} = 0 \\ \lambda_{1}^{Y_1} &= \lambda_{1}^{Y_2} \\ \lambda_{11}^{Y_1X_1} &= \lambda_{11}^{Y_2X_2} \end{aligned} \tag{19}$$

b) Zwei clusterspezifische Einflußgrößen
Ähnlich wie in Fall a) muß bei zwei clusterspezifischen Einflußgrößen X_1, X_2 gelten:

$$\begin{aligned} \lambda_{111}^{Y_1Y_2X_1} = \lambda_{111}^{Y_1X_1X_2} = \lambda_{111}^{Y_1Y_2X_2} &= \lambda_{111}^{Y_2X_1X_2} = \lambda_{1111}^{Y_1Y_2X_1X_2} = 0 \\ \lambda_{1}^{Y_1} &= \lambda_{1}^{Y_2} \\ \lambda_{11}^{Y_1X_1} &= \lambda_{11}^{Y_2X_1} \\ \lambda_{11}^{Y_1X_2} &= \lambda_{11}^{Y_2X_2} \,. \end{aligned} \tag{20}$$

3.3 Kombination der Modellformen

Die Kombination loglinearer Modelle mit dem spezifischen SRM erfolgt durch eine Schachtelung beider Modellformen:
Zunächst wird an die vorliegenden Daten ein sogenanntes loglineares Grundmodell angepaßt, d.h. ein loglineares Modell, das den Nullrestriktionen in der Form entspricht, daß es alle λ-Parameter, die gemäß der jeweiligen Datensituation Null sein müssen, nicht enthält. Für das Grundmodell wird dann der G^2-Wert als Anpassungswert und der p-value berechnet. Wird das Grundmodell aufgrund des G^2-Wertes abgelehnt, so bricht das Verfahren hier ab, da dann bereits die Grundvoraussetzung für die Anwendung des SRM nicht gegeben ist. Es muß sowohl die Definition der Einflußgrößen als cluster- oder einheitsspezifisch, als auch die symmetrische Betrachtungsweise der Korrelation der Responsevariablen überprüft und eventuell ein völlig anderer regressiver Ansatz gewählt werden.
Führt dagegen der G^2-Wert nicht zur Ablehnung, so werden die jeweils verbleibenden Restriktionen für das Modell mittels eines geeigneten Tests geprüft. Hier ist sowohl die Verwendung des Wald-Tests als auch des Score-Tests möglich. Eine Ablehnung der Hypothese führt wie oben zum Abbruch des Verfahrens. Anderenfalls können jetzt die interessierenden Parameter $\alpha(0), \alpha(1), \beta$ des SRM direkt aus den einfach zu schätzenden loglinearen λ-Parametern berechnet werden. Auch deren Varianzen lassen sich aus den Varianzen der λ-Parameter, die man beispielsweise bei der Anwendung des Programms „Loggy 1“ (1993), [7], erhält, bestimmen. Diese relativ einfache Methode der Schätzung der Parameter des SRM umgeht die bisher dabei aufgetretenen Schwierigkeiten, daß sowohl bei der direkten Schätzung mittels der ML-Methode, als auch unter Verwendung von

a priori festgelegten Verteilungen der Responsevariablen, der Rechenaufwand erheblich ist und zudem subjektive Annahmen in die Schätzung miteingehen.

Das neue zweistufige Schätzverfahren ist mittels Standardmethoden durchführbar und kann als parameterfrei angesehen werden, da ausschließlich die aus den Daten direkt hervorgehende Information bei der Modellbildung und Schätzung verwendet wird. Zudem unterscheidet es sich von der bisherigen Vorgehensweise dadurch, daß es durch die primäre Analyse der Zusammenhänge zwischen den Variablen mittels loglinearer Modelle nicht mehr notwendig ist, vor Beginn der Analyse eine feste Korrelationsstruktur als gegeben anzusehen, wodurch eine passende Verteilung der Responsevariablen als Basis des SRM impliziert wird. Vielmehr wird über die Restriktionen die Abhängigkeitsstruktur in non- bzw. semiparametrischer Form überprüft und damit eine Fehlspezifikation und eine daraus resultierende falsche Einschätzung der echten Einflußgröße X vermieden.

4 Schätzung und Prüfung des SRM

4.1 Schätzung

Für die Schätzung der Modellparameter des SRM wird im ersten Schritt mit Hilfe von Backward-Elimination ein den Daten am besten angepaßtes loglineares Modell, das BE-Modell, bestimmt. Es beschreibt die für die jeweiligen Daten grundlegenden Zusammenhänge bei möglichst kleinem G^2-Wert und möglichst großem p-value. Da es jedoch vorkommen kann, daß größere Modelle Parameter enthalten, die zwar im BE-Modell nicht vorkommen aber trotzdem signifikant sind, wird zur Überprüfung der für die Anwendung des SRM notwendigen Restriktionen ein der Datensituation entsprechendes Grundmodell als Ausgangsmodell gewählt, d.h.

Fall 1: bei einer clusterspezifischen Einflußgröße:

$$\ln \pi_{klm} = \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X} + \lambda_{kl}^{Y_1 Y_2} + \lambda_{kr}^{Y_1 X} + \lambda_{lr}^{Y_2 X} + \ln n$$

Fall 2: bei einer einheitsspezifischen Einflußgröße:

$$\begin{aligned} \ln \pi_{klmr} &= \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_1} + \lambda_r^{X_2} \\ &\quad + \lambda_{kl}^{Y_1 Y_2} + \lambda_{km}^{Y_1 X_1} + \lambda_{lr}^{Y_2 X_2} + \lambda_{mr}^{X_1 X_2} + \ln n \end{aligned}$$

Fall 3: bei zwei clusterspezifischen Einflußgrößen:

$$\begin{aligned} \ln \pi_{klmr} &= \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_1} + \lambda_r^{X_2} \\ &\quad + \lambda_{kl}^{Y_1 Y_2} + \lambda_{km}^{Y_1 X_1} + \lambda_{lm}^{Y_2 X_1} + \lambda_{kr}^{Y_1 X_2} + \lambda_{lr}^{Y_2 X_2} + \lambda_{mr}^{X_1 X_2} + \ln n \,. \end{aligned}$$

Der Grund, warum nicht sofort mit dem Grundmodell begonnen wird, liegt darin, daß das BE-Modell einen sehr genauen Anhaltspunkt über die gesamten Zusammenhänge aller Variablen gibt und zudem gar nicht erst mit dem gesamten Verfahren begonnen werden muß, wenn im BE-Modell signifikante Dreifach-Wechselwirkungsparameter auftreten, da dann ein regressiver Ansatz in Form des SRM nicht passend ist.

