Automatische Modellselektion bei der loglinearen Modellierung in SAS

Berit Kalina
IMBE Universität
Erlangen-Nürnberg
berit.kalina@imbe.imed.uni-erlangen.de

Annette Pfahlberg
IMBE Universität
Erlangen-Nürnberg
annette.pfahlberg@rzmail.uni-erlangen.de

Kira Klenke
Fachhochschule
Hannover
kira.klenke@ik.fh-hannover.de

Olaf Gefeller
IMBE Universität
Erlangen-Nürnberg
olaf.gefeller@rzmail.uni-erlangen.de

Zusammenfassung

Im Rahmen der Erlanger Kindergartenstudie (ErlKing) wurde im Herbst 2001 in Kindergärten der Stadt Erlangen sowie im Landkreis Erlangen-Höchstadt eine Befragung der Eltern von Kleinkindern durchgeführt. Dabei wurde der Kenntnisstand hinsichtlich der bekannten Melanom-risikofaktoren, sowie das präventionsrelevante Schutzverhalten und nicht zuletzt die Konsequenzen für die Kinder (Auftreten von Sonnenbränden bei den Kindern als Ausdruck einer überschwelligen Schädigung) erfragt. Die Daten aus der Befragung wurden mit dem Ziel erhoben, aussagekräftige Ergebnisse für die Region Erlangen und Erlangen-Höchstadt zu gewinnen, um Basisdaten für zukünftige Informations- und Aufklärungskampagnen zu erarbeiten. Eine Möglichkeit die Assoziation zwischen den interessierenden Variablen aufzudecken, ist die Untersuchung von Zusammenhängen zwischen den kategorialen Variablen mittels der loglinearen Modellierung.
Für die loglineare Modellierung der Assoziationsstrukturen kategorialer Variablen steht in SAS die Prozedur CATMOD zur Verfügung. Die konkrete statistische Analyse, welches Modell die in den Daten enthaltenen Assoziationsstrukturen hinreichend gut beschreiben könnte, bedeutet allerdings mit der CATMOD Prozedur eine aufwendige und zeitintensive Suche, da keine automatischen Selektionsalgorithmen zur Verfügung stehen. Um die Suche nach einem geeigneten Modell zu vereinfachen, wurde für den Prozess der schrittweisen Modell-vereinfachung ein SAS Macro (*amelli*) geschrieben. Die Aufgabe dieses Macros besteht darin, die sukzessive Reduzierung von nichtsignifikanten Interaktionen und Haupteffekten unter Berücksichtigung des Hierarchie-prinzips aus einem saturierten loglinearen Modell vorzunehmen. Diese Modellselektion soll anhand der Erlanger Kindergartenstudie illustriert werden.

Keywords: malignes Melanom, Erlanger Kindergartenstudie, SAS, loglineare Modellierung, Modellselektion, Macro, CATMOD.

1 Malignes Melanom

In der Bundesrepublik Deutschland beträgt die Neuerkrankungsrate des malignen Melanoms derzeit etwa 10 bis 12 pro 100000 Einwohner pro Jahr. Als Risikofaktoren für die Entwicklung des malignen Melanoms gelten sowohl eine große Anzahl der unauffälligen als auch der atypischen pigmentierten Naevi (Leberflecken), ein heller Hauttyp, aber auch eine genetische Disposition (familiäres Auftreten). Die Naevi stellen dabei einen besonders großen Risikofaktor dar, da fast 50% aller Melanome sich aus pigmentierten Naevi entwickeln. Obwohl keine Dosis-Wirkungs-Beziehung definiert werden konnte, scheint starke Sonnenexposition, insbesondere in der Kindheit, die Entwicklung der pigmentierten Naevi als auch deren maligne Transformation zu fördern. Da bis zu 80% der individuellen Sonnenlichtbelastung vor dem 18. Lebensjahr erreicht wird, sollte besonders die kind- und jugendliche Haut vor übermäßiger UV-Exposition und Sonnenbränden bewahrt werden [4].