Wenn in den Fällen 1–3 das Grundmodell aufgrund des G^2-Wertes nicht abgelehnt wird und somit nichts gegen die Nullhypothesen spricht, erfolgt die Überprüfung der weiteren Restriktionen, die nur bei Einhaltung der Nullrestriktionen erfüllt sein können. Die dafür allgemein formulierte Wald-Statistik lautet:

$$W = \delta' A' (A \Sigma_\delta A')^{-1} A \delta \,,$$

mit δ als p-Parametervektor bestehend aus λ-Parametern des jeweiligen Grundmodells, A als einer an δ angepaßten $c \times p$-Matrix mit Rang c und Σ_δ als Kovarianzmatrix des Parametervektors δ. Unter der Annahme, daß der Schätzer von δ bei Gültigkeit der Nullhypothese $H_0 : A\delta = 0$ asymptotisch normalverteilt ist mit Mittelwert 0 und Kovarianzmatrix $A\Sigma_{\hat{\delta}}A'$, ist W χ^2-verteilt mit c Freiheitsgraden. Im Einzelfall besitzt H_0 folgende Form:

Fall 1:

$$\begin{pmatrix} \lambda_1^{Y_1} - \lambda_1^{Y_2} \\ \lambda_{11}^{Y_1 X} - \lambda_{11}^{Y_2 X} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Fall 2:

$$\begin{pmatrix} \lambda_1^{Y_1} - \lambda_1^{Y_2} \\ \lambda_{11}^{Y_1 X_1} - \lambda_{11}^{Y_2 X_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Fall 3:

$$\begin{pmatrix} \lambda_1^{Y_1} - \lambda_1^{Y_2} \\ \lambda_{11}^{Y_1 X_1} - \lambda_{11}^{Y_2 X_1} \\ \lambda_{11}^{Y_1 X_2} - \lambda_{11}^{Y_2 X_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Die Ablehnung von H_0 ist gleichbedeutend damit, daß das SRM auf die jeweiligen Daten nicht angewendet werden kann. Kann H_0 nicht signifikant abgelehnt werden, können die Parameter $\alpha(0), \alpha(1), \beta$ bzw. β_1, β_2 unter Anwendung des bereits beschriebenen Modellvergleichs aus den geschätzten λ-Parametern berechnet werden.

Dazu wird Gleichung (16) durch Logarithmieren in die Form

$$\ln P(a, b|k) - \ln P(c, d|k) = \ln P(a, b, k) - \ln P(c, d, k) \qquad (21)$$

umgewandelt. Einsetzen der entsprechenden möglichen Formen des SRM auf der linken Seite und des loglinearen Modells auf der rechten Seite von (21) führt unter Einhaltung der Restriktionen und unter Berücksichtigung der in loglinearen Modellen geltenden Reparametrisierungsbedingungen zu einem eindeutig lösbaren Gleichungssystem.

Dessen Auflösung nach $\alpha(0), \alpha(1)$ und β bzw. β_1, β_2 unter Verwendung der Parameterschätzungen $\widehat{\lambda}_1^{Y_1}, \widehat{\lambda}_1^{Y_2}, \widehat{\lambda}_{11}^{Y_1Y_2}$ und $\widehat{\lambda}_{11}^{Y_1X}$ des loglinearen Grundmodells ergibt folgende Schätzungen:

Fall 1:

$$\begin{aligned} \widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X} \\ \widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X} \\ \widehat{\beta} &= 4\widehat{\lambda}_{11}^{Y_1X} \end{aligned} \tag{22}$$

Fall 2:

$$\begin{aligned} \widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} \\ \widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} \\ \widehat{\beta} &= 4\widehat{\lambda}_{11}^{Y_1X_1} \end{aligned} \tag{23}$$

Fall 3:

$$\begin{aligned} \widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} - 2\widehat{\lambda}_{11}^{Y_1X_2} \\ \widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} - 2\widehat{\lambda}_{11}^{Y_1X_2} \\ \widehat{\beta_1} &= 4\widehat{\lambda}_{11}^{Y_1X_1} \\ \widehat{\beta_2} &= 4\widehat{\lambda}_{11}^{Y_1X_2} \end{aligned} \tag{24}$$

Die Schätzungen (22), (23), (24) beruhen auf der Annahme der Hypothesen $\lambda_1^{Y_1} = \lambda_1^{Y_2}$, $\lambda_{11}^{Y_1X} = \lambda_{11}^{Y_2X}$, $\lambda_{11}^{Y_1X_1} = \lambda_{11}^{Y_2X_2}$, $\lambda_{11}^{Y_1X_1} = \lambda_{11}^{Y_2X_1}$ und $\lambda_{11}^{Y_1X_2} = \lambda_{11}^{Y_2X_2}$, d.h. prinzipiell ist es egal, welche der Parameter für die Bestimmung von $\alpha(0), \alpha(1), \beta$ bzw. β_1 und β_2 verwendet werden. Allderdings können bei der genauen Berechnung mittels der Parameterschätzungen der λ's Schwankungen zwischen den Parametern des SRM auftreten. Die Unterschiede kommen dadurch zustande, daß zwar die „Gleichheitshypothesen“ nicht signifikant abgelehnt werden können, die tatsächlichen Schätzwerte aber trotzdem nicht exakt übereinstimmen.

Die hier beschriebene Methode, bei der es mehrere Möglichkeiten gibt, die Parameterschätzungen des SRM zu berechnen, ist also nur dann ausreichend, wenn das SRM ausschließlich der Analyse und der Entdeckung gewisser Trends zwischen den Variablen dient. Falls jedoch exakte Schätzwerte benötigt werden,

bieten sich zwei Alternativen: Entweder werden die SRM-Parameter aus den Summen der Werte, d.h. aus $\lambda_1^{Y_1} + \lambda_1^{Y_2}$, $\lambda_{11}^{Y_1 X} + \lambda_{11}^{Y_2 X}$ etc. berechnet, oder die Schätzung wird mit Parametern eines spezifizierten Grundmodells durchgeführt. Der erste Weg umgeht zwar das Problem der Entscheidung zwischen den Parametern und liefert sehr ähnliche Ergebnisse wie das zweite Verfahren, besitzt aber den Nachteil, daß für die Berechnungen der Varianzen und für die Signifikanzbestimmung größere Kovarianzmatrizen verwendet werden müssen und somit die Berechnungen erheblich aufwendiger werden. Im Gegensatz hierzu vereinfacht die Anwendung des spezifizierten Grundmodells den gesamten Schätz- und Testvorgang. Es ergibt sich aus dem „normalen" Grundmodell (Modell ohne Parameter, die gemäß den Restriktionen Null sein müssem) indem alle Parameter zu einem neuen Parameter zusammengefaßt werden, die gemäß den Restriktionen übereinstimmen sollen. So entsteht ein bedingtes loglineares Modell, das bereits alle geforderten Restriktionen enthält.

Bei der Überprüfung der Güte des Modells werden somit gleichzeitig alle Restriktionen getestet, d.h. die explizite Durchführung des Wald-Tests entfällt, da der LQ-Test für das spezifizierte Grundmodell auch die jeweils zu erfüllenden Restriktionen überprüft. Falls der Test nicht zur Ablehnung führt, können die für den loglinearen Ansatz geschätzten Parameter direkt zur Berechnung der SRM-Parameter verwendet werden, ohne daß vorher zwischen irgendwelchen λ-Parametern differenziert werden muß. Das spezifizierte Grundmodell enthält weniger Parameter, ist einfacher zu schätzen und in Bezug auf Zusammenhangsstrukturen zwischen den Variablen leichter zu interpretieren als das normale Grundmodell. Allgemein muß jedoch bei der Wahl des Schätzverfahrens berücksichtigt werden, daß alle auf demselben theoretischen Hintergrund aufbauen und rein asymptotisch geltende Schätzungen und Aussagen liefern. Daher ist es meist sinnvoll, nicht nur eine der Schätzmöglichkeiten anzuwenden, sondern alle parallel durchzuführen, um dann die Ergebnisse vergleichen zu können. In der Mehrzahl der Fälle wird man zu ähnlichen Resultaten gelangen, d.h. entweder führen alle Verfahren zum Abbruch oder es ergeben sich sinnvolle, interpretierbare Schätzungen der Parameter des SRM. Vorsicht bei der Auswertung ist in den Grenzfällen geboten, in denen ein Verfahren zum Abbruch führt, das andere aber nicht.

4.2 Varianzbestimmung

Die Varianzen der Parameterschätzungen $\widehat{\alpha(0)}, \widehat{\alpha(1)}, \widehat{\beta}$ bzw. $\widehat{\beta}_1, \widehat{\beta}_2$ werden wieder unter Verwendung der loglinearen Parameterschätzungen und deren Varianzen bestimmt, die sich beispielsweise bei Verwendung des Programms „Loggy 1" ergeben. Die Schätzungen $\widehat{\alpha(0)}, \widehat{\alpha(1)}$ und $\widehat{\beta}, \widehat{\beta}_1, \widehat{\beta}_2$ sind im einzelnen Fall als Linearkombinationen des jeweiligen Parametervektors $\hat{\delta}$ der loglinearen Parameter des Grundmodells darstellbar.

Es gilt beispielsweise im Fall 1 einer clusterspezifischen Einflußgröße mit $\delta = (\hat{\mu}, \hat{\lambda}_1^{Y_1}, \hat{\lambda}_1^{Y_2}, \hat{\lambda}_1^{X}, \hat{\lambda}_{11}^{Y_1Y_2}, \hat{\lambda}_{11}^{Y_1X}, \hat{\lambda}_{11}^{Y_2X})'$ folgender Zusammenhang:

$$\begin{aligned}
\widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X} = (0,2,0,0,-2,-2,0)\widehat{\delta} \\
&= C_1'\widehat{\delta} \\
\widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X} = (0,2,0,0,2,-2,0)\widehat{\delta} \\
&= C_2'\widehat{\delta} \\
\widehat{\beta} &= 4\widehat{\lambda}_{11}^{Y_1X} = (0,0,0,0,0,4,0)\widehat{\delta} = C_3'\widehat{\delta}\,.
\end{aligned} \tag{25}$$

Unter Verwendung der Kovarianzmatrix $\widehat{\Sigma}_{\hat{\delta}}$ des loglinearen Modells resultiert daraus folgendes Ergebnis für die Varianzen:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\alpha(0)}) &= C_1'\widehat{\Sigma}_{\hat{\delta}}C_1 \\
\widehat{\mathrm{Var}}(\widehat{\alpha(1)}) &= C_2'\widehat{\Sigma}_{\hat{\delta}}C_2 \\
\widehat{\mathrm{Var}}(\widehat{\beta}) &= C_3'\widehat{\Sigma}_{\hat{\delta}}C_3\,.
\end{aligned} \tag{26}$$

Die Vektoren $\widehat{\delta}, C_1, C_2, C_3$ variieren entsprechend des Typs der Einflußgröße im SRM und damit entsprechend der Anzahl der Variablen des loglinearen Grundmodells. Im Fall 2 (eine einheitsspezifische Einflußgröße) mit

$$\widehat{\delta} = (\hat{\mu}, \hat{\lambda}_1^{Y_1}, \hat{\lambda}_1^{Y_2}, \hat{\lambda}_1^{X_1}, \hat{\lambda}_1^{X_2}, \hat{\lambda}_{11}^{Y_1Y_2}, \hat{\lambda}_{11}^{Y_1X_1}, \hat{\lambda}_{11}^{Y_2X_2}, \hat{\lambda}_{11}^{X_1X_2})'$$

gilt

$$\begin{aligned}
\widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} \\
&= (0,2,0,0,0,-2,-2,0,0)\widehat{\delta} = C_1'\widehat{\delta} \\
\widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} \\
&= (0,2,0,0,0,2,-2,0,0)\widehat{\delta} = C_2'\widehat{\delta} \\
\widehat{\beta} &= 4\widehat{\lambda}_{11}^{Y_1X_1} \\
&= (0,0,0,0,0,0,4,0,0)\widehat{\delta} = C_3'\widehat{\delta}\,.
\end{aligned} \tag{27}$$

Ähnlich wie im Fall mit zwei clusterspezifischen Einflußgrößen besitzt $\widehat{\delta}$ im Fall 3 die Form

$$\widehat{\delta} = (\hat{\mu}, \hat{\lambda}_1^{Y_1}, \hat{\lambda}_1^{Y_2}, \hat{\lambda}_1^{X_1}, \hat{\lambda}_1^{X_2}, \hat{\lambda}_{11}^{Y_1Y_2}, \hat{\lambda}_{11}^{Y_1X_1}, \hat{\lambda}_{11}^{Y_1X_2}, \hat{\lambda}_{11}^{Y_2X_1}, \hat{\lambda}_{11}^{Y_2X_2}, \hat{\lambda}_{11}^{X_1X_2})'$$

und es gilt

$$\begin{aligned}
\widehat{\alpha(0)} &= 2\widehat{\lambda}_1^{Y_1} - 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} - 2\widehat{\lambda}_{11}^{Y_1X_2} \\
&= (0,2,0,0,0,-2,-2,-2,0,0,0)\widehat{\delta} = C_1'\widehat{\delta}
\end{aligned}$$

$$
\begin{aligned}
\widehat{\alpha(1)} &= 2\widehat{\lambda}_1^{Y_1} + 2\widehat{\lambda}_{11}^{Y_1Y_2} - 2\widehat{\lambda}_{11}^{Y_1X_1} - 2\widehat{\lambda}_{11}^{Y_1X_2} \qquad (28)\\
&= (0,2,0,0,0,2,-2,-2,0,0,0)\widehat{\delta} = C_2'\widehat{\delta}\\
\widehat{\beta_1} &= 4\widehat{\lambda}_{11}^{Y_1X_1}\\
&= (0,0,0,0,0,0,4,0,0,0,0,0)\widehat{\delta} = C_3'\widehat{\delta}\\
\widehat{\beta_2} &= 4\widehat{\lambda}_{11}^{Y_1X_2}\\
&= (0,0,0,0,0,0,0,4,0,0,0,0)\widehat{\delta} = C_4'\widehat{\delta}\,.
\end{aligned}
$$

Die Varianzen werden in den einzelnen Fällen gemäß der Gleichungen (26) unter Verwendung der jeweiligen Vektoren $\widehat{\delta}, C_1, C_2, C_3$ bzw. C_4 bestimmt.

4.3 Prüfung des Modells

Die Anpassung des SRM an die Daten wird bereits während der Schätzung der Modellparameter durchgeführt. Die Signifikanz der einzelnen Parameter $\alpha(0), \alpha(1), \beta$ bzw. β_1, β_2 kann durch die Bildung von standardisierten Parameterschätzungen getestet werden, die asymptotisch $N(0;1)$-verteilt sind.