Das maligne Melanom ist ein Tumor, dem mit Prävention erfolgreich begegnet werden kann. Hierbei kann zwischen zwei Präventionskonzepten unterschieden werden. Unter der primären Prävention wird die Vermeidung des Entstehens maligner Melanome verstanden. Dies kann im wesentlichen durch eine gesundheitliche Aufklärung der Bevölkerung erreicht werden. Das Ziel solch einer Aufklärung ist, den Wissensstand der Bevölkerung über die Risikofaktoren des malignen Melanoms zu vermehren und somit die ungeschützte Sonnenexposition zu reduzieren. Die zweite Präventionsmaßnahme, die sekundäre Prävention, hat als Ziel, eine möglichst frühe Diagnosestellung zu ermöglichen. Hiermit soll die Metastasierung und Tumorprogression vorzeitig verhindert werden [1].

2 Erlanger Kindergartenstudie

Bezug nehmend auf die primäre Prävention wurde vom Institut für Medizininformatik, Biometrie und Epidemiologie der Friedrich-Alexander-Universität Erlangen-Nürnberg, in Zusammenarbeit mit dem Gesundheitsamt in Erlangen sowie dem Tumorzentrum Erlangen-Nürnberg, eine Befragung von Eltern in ausgewählten Kindergärten in der Stadt Erlangen sowie im Landkreis Erlangen-Höchstadt initiiert.

Im Herbst 2001 wurde eine Befragung von Eltern zu ihrem Wissen über die Gefährdung durch UV-Strahlung, dem Verhalten ihrer Kinder beim Aufenthalt im Freien und den möglicherweise aufgetretenen Hautrötungen und Sonnenbränden bei ihren Kindern im vergangenen Sommer durchgeführt. In der Erlanger Kindergartenstudie nahmen 59 durch eine gewichtete Zufallsstichprobe ausgewählte Kindergärten (4146 Elternpaare) teil. Mit Hilfe eines selbst auszufüllenden, standardisierten Erhebungsbogens wurden alle Eltern der ausgewählten Kindergärten befragt. Bei mehr als einem Kindergartenkind pro Familie sollte nur für das älteste Kind der Fragebogen ausgefüllt werden.

Ziel dieser Befragung war es, aussagekräftige Ergebnisse für die Region Erlangen und Erlangen-Höchstadt über den Umgang der Eltern und ihrer Kinder mit der Sonne zu gewinnen, um daraus Basisdaten für zukünftige Informations- und Aufklärungskampagnen zu erarbeiten.

Die Rücklaufquote, die bei dieser Querschnittsstudie erzielt wurde, liegt bei 64,69%. Das heißt, von 4146 Fragebögen, welche an die 59 Kindergärten verteilt wurden, kamen 2682 Fragebögen zurück. Davon standen 2667 Fragebögen für die Auswertung zur Verfügung.

3 Loglineare Modellierung

Im Fragebogen wurde zunächst versucht, den Grundhauttyp des Kindes zu bestimmen. Diese Zuordnung zu einem der vier Hauttypen ist dahingehend wichtig, da eine lichtempfindlichere Haut ein erhöhtes Melanomrisiko darstellt. Lichtempfindlichkeit der Haut ist vor allem bei Menschen mit roten oder blonden Haaren, hellen Augen und einer hohen Sommersprossenanzahl anzutreffen. Mit dieser Thematik wurde die Einschätzung der Eltern und Kinder gegenüber der Sonnenempfindlichkeit (d.h. wie viel Sonne die Haut verträgt, ohne rot zu werden) erfasst.

Anhand der gewonnenen Daten zur Lichtempfindlichkeit der Haut konnten nun Zusammenhänge zwischen den kategorialen Variablen mittels der loglinearen Modellierung untersucht werden.

In die Auswertung gingen die folgenden fünf Variablen ein:

Label	**Variablen**	**Ausprägungen**	
A = Geschlecht des Kindes	ges	i=1,2	weiblich männlich
B = Haarfarbe des Kindes	haar	j=1,2	rot oder blond braun oder schwarz
C = Augenfarbe des Kindes	auge	k=1,2,3	blau oder blaugrau grün oder grünbraun braun oder braunschwarz
D = Sommersprossen im Gesicht	sp_g	l=1,2,3	0 10 ≥ 20
E = Sommersprossen am Arm	sp_a	m=1,2	0 ≥ 10

In der zweiten Spalte befinden sich die innerhalb von SAS verwendeten Variablenbezeichnungen und die Buchstaben in der dritten Spalte zeigen die Anzahl der Ausprägungen pro Variable.