5 Beispiele

Für die praktische Darstellung des Verfahrens werden zwei Datensätze verwendet: der sogenannte „Zwillingszahndatensatz“ aus einer Studie von Dr. W. Walther, die von 1983-1989 durchgeführt wurde und der „Wankdatensatz“ aus der Studie „Der Kronenzustand der Fichten am Wank bei Garmisch-Partenkirchen 1986-1992“ , die von der Bayerischen Landesanstalt für Wald- und Forstwirtschaft durchgeführt wurde.

5.1 Die Zwillingszahndaten

In der Zahnärztlichen Poliklinik Karlsruhe wurden im Zeitraum 1983 bis 1989 insgesamt 331 „Zwillingspaare“ (d.h. zwei Versorgungen je Patient, Cluster vom Umfang 2) mit herausnehmbaren Konuskronenkonstruktionen versorgt. Von den 662 Konuskronen weisen 50 fehlende Werte auf, daher wurden diese Patienten ausgeschlossen und nur die verbleibenden 612 vollständigen Zwillingsdaten für die Schätzung der Regressionsparameter berücksichtigt. Die Cluster werden durch die Zwillingspaare gebildet, wobei die Zwillinge selbst die untersuchten Einheiten sind. Die Responsevariable Y_{ij}, $(i = 1, \ldots, 306,\ j = 1, 2)$, beschreibt

das Ereignis des Verlustes eines Pfeilerzahnes, d.h.

$$Y_{ij} = \begin{cases} 1 & \text{falls Pfeilerverlust bei Zwilling } j \\ & \text{im } i\text{-ten Zwillingspaar auftritt} \\ 0 & \text{falls kein Pfeilerverlust bei Zwilling } j \\ & \text{im } i\text{-ten Zwillingspaar auftritt}. \end{cases}$$

Ziel der regressiven Analyse ist die Untersuchung des Zusammenhangs zwischen dem Zielereignis „Verlust eines Pfeilerzahns“ und den prognostischen Faktoren Alter, Geschlecht und Form unter Berücksichtigung der zwischen den beiden Zwillingen eines Zwillingspaares vorliegenden Korrelation. Der prognostische Faktor „Alter“ wurde durch einen Splitpunkt a ($a = 60$ bzw. $a = 65$) in zwei Klassen geteilt, der Faktor „Form“ beschreibt das dentoalveoläre oder transversale Design der Konuskronenkonstruktion. Es wurde 146 mal ein dentoalveoläres und 160 mal ein transversales Design verwendet. Die 306 untersuchten Patienten setzen sich aus 139 Männern und 167 Frauen zusammen, wobei zum Zeitpunkt der prothetischen Versorgung der jüngste Patient 29.63, der älteste 85.98 Jahre alt war und das mittlere Alter aller Patienten 58.60 Jahre betrug. Alle Kovariablen sind binär und clusterspezifisch. Beide Zwillinge besitzen dasselbe Alter, dasselbe Geschlecht und bei jedem Paar wurde eine identische Konuskronenkonstruktion verwendet.

Die folgende Tafel gibt einen Überblick über die Anzahl und die Kombination der aufgetretenen Verluste von Pfeilerzähnen bei Zwillingspaaren:

		Y_{i1}		
		1	0	
Y_{i2}	1	23	12	35
	0	9	262	271
		32	274	306

Durch Hinzunahme jeweils einer der drei interessierenden Einflußgrößen (Alter, Geschlecht, Form) wird diese Tafel spezifiziert. Daraus ergeben sich vier Fälle, die in den folgenden vier Tafeln zusammengefaßt sind:

			Y_{i1}	
			1	0
$X_F = 1$	Y_{i2}	1	18	2
		0	8	118
$X_F = 0$	Y_{i2}	1	5	7
		0	4	144

Tabelle 1: (Fall 1): $X_F = 1$: dentoalveoläres Design; $X_F = 0$: transversales Design

			Y_{i1}	
			1	0
$X_{A1}=1$	Y_{i2}	1	7	4
		0	5	147
$X_{A1}=0$	Y_{i2}	1	16	5
		0	7	115

Tabelle 2: (Fall 2): $X_{A1}=1$: Alter ≤ 60 Jahre; $X_{A1}=0$: Alter >60 Jahre

			Y_{i1}	
			1	0
$X_{A2}=1$	Y_{i2}	1	11	6
		0	7	193
$X_{A2}=0$	Y_{i2}	1	12	3
		0	5	69

Tabelle 3: (Fall 3): $X_{A2}=1$: Alter ≤ 65 Jahre; $X_{A2}=0$: Alter >65 Jahre

			Y_{i1}	
			1	0
$X_G=1$	Y_{i2}	1	11	4
		0	6	118
$X_G=0$	Y_{i2}	1	12	5
		0	6	144

Tabelle 4: (Fall 4): $X_G=1$: Geschlecht männlich; $X_G=0$: Geschlecht weiblich

Die Durchführung des Verfahrens für ein Modell mit zwei Einflußgrößen ist hier nicht sinnvoll, da die dafür gebildeten Kontingenztafeln Nullzellen und viele nur sehr schwach besetzte Zellen enthalten und das Verfahren vorzeitig abbricht oder zu nutzlosen und schwer interpretierbaren Ergebnissen führt. In den Fällen 1–4 können sowohl der Restriktionentest als auch die Parameterschätzungen durchgeführt werden, wobei die erste Analyse mittels loglinearer Modelle folgende BE-Modelle ergibt:

1	$\ln \pi_{klm} = \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_F} + \lambda_{kl}^{Y_1 Y_2} + \lambda_{lm}^{Y_2 X_F} + \ln n$ $G^2 = 2.5144 < 7.38 = \chi^2_{2;0.975}$ $\quad p$-value $= 0.2844$
2	$\ln \pi_{klm} = \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_{A1}} + \lambda_{kl}^{Y_1 Y_2} + \lambda_{lm}^{Y_2 X_{A1}} + \ln n$ $G^2 = 0.9117 < 7.38 = \chi^2_{2;0.975}$ $\quad p$-value $= 0.6339$
3	$\ln \pi_{klm} = \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_{A2}} + \lambda_{kl}^{Y_1 Y_2} + \lambda_{lm}^{Y_2 X_{A2}} + \ln n$ $G^2 = 0.5983 < 7.38 = \chi^2_{2;0.975}$ $\quad p$-value $= 0.7564$
4	$\ln \pi_{klm} = \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_G} + \lambda_{kl}^{Y_1 Y_2} + \ln n$ $G^2 = 0.1736 < 9.35 = \chi^2_{3;0.975}$ $\quad p$-value $= 0.9817$

Das Resultat der Parameterschätzung ist in folgender Tafel zusammengefaßt:

		Schätzung (Schätzung/Standardabweichung)				
		mit $\widehat{\lambda}_1^{Y_1}$ $\widehat{\lambda}_{11}^{Y_1X}$	mit $\widehat{\lambda}_1^{Y_2}$ $\widehat{\lambda}_{11}^{Y_2X}$	mit $\widehat{\lambda}_1^{Y_1}+\widehat{\lambda}_1^{Y_2}$ $\widehat{\lambda}_{11}^{Y_1X}+\widehat{\lambda}_{11}^{Y_2X}$	mit Alternativmethode $\widehat{\lambda}_1^{*}$ $\widehat{\lambda}_{11}^{**}$	G^2-Wert (p-value)
Fall 1	$\widehat{\alpha(0)}$	-0.8765 (-1.614)	-0.1184 (-0.241)	-0.4975 (-1.341)	-0.3840 (-1.085)	6.2668 (0.0993)
	$\widehat{\alpha(1)}$	3.2840 (8.234)	4.006 (7.846)	3.6274 (13.001)	3.5370 (13.606)	
	$\widehat{\beta}$	0.3000 (0.559)	-1.4832 (-2.752)	-0.5915 (-2.654)	-0.6025 (-2.732)	
	$s(0)$	0.0797	0.0194	0.0398	0.0391	
	$s(1)$	0.0462	0.2553	0.1152	0.1148	
	OR	64.1036	61.8307	61.8617	50.4509	
Fall 2	$\widehat{\alpha(0)}$	-0.8223 (-2.033)	-1.1577 (-2.627)	-0.9899 (-3.224)	-0.9795 (-3.207)	0.4463 (0.9305)
	$\widehat{\alpha(1)}$	3.1312 (7.702)	2.7958 (7.796)	2.9634 (11.870)	2.9521 (11.932)	
	$\widehat{\beta}$	0.4786 (0.953)	0.5896 (1.224)	0.5341 (2.483)	0.5355 (2.492)	
	$s(0)$	0.0838	0.1476	0.1119	0.1119	
	$s(1)$	0.0347	0.0532	0.0426	0.0428	
	OR	52.1070	52.1175	52.1070	50.9885	
Fall 3	$\widehat{\alpha(0)}$	-0.8514 (-1.885)	-1.3511 (-2.723)	-1.1012 (-3.503)	-1.0889 (-3.487)	0.5832 (0.9003)
	$\widehat{\alpha(1)}$	3.106 (6.518)	2.6065 (6.363)	2.8564 (10.704)	2.8401 (10.770)	
	$\widehat{\beta}$	0.3786 (0.748)	0.72230 (1.494)	0.5512 (2.669)	0.5555 (2.695)	
	$s(0)$	0.0878	0.1987	0.1343	0.1345	
	$s(1)$	0.0442	0.05894	0.0511	0.0508	
	OR	52.3210	52.3316	52.3316	50.8561	
Fall 4	$\widehat{\alpha(0)}$	-0.6770 (-2.398)	-0.8572 (-1.916)	-0.7671 (-2.398)	-0.7537 (-2.371)	0.5033 (0.9182)
	$\widehat{\alpha(1)}$	3.3471 (8.328)	3.1669 (8.418)	3.2569 (13.164)	3.2463 (13.233)	
	$\widehat{\beta}$	0.0542 (0.110)	-0.1766 (-0.373)	-0.0612 (-0.302)	-0.0640 (-0.317)	
	$s(0)$	0.0608	0.0839	0.0715	0.0713	
	$s(1)$	0.0551	0.1138	0.0797	0.0799	
	OR	55.9299	55.9299	55.9244	54.5982	

Das für die Modellprüfung und Parameterschätzung entwickelte Programm „Korr“ berechnet die Schätzungen sowohl mit $\lambda_1^{Y_1}$, als auch mit $\lambda_1^{Y_2}$ und unter Verwen-

dung der Summe beider λ-Parameter sowie mittels der in Abschnitt 4.1 geschilderten Methode eines spezifizierten Grundmodells (Alternativmethode). Aus den geschätzten Parametern ergibt sich, daß die Variablen „Konuskronenkonstruktion“ und „Alter“ einen schwachen Einfluß auf das Zielereignis „Verlust eines Pfeilerzahns“ haben. Das Geschlecht hat hingegen keine Wirkung. Die Odds-Ratios sind in allen vier Fällen sehr groß, d.h. beide Zielgrößen Y_{i1}, Y_{i2}, sind stark positiv korreliert. Eine Umrechnung der Odds-Ratios in den Parameter $k = \ln OR$ der Modellform (10) zeigt, daß dieser immer in der Nähe von 4 liegt und somit die Wirkung der „Einflußgröße“ Y_{ij} auf Y_{ik} sehr groß ist. Zudem wird beim Vergleich der Odds-Ratios für die einzelnen Berechnungsmethoden deutlich, daß sie kaum voneinander abweichen und es somit für Trendausagen ohne Bedeutung ist, welcher loglinearer Parameter zur Berechnung und Schätzung verwendet wird.

5.2 Die Wankdaten

Dieser Datendatz entstammt der Studie „Der Kronenzustand der Fichten am Wank bei Garmisch-Partenkirchen 1986 – 1989 – 1992“ , die von der Bayerischen Landesanstalt für Wald- und Forstwirtschaft durchgeführt wurde. Über das gesamte Untersuchungsgebiet wurde ein geographisches Rasternetz gelegt, so daß jedes Raster vier Bäume enthält. Aus jedem Raster werden zufällig zwei Bäume ausgewählt, d.h die einzelnen Raster bilden die Cluster und die jeweils darin ausgewählten zwei Bäume die untersuchten Einheiten. Aufgrund der geographischen Lage und desselben Alters der beiden Bäume im Cluster muß von einer Korrelation zwischen ihnen ausgegangen werden. Insgesamt wurden 1282 Bäume bzw. 641 Cluster untersucht. Die Beurteilung des Kronenzustandes der ausgewählten Fichten wurde von zwei Faktoren abhängig gemacht: dem Nadelverlust und der Vergilbung der Assimilationsorgane (Chlorose). Beide Responsevariablen wurden optisch mit Hilfe von Infrarot-Luftbildern erhoben und in Stufen eingeteilt. Die Vergilbung der Assimilationsorgane wurde zwei Chlorosestufen mit Splitpunkt $< 26\%$ und $> 26\%$ zugeordnet. Da ein bestimmter Teil der Bäume bereits so stark entnadelt war, daß keine Einstufung der Chlorose mehr möglich war, existiert für die Responsevariable „Chlorose“ noch eine dritte Kategorie, die jedoch der hohen Chlorosestufe ($> 26\%$) zugerechnet wird. Die binäre Responsevariable „Chlorose“ besitzt demnach die Form

$$Y^C = \begin{cases} 0 & \text{bei weniger als 26\% vergilbter Nadelmasse} \\ 1 & \text{bei mehr als 26\% vergilbter Nadelmasse.} \end{cases}$$

Für den Nadelverlust wurden die vier Entnadelungsstufen 0% - 25%, 26% - 45%, 46% - 60% und $> 60\%$ gebildet. Die Variable „Entnadelung“ wurde zur Analyse der Daten dichotomisiert gemäß

$$Y^E = \begin{cases} 0 & \text{bei weniger als 26\% Entnadelung} \\ 1 & \text{bei mehr als 26\% Entnadelung.} \end{cases}$$

Folgende Einflußgrößen wurden erhoben:

- Alter

$$X_A = \begin{cases} 0 & \text{Alter } 60 - 100 \text{ Jahre} \\ 1 & \text{Alter} > 100 \text{ Jahre} \end{cases}$$

- Beschirmungsgrad

$$X_B = \begin{cases} 0 & \text{Beschirmungsgrad} \leq 60\% \\ 1 & \text{Beschirmungsgrad} > 60\% \end{cases}$$

- Höhe (Höhenlage der Fichten)

$$X_H = \begin{cases} 0 & \text{Höhe} \leq 1200\text{m} \\ 1 & \text{Höhe} > 1200\text{m} \end{cases}$$