Das aus diesen fünf Variablen entstandene saturierte loglineare Modell sieht wie folgt aus:

$$
\begin{aligned}
\ln m_{ijklm} &= u \\
&+ u_{A(i)} + u_{B(j)} + u_{C(k)} + u_{D(l)} + u_{E(m)} && \textbf{Haupteffekte} \\
&+ u_{AB(ij)} + u_{AC(ik)} + u_{AD(il)} + u_{AE(im)} && \textbf{Interaktion 1. Ordnung} \\
&+ u_{BC(jk)} + u_{BD(jl)} + u_{BE(jm)} \\
&+ u_{CD(kl)} + u_{CE(km)} + u_{DE(lm)} \\
&+ u_{ABC(ijk)} + u_{ABD(ijl)} + u_{ABE(ijm)} && \textbf{Interaktion 2. Ordnung} \\
&+ u_{ACD(ikl)} + u_{ACE(ikm)} + u_{ADE(ilm)} \\
&+ u_{BCD(jkl)} + u_{BCE(jkm)} + u_{BDE(jlm)} \\
&+ u_{CDE(klm)} \\
&+ u_{ABCD(ijkl)} + u_{ABCE(ijkm)} && \textbf{Interaktion 3. Ordnung} \\
&+ u_{ABDE(ijlm)} + u_{ACDE(iklm)} \\
&+ u_{BCDE(jklm)} \\
&+ u_{ABCDE(ijklm)} && \textbf{Interaktion 4. Ordnung}
\end{aligned}
$$

Aufgezeigt ist hier jede Hierarchieebene des saturierten Modells, angefangen mit den fünf Haupteffekten, bei der Annahme, dass A, B, C, D und E für die fünf Variablen stehen. Weiterhin unterteilt sich das saturierte Modell in vier Interaktionsordnungen, in denen die Assoziation zwischen den verschiedenen Variablen aufgezeigt ist. Je nach Anzahl der involvierten Variablen, wird zwischen Interaktion 1. Ordnung bis zur Interaktion 4. Ordnung unterschieden.

Das Ziel der loglinearen Modellierung ist die Identifikation eines Modells, das die Variationen der Zellhäufigkeiten möglichst einfach und gut erklärt. Um ein angemessenes Modell zu finden, stehen verschiedene Selektionsalgorithmen zur Verfügung.

In der loglinearen Modellierung im Zusammenhang mit der Erlanger Kindergartenstudie wurde das „backward elimination"- Verfahren nach Goodman zugrunde gelegt [2].

3.1 „backward elimination" – Verfahren

Das Grundprinzip des „backward elimination"- Verfahrens nach Goodman ist, aus einem komplexeren Basismodell ein einfaches Modell entstehen zu lassen. Das heißt, hier werden die nichtsignifikanten Interaktionseffekte, sowie die Haupteffekte unter Berücksichtigung des Hierarchieprinzips, schrittweise aus dem Basismodell herausgenommen. Unter Berücksichtigung des Hierarchieprinzips bedeutet dabei, dass der Prozess der schrittweisen Modellvereinfachung mit der Interaktion höchster Ordnung, also mit der Interaktion 4. Ordnung beginnt und sich später auf die Interaktionen 3. Ordnung, hiernach auf die Interaktionen 2. Ordnung sowie auf die Interaktionen 1. Ordnung und als letztes auf die Haupteffekte ausweitet.

Dabei wird bei jeder Interaktionsordnung überprüft, ob die Modellparameter der jeweiligen Ordnung einen signifikanten Beitrag zur Beschreibung der Kontingenztafel leisten können. Für diese Überprüfung muss vorher ein Signifikanzniveau zur Eliminierung von Modellparametern festgelegt werden. Sollte sich innerhalb dieser Überprüfung herausstellen, dass kein signifikanter Beitrag besteht, so werden die nichtsignifikanten Modellparameter der gerade zu untersuchenden Interaktionsordnung miteinander verglichen und die Interaktion mit dem höchsten über α liegenden p-Wert würde aus dem Modell entfernt werden, wenn sie in keiner signifikanten Interaktion höherer Ordnung beinhaltet ist.
Nach der Entfernung eines Modellparameters aus dem aktuellen loglinearen Modell wird ein neues Modell erzeugt, und die Überprüfung der Modellparameter einer Ordnung beginnt von neuem.