Von den 1282 beobachteten Bäumen waren 442 älter als 100 Jahre, 840 jünger. 376 der insgesamt 641 Raster lagen unterhalb einer Höhe von 1200 m, 265 darüber. Bei 212 Bäumen konnte ein Beschirmungsgrad von mehr als 60% festgestellt werden, dagegen war dieser bei 429 Bäumen kleiner als 60%. 75.59% der Bäume (969) wiesen eine nur geringe Vergilbung auf (Chlorose $< 26\%$), dagegen waren bereits 77% (988) der Bäume stark entnadelt. Die nachstehenden Tafeln zeigen die Häufigkeitsverteilung der beiden interessierenden Responsevariablen Y_{ij}^C und Y_{ij}^E, die jeweils an beiden Bäumen ($j = 1, 2$) eines Clusters i ($i = 1, \ldots, 641$) erhoben wurden:

		Y_{i1}^E		
		1	0	
Y_{i2}^E	1	412	78	490
	0	86	65	151
		498	143	641

		Y_{i1}^C		
		1	0	
Y_{i2}^C	1	69	93	162
	0	82	397	479
		151	490	641

Die Untersuchung der Wirkung der Faktoren „Alter“ „Beschirmungsgrad“ und „Höhenlage“ auf die Zielgrößen wird zur Demonstration des Verfahrens beispielhaft für vier verschiedene Modelle durchgeführt, die in den folgenden Tafeln (mit zwei clusterspezifischen Einflußgrößen) zusammengefaßt sind:

				Y_{i1}^C	
				1	0
$X_H = 1$	$X_A = 1$	Y_{i2}^C	1	25	20
			0	20	66
$X_H = 1$	$X_A = 0$	Y_{i2}^C	1	16	24
			0	20	74
$X_H = 0$	$X_A = 1$	Y_{i2}^C	1	11	16
			0	18	45
$X_H = 0$	$X_A = 0$	Y_{i2}^C	1	17	33
			0	24	212

Tabelle 5: (Fall 5): X_A: Alter $>$ 100 Jahre bzw. $\leq$ 100 Jahre; X_H: Höhe $>$ 1200 m bzw. $\leq$ 1200 m

				Y_{i1}^E	
				1	0
$X_H = 1$	$X_A = 1$	Y_{i2}^E	1	93	21
			0	15	2
$X_H = 1$	$X_A = 0$	Y_{i2}^E	1	75	19
			0	25	15
$X_H = 0$	$X_A = 1$	Y_{i2}^E	1	79	1
			0	8	2
$X_H = 0$	$X_A = 0$	Y_{i2}^E	1	165	37
			0	38	46

Tabelle 6: (Fall 6): X_A: Alter $>$ 100 Jahre bzw. $\leq$ 100 Jahre; X_H: Höhe $>$ 1200 m bzw. $\leq$ 1200 m

				Y_{i1}^C	
				1	0
$X_B = 1$	$X_H = 1$	Y_{i2}^C	1	9	11
			0	6	32
$X_B = 1$	$X_H = 0$	Y_{i2}^C	1	9	14
			0	17	114
$X_B = 0$	$X_H = 1$	Y_{i2}^C	1	32	33
			0	34	108
$X_B = 0$	$X_H = 0$	Y_{i2}^C	1	19	35
			0	25	143

Tabelle 7: (Fall 7): X_H: Höhe $>$ 1200 m bzw. $\leq$ 1200 m; X_B: Beschirmungsgrad $>$ 60% bzw. $\leq$ 60%

				Y_{i1}^E	
				1	0
$X_B = 1$	$X_A = 1$	Y_{i2}^E	1	45	3
			0	3	1
$X_B = 1$	$X_A = 0$	Y_{i2}^E	1	98	18
			0	23	21
$X_B = 0$	$X_A = 1$	Y_{i2}^E	1	127	19
			0	20	3
$X_B = 0$	$X_A = 0$	Y_{i2}^E	1	142	38
			0	40	40

Tabelle 8: (Fall 8): X_A: Alter > 100 Jahre bzw. ≤ 100 Jahre; X_B: Beschirmungsgrad $> 60\%$ bzw. $\leq 60\%$

Für jeden der Fälle 5–8 wird ausgehend vom saturierten Modell mit Hilfe der Backward-Elimination das BE-Modell, d.h. ein den Daten am besten angepaßtes Modell bestimmt. Der kritische Wert für den p-value ist $\alpha = 0.025$ und für den G^2-Wert ist das jeweilige $\chi^2_{df;0.975}$-Fraktil (mit df = Zahl der Parameter im saturierten Modell − Zahl der Parameter im aktuellen Modell) ausschlaggebend für die Ablehnung des Modells. Die folgende Tabelle zeigt die BE-Modelle, deren G^2-Wert und den p-value für die Fälle 5–8 ($\ln \pi_{klmr} = P(Y_1^* = k, Y_2^* = l, X_* = m, X_{**} = r)$, $k, l, m, r = 0, 1$):

1	$\ln \pi_{klmr} = \mu + \lambda_k^{Y_1^C} + \lambda_l^{Y_2^C} + \lambda_m^{X_A} + \lambda_r^{X_H} +$ $+\lambda_{kl}^{Y_1^C Y_2^C} + \lambda_{lr}^{Y_2^C X_H} + \lambda_{km}^{Y_1^C X_A} + \lambda_{mr}^{X_A X_H} + \ln n$ $G^2 = 12.8359 < 16.00 = \chi^2_{7;0.975}$ p-value $= 0.0762$
2	$\ln \pi_{klmr} = \mu + \lambda_k^{Y_1^E} + \lambda_l^{Y_2^E} + \lambda_m^{X_A} + \lambda_r^{X_H} + \lambda_{kl}^{Y_1^E Y_2^E} + \lambda_{lr}^{Y_2^E X_H} + \lambda_{km}^{Y_1^E X_A}$ $+\lambda_{kr}^{Y_1^E X_H} + \lambda_{lm}^{Y_2^E X_A} + \lambda_{mr}^{X_A X_H} + \lambda_{klr}^{Y_1^E Y_2^E X_H} + \lambda_{kmr}^{Y_1^E X_A X_H} + \ln n$ $G^2 = 4.2887 < 9.35 = \chi^2_{3;0.975}$ p-value $= 0.2319$
3	$\ln \pi_{klmr} = \mu + \lambda_k^{Y_1^C} + \lambda_l^{Y_2^C} + \lambda_m^{X_A} + \lambda_r^{X_H} +$ $+\lambda_{kl}^{Y_1^C Y_2^C} + \lambda_{lm}^{Y_2^C X_A} + \lambda_{km}^{Y_1^C X_A} + \lambda_{mr}^{X_A X_H} + \ln n$ $G^2 = 6.9596 < 16.00 = \chi^2_{7;0.975}$ p-value $= 0.4331$
4	$\ln \pi_{klmr} = \mu + \lambda_k^{Y_1^E} + \lambda_l^{Y_2^E} + \lambda_m^{X_A} + \lambda_r^{X_B} +$ $+\lambda_{kl}^{Y_1^E Y_2^E} + \lambda_{lm}^{Y_2^E X_A} + \lambda_{km}^{Y_1^E X_A} + \lambda_{mr}^{X_A X_B} + \ln n$ $G^2 = 9.1608 < 16.00 = \chi^2_{7;0.975}$ p-value $= 0.2413$

Mit Ausnahme des Falles 6 ist das BE-Modell kleiner als das zur Schätzung verwendete Grundmodell

$$\begin{aligned}\ln \pi_{klmr} &= \mu + \lambda_k^{Y_1} + \lambda_l^{Y_2} + \lambda_m^{X_1} + \lambda_r^{X_2} + \lambda_{kl}^{Y_1 Y_2} + \\ &\quad + \lambda_{km}^{Y_1 X_1} + \lambda_{lm}^{Y_2 X_1} + \lambda_{kr}^{Y_1 X_2} + \lambda_{lr}^{Y_2 X_2} + \lambda_{mr}^{X_1 X_2} + \ln n \qquad (29)\end{aligned}$$

Die Überprüfung der Hypothesen

$$H_0^{\circ}: \quad \begin{pmatrix} \lambda_{111}^{Y_1 Y_2 X_1} \\ \lambda_{111}^{Y_1 Y_2 X_2} \\ \lambda_{111}^{Y_1 X_1 X_2} \\ \lambda_{111}^{Y_2 X_1 X_2} \\ \lambda_{1111}^{Y_1 Y_2 X_1 X_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (\text{kritischer Wert } \chi^2_{5;0.975} = 12.80)$$

und

$$H_0^{\circ\circ}: \quad \begin{pmatrix} \lambda_1^{Y_1} - \lambda_1^{Y_2} \\ \lambda_{11}^{Y_1 X_1} - \lambda_{11}^{Y_2 X_1} \\ \lambda_{11}^{Y_1 X_2} - \lambda_{11}^{Y_2 X_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad (\text{kritischer Wert } \chi^2_{3;0.975} = 9.35)$$

führt zu folgendem Ergebnis:

	G^2 Grundmodell (p-value)	H_0°	Wald-Statistik W (p-value)	$H_0^{\circ\circ}$
5	7.5013 (0.1859)	kann nicht abgelehnt werden	1.6820 (0.6410)	kann nicht abgelehnt werden
6	24.1825 (0.0002)	wird abgelehnt	— —	wird nicht mehr geprüft
7	4.2162 (0.5187)	kann nicht abgelehnt werden	0.7860 (0.8529)	kann nicht abgelehnt werden
8	6.2126 (0.2991)	kann nicht abgelehnt werden	0.5640 (0.9045)	kann nicht abgelehnt werden

Die standardisierten Parameterschätzungen zeigen, daß in allen geschilderten Fällen die Parameter $\lambda_k^{Y_1^C}, \lambda_l^{Y_2^C}, \lambda_{kl}^{Y_1^C Y_2^C}, \lambda_k^{Y_1^E}, \lambda_l^{Y_2^E}, \lambda_{kl}^{Y_1^E Y_2^E}$ signifikant sind. Mit Ausnahme folgender Parameter sind auch alle anderen λ-Parameter signifikant:

Fall 5	$\lambda_r^{X_H}, \lambda_{lm}^{Y_2^C X_A}$
Fall 6	$\lambda_r^{X_H}, \lambda_{kr}^{Y_1^E X_H}, \lambda_{lr}^{Y_2^E X_H}$
Fall 7	$\lambda_{kr}^{Y_1^C X_B}, \lambda_{lr}^{Y_2^C X_B}$
Fall 8	$\lambda_{kr}^{Y_1^E X_B}, \lambda_{lr}^{Y_2^E X_B}$

		Schätzung (Schätzung/Standardabweichung)				
		mit $\widehat{\lambda}_1^{Y_1^\bullet}$ $\widehat{\lambda}_{11}^{Y_1^\bullet X}$	mit $\widehat{\lambda}_1^{Y_2^\bullet}$ $\widehat{\lambda}_{11}^{Y_2^\bullet X}$	mit $\widehat{\lambda}_1^{Y_1^\bullet}+\widehat{\lambda}_1^{Y_2^\bullet}$ $\widehat{\lambda}_{11}^{Y_1^\bullet X}+\widehat{\lambda}_{11}^{Y_2^\bullet X}$	mit Alternativmethode $\widehat{\lambda}_1^{*}$ $\widehat{\lambda}_1^{**}$	G^2-Wert (p-value)
Fall 5	$\widehat{\alpha(0)}$	-0.2111 (-1.035)	-0.1626 (-0.804)	-0.1869 (-1.212)	-0.1816 (-1.179)	9.1920 (0.3264)
	$\widehat{\alpha(1)}$	0.9644 (5.148)	1.0128 (5.378)	0.9886 (7.419)	0.9847 (7.399)	
	$\widehat{\beta}_1$	0.6081 (2.984)	0.2950 (1.464)	0.4516 (3.615)	0.4495 (3.560)	
	$\widehat{\beta}_2$	0.3848 (1.902)	0.4062 (2.066)	0.3955 (3.203)	0.3958 (3.206)	
	$s(0,0)$	0.1525	0.1440	0.1482	0.1483	
	$s(0,1)$	0.0941	0.0856	0.0898	0.0898	
	$s(1,0)$	0.0691	0.0994	0.0832	0.0834	
	$s(1,1)$	0.0389	0.0563	0.0469	0.0470	
	OR	3.2404	3.2394	3.2340	3.2101	
Fall 7	$\widehat{\alpha(0)}$	0.1927 (0.753)	0.1628 (0.641)	0.1778 (0.952)	0.1809 (0.969)	5.0041 (0.7572)
	$\widehat{\alpha(1)}$	1.3999 (5.983)	1.3700 (6.955)	1.3849 (8.988)	1.3837 (8.983)	
	$\widehat{\beta}_1$	0.5027 (2.534)	0.4316 (2.227)	0.4672 (3.861)	0.4665 (3.856)	
	$\widehat{\beta}_2$	-0.1732 (-0.796)	-0.2566 (-1.206)	-0.2149 (-1.611)	-0.2158 (-1.620)	
	$s(0,0)$	0.0894	0.0931	0.0912	0.0912	
	$s(0,1)$	0.1123	0.1294	0.1206	0.1207	
	$s(1,0)$	0.0432	0.0504	0.0467	0.0467	
	$s(1,1)$	0.0561	0.07322	0.0642	0.0643	
	OR	3.3441	3.3441	3.3438	3.3294	
Fall 8	$\widehat{\alpha(0)}$	-2.4631 (-8.832)	-2.3074 (-8.549)	-2.3853 (-13.494)	2.3881 (-13.491)	6.6396 (0.5760)
	$\widehat{\alpha(1)}$	-1.2279 (-3.840)	-1.0722 (-3.380)	-1.1500 (-4.802)	-1.1513 (-4.809)	
	$\widehat{\beta}_1$	0.8985 (3.666)	0.9381 (3.902)	0.9183 (5.896)	0.9188 (5.900)	
	$\widehat{\beta}_2$	0.2891 (1.331)	0.1587 (0.751)	0.2239 (1.670)	0.2226 (1.692)	
	$s(0,0)$	0.7127	0.6776	0.6955	0.6956	
	$s(0,1)$	0.6455	0.6392	0.6432	0.6427	
	$s(1,0)$	0.4810	0.4253	0.4532	0.4531	
	$s(1,1)$	0.3987	0.3805	0.3896	0.3899	
	OR	3.4391	3.4391	3.4394	3.4274	