Der Prozess zur Reduktion von Modellparametern ist erst beendet, wenn nur noch signifikante Parameter aller Ordnungen und nichtsignifikante Parameter mit Beinhaltung in Interaktionen höherer Ordnung in dem Modell enthalten sind und alle nichtsignifikanten Modellparameter ohne eine Beinhaltung in Interaktionen höherer Ordnung entfernt wurden.

4 Implementierung in SAS

Um das „backward elimination"- Verfahren auch unter SAS anwenden zu können, wurden die fünf Variablen aus der Erlanger Kindergartenstudie in die SAS Prozedur CATMOD integriert, wobei das Basismodell ein saturiertes loglineares Modell war [5].

```
PROC CATMOD;
MODEL ges*haar*auge*sp_g*sp_a = _RESPONSE_ / PRED=FREQ
                                             NORESPONSE
                                             NOITER NOPARM;
LOGLIN ges|haar|auge|sp_g|sp_a;
RUN;
```

Bei der Modellierung der loglinearen Regression mittels der CATMOD Prozedur wurden auch zusätzliche Optionen im MODEL-Statement angegeben, mit denen verschiedene Tabellen im Output unterdrückt werden sollten, da diese Tabellen nur eine untergeordnete Bedeutung im Rahmen der loglinearen Modellierung haben. Wie zum Beispiel die Option NORESPONSE, welche die Tabelle „_Response_Matrix" unterdrückte.

Bereits während der Aufnahme der fünf Variablen zeigte sich, dass die Analyse, welches loglineare Modell die in den Daten enthaltenen Assoziationsstrukturen hinreichend gut beschreiben könnte, mit der CATMOD Prozedur eine aufwendige und zeitintensive Suche nach einem angemessenen Modell ist. Das beruht vor allem auf der mehrmaligen Ausführung der CATMOD Prozedur und der damit verbundenen individuellen Reduktion der Modellparameter. Ein Lösungsvorschlag der daraufhin entstanden ist, war es, automatische Variablen- bzw. Modellselektionsalgorithmen in SAS zu implementieren. Hieraus motiviert das Macro *amelli* in der SAS Version 8.2, das die Aufgabe hat, die sukzessive Reduzierung von nichtsignifikanten Interaktionen und Haupteffekten unter Berücksichtigung des Hierarchieprinzips aus einem saturierten loglinearen Modell vorzunehmen.

5 Macro amelli

Der Aufruf des Macros *amelli* lautet wie folgt:

```
%amelli(data,alpha,var);
```

Es besteht aus drei Elementen, wobei jedes Element ein Pflichtbestandteil ist. Das erste Element gibt die Datei an, in welcher die Daten gespeichert sind, das zweite gibt das Signifikanzniveau für den Eliminationsschritt in der „backward elimination“ an und der letzte Bestandteil des Aufrufs umfasst die Variablen, welche in das Grundmodell aufgenommen werden sollen, wobei sie von einem Leerzeichen getrennt angegeben werden müssen [6].

Der Aufruf des Macros *amelli* für die Erlanger Kindergartenstudie sieht wie folgt aus:

```
%amelli(ErlKing_Daten,0.10,ges haar auge sp_g sp_a);
```

An einer Illustration wird die Modellselektion des SAS Macros *amelli* für den Haupteffekt *Geschlecht*, ausgehend von einem bereits vereinfachten loglinearen Modell, dargestellt.

Step 23

Source	DF	ChiSq	Prob ChiSq
ges	1	0.07	0.7901
haar	1	109.89	<.0001
auge	2	169.47	<.0001
haar*auge	2	346.45	<.0001
sp_g	2	3.85	0.1459
auge*sp_g	4	12.91	0.0117
sp_a	1	482.93	<.0001
sp_g*sp_a	2	338.36	<.0001
Likelihood Ratio	56	44.06	0.8761

Bei diesem Modell ist zu erkennen, dass die Interaktion höchster Ordnung (Interaktion 4. Ordnung), sowie die Interaktionen der 3. und 2. Ordnung aus dem Modell entfernt wurden und nur noch einige wenige Interaktionen 1. Ordnung sowie die Haupteffekte im Modell verblieben sind.