Tabelle 9: Parameterschätzungen der Fälle 5,7 und 8

Die Schätzungen $\widehat{\alpha(0)}$, $\widehat{\alpha(1)}$, $\widehat{\beta}_1$ und $\widehat{\beta}_2$ können mit Ausnahme von Fall 6 für alle Fälle berechnet werden, da die Restriktionen für das SRM immer erfüllt werden. Für den Fall 6 wird das Grundmodell bei einem Signifikanzniveau von 0.025 abgelehnt. Die Anwendung des SRM auf diese Kontingenztafel ist aufgrund der Signifikanz des Dreifachinteraktionsparameters ($\widehat{\lambda}_{111}^{Y_1^E X_A X_H}$) nicht möglich, da die Hypothese H_0° in der die Nullrestriktionen zusammengefaßt sind, abgelehnt wird. Die nachstehende Tabelle zeigt für alle Fälle mit Ausnahme des Falles 6 die Resultate der Parameterschätzungen, die sich bei Verwendung der verschiedenen λ-Parameter ergeben ($Y^\bullet$ steht für Y^C bzw. Y^E).

Die Ergebnisse der Analyse der Wirkung der Kovariablen auf die Responsevariablen unter Berücksichtigung derer Korrelation zeigen, daß das Alter und die Höhe signifikant für beide Responsevariablen sind und dem Beschirmungsgrad jeweils keine Bedeutung zukommt. Die Odds-Ratios sind im Vergleich zum Zahndatensatz relativ klein und deuten auf eine nur schwache positive Korrelation der Responsevariablen hin. Auch der gegenseitige Einfluß der Zielvariablen aufeinander, der durch den Parameter $k = \ln \mathrm{OR}$ charakterisiert wird, ist geringer, da der Wert von k immer nahe bei 1 liegt.

6 Zusammenfassung und Ausblick

Das hier definierte symmetrische Regressionsmodell (SRM) bietet sicher nicht die einzige Möglichkeit der Bearbeitung korrelierter Größen, aber in Verbindung mit dem dafür entwickelten Schätzverfahren ist es eine grundlegend neue Methode, die im Gegensatz zu anderen Modellen und Schätzverfahren mit relativ einfachen Mitteln und ohne übermäßig großen Rechenaufwand durchführbar ist. Das SRM berücksichtigt im Gegensatz zu manchen anderen Regressionsmodellen mehrere Richtungen bei der Modellierung der Zusammenhangsstruktur der vorliegenden Daten. Es bietet die Möglichkeit, sowohl die Zusammenhänge zwischen den Responsevariablen, als auch diejenigen zwischen den Responsevariablen und den Einflußgrößen in sinnvoller und interpretierbarer Form darzustellen. Die zusätzliche Verwendung loglinearer Modelle bei der Schätzung der Parameter des SRM bringt weitere Vorteile mit sich: Der Informationsverlust bei der Modellbildung wird reduziert, da loglineare Modelle wesentlich sensibler bei der Analyse von Zusammenhängen sind und alle in den Daten vorhandene Informationen dazu nutzen. Durch die primäre Analyse der Daten in Form einer Backward-Elimination der loglinearen Modelle erhält der Anwender einen grundlegenden Überblick über die tatsächlich vorliegenden Zusammenhangsstrukturen. Das Problem loglinearer Modelle, daß die Parameter nicht explizit interpretierbar sind, weswegen loglineare Modelle für die Regressionsanalyse eher ungeeignet sind, wird durch die Kombination mit dem SRM gelöst. Gleichzeitig wird dadurch vermieden, daß für die Schätzung des SRM subjektive

Annahmen, wie beispielsweise die Annahme einer Basisverteilung der Responsevariablen, verwendet werden. In die Schätzung gehen nur die tatsächlich in den Daten vorhandenen Informationen ein. Die Folge ist eine größere Flexibilität und Sicherheit bei der Anwendung des SRM, da mit den Ergebnissen aus den loglinearen Modellen die für das SRM notwendigen Voraussetzungen und Annahmen überprüft werden können und somit die unzutreffende Verwendung des SRM vermieden wird, was bei allen anderen Modellen nicht möglich ist. Fehlinterpretationen der Parameter und falsche Schlußfolgerungen werden verhindert.

Natürlich gibt es in Bezug auf das Schätzverfahren noch offene Fragen. Dazu gehört z.B. die Erweiterung der Methode auf komplexere Datenstrukturen, d.h. auf Datensätze, die mehr als zwei Einflußgrößen enthalten. Auch die Anwendung des SRM auf mehrkategoriale Variablen und auf das Problem fehlender Daten stellt ein weiteres, noch zu bearbeitendes Thema dar.

Literatur

[1] A. Agresti (1990): *Categorical Data Analysis.* Wiley, New-York

[2] Bayerische Landesanstalt für Wald- und Forstwirtschaft (1992): *Der Kronenzustand der Fichten am Wank bei Garmisch-Partenkirchen 1986-1992.*

[3] G. E. Bonney (1987): *Logistic regression for dependent binary observations.* Biometrics, 43, 951–973

[4] M. A. Connolly and K.-Y. Liang (1988): *Conditional logistic regression models for correlated binary data.* Biometrika, 75, 501–506

[5] L. Fahrmeir and H. Kaufmann (1987): *Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models.* Jour. of Time Series Analysis, 8, 147-160

[6] L. Fahrmeir and G. Tutz (1994): *Multivariate statistical modelling based on generalized linear models.* Springer, New York

[7] Christian Heumann: *Loggy 1*, Handbuch, Institut für Statistik, Universität München, 1993

[8] K.-Y. Liang, S. L. Zeger and B. F. Qaqish (1992): *Multivariate regression analyses for categorical data.* Jour. Royal Statist. Soc. B, 54, Nr. 1, 3–40

[9] K.-Y. Liang and S. L. Zeger (1993): *Regression analyses for correlated data.* Annu. Rev. Pub. Health 1993, 14, 43–68

[10] R. L. Prentice (1988): *Correlated binary regression with covariates specific to each binary observation.* Biometrics, 44, 1033–1048

[11] Y. S. Qu, G. W. Williams, G. J. Beck and M. Goormastic (1987): *A generalized model of logistic regression for correlated data.* Communications in Statistics – A, 16, 3447–3477

[12] B. Rosner (1982): *Statistical methods in ophthalmology: An adjustment for the intra-class correlation between eyes.* Biometrics, 38, 105–114

[13] B. Rosner (1984): *Multivariate methods in ophthalmology with application to other paired-data situations.* Biometrics, 40, 961–971

[14] B. Rosner (1989): *Multivariate methods for clustered binary data with more than one level of nesting.* Jour. Amer. Statist. Ass., 84, 373–380

[15] B. Rosner and T. D. Tosteson (1990): *Response to the paper by Neuhaus and Jewell.* Biometrics, 46, 531–534

[16] B. Rosner (1992): *Multivariate methods for clustered binary data with multiple subclasses, with application to binary longitudinal data.* Biometrics, 48, 721–731

[17] W. Walther (1992): *Ein Modell zur Erfassung und statistischen Bewertung klinischer Therapieverfahren – entwickelt durch Evaluation des Pfeilerverlusts bei Konuskronenersatz.* Habilitationsschrift, Medizinische Fakultät der Universität des Saarlandes, 1992