In diesem vereinfachten Modell zeigt sich, dass die Haupteffekte *Geschlecht* und *Sommersprossen im Gesicht* (*ges* und *sp_g*) beide über dem 10% Signifikanzniveau liegen. Da sie also beide keinen signifikanten Beitrag zur Beschreibung der Kontingenztafel leisten können, müssen diese beiden Modellparameter hinsichtlich des höchsten über α liegenden p-Wertes

miteinander verglichen werden. Bereits ein Blick in den Output lässt erkennen, dass der Haupteffekt *Geschlecht* einen weit höheren p-Wert besitzt als *Sommersprossen im Gesicht*, und da er in keiner signifikanten Interaktion höherer Ordnung beinhaltet ist, kann dieser Modellparameter aus dem Modell entfernt werden.

Welcher Modellparameter bei der Modellselektion aus dem aktuellen Modell entfernt wurde, wird vom Macro in einem zusätzlichen Output angezeigt.

Folgender Modellparameter wird aus dem loglinearen Modell entfernt:

Source	DF	ChiSq	Prob ChiSq2
ges	1	0.07	0.7901

Das aus dieser Modellselektion entstehende aktuelle loglineare Modell, welches gleichzeitig das Endmodell der Erlanger Kindergartenstudie ist, da keine weiteren Parameter aus dem Modell entfernt werden können, sieht wie folgt aus:

Step 24

Source	DF	ChiSq	Prob ChiSq
haar	1	109.89	<.0001
auge	2	169.47	<.0001
haar*auge	2	346.45	<.0001
sp_g	2	3.85	0.1459
auge*sp_g	4	12.91	0.0117
sp_a	1	482.93	<.0001
sp_g*sp_a	2	338.36	<.0001
Likelihood Ratio	57	44.13	0.8937

In diesem Modell zeigt sich, dass der Haupteffekt *Sommersprossen im Gesicht* (*sp_g*) immer noch keinen signifikanten Beitrag zur Beschreibung der Kontingenztafel leisten kann, da er aber in zwei Interaktionen 1. Ordnung beinhaltet ist und diese beiden signifikant sind, bleibt dieser Haupteffekt im Modell enthalten.

Erkennbar aus diesem Modell ist, dass in der Erlanger Kindergartenstudie mittels des SAS Macros *amelli* ein Zusammenhang zwischen den Variablen

- *Haarfarbe* und *Augenfarbe* (*haar*auge*),
- *Augenfarbe* und *Sommersprossen im Gesicht* (*auge*sp_g*) und

- *Sommersprossen im Gesicht* und *Sommersprossen am Arm* (*sp_g*sp_a*)

festgestellt werden konnte.

Das Macro ist, für nichtkommerzielle Anwendungen kostenlos, mit einer genauen Beschreibung der Funktionsweise unter folgender Web-Adresse herunterzuladen:
http://www.imbe.med.uni-erlangen.de/issan/SAS/amelli/amelli.htm.

Diese Web-Seite wird kontinuierlich gepflegt, so dass dort auch eine aktuelle Fassung des Macros zu finden sein wird.

Literatur

1. Blum, A., Garbe, C. and Rassner, G. (1998). Prevention of malignant melanoma. *Hautarzt*, 49, 826-834.

2. Goodman, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrisc*, 13, 33-61.

3. Kalina, B. (2002). Automatische Modellselektion bei der loglinearen Modellierung in SAS: Praktische Anwendung in der Erlanger Kindergartenstudie. Diplomarbeit. Fachhochschule Hannover.

4. Kölmel, K.F., Pfahlberg, A. and Gefeller, O. (1997). Prevention of melanoma by sun protective measures in childhood. Temporal changes in awareness of parents. *Hautarzt*, 48, 391-396.

5. SAS Institute Inc. (1999). *SAS/STAT Users Guide, Version 8, First Edition*. Cary, NC: SAS Institute Inc.

6. SAS Institute Inc. (1999). *SAS Macro Language: Reference, Version 8, First Edition*. Cary, NC: SAS Institute Inc.