Die Anwendung des SAS-Makros GLIMMIX bei der statistischen Auswertung von Langzeitstudien in der Ernährungsmedizin am Beispiel der DONALD-Studie

V. Schultze-Pawlitschko, M. Kersting
Forschungsinstitut für Kinderernährung
Heinstück 11
44225 Dortmund
pawl@fke-do.de
kersting@fke-do.de

Zusammenfassung

Im Rahmen der DONALD-Studie werden seit 1985 Kinder im Alter von 3 Monaten bis 18 Jahren zu bestimmten Zeitpunkten ärztlich untersucht und befragt. Zusätzlich werden an drei aufeinanderfolgenden Tagen Verzehrsprotokolle erstellt. Die Daten aus den Untersuchungen, Befragungen und Protokollen werden mit dem Ziel erhoben, langfristige Trends in den Lebens- und Ernährungsgewohnheiten von Kindern zu untersuchen.
Bei der statistischen Modellbildung treten hierbei zwei Besonderheiten auf. Zum einen erfolgen die Messungen bei den Kindern zu nicht äquidistanten Zeitpunkten, zum anderen existieren stochastisch abhängige Messverläufe (z. B. bei Geschwisterkindern in der Studie).
Im Falle stetiger Zielgrößen sind zur statistischen Modellbildung Gemischte Lineare Modelle geeignet, zu deren Analyse eine geeignete SAS-Prozedur (PROC MIXED) zur Verfügung steht. Im Fall diskreter Zielgrößen sieht die Situation anders aus. Für die hier einzusetzenden Generalisierten Gemischten Linearen Modelle gibt es – soweit bekannt – keine SAS-Prozeduren mit denen diese korrekt behandelt werden können.
Daher muss auf ein SAS-Makro (GLIMMIX) zurückgegriffen werden. Die Möglichkeiten und Grenzen dieses Makros werden erläutert und anhand zweier Beispieldatensätze aus der DONALD-Studie illustriert.

Keywords: Längsschnittstudie, Trends, Generalisierte Gemischte Lineare Modelle.

1 Die DONALD-Studie

In der DONALD-Studie werden seit 1985 gesunde Säuglinge, Kinder und Jugendliche im Alter von 3 Monaten bis 18 Jahre zu 29 Zeitpunkten (siehe Abbildung 1) untersucht. Der Studienname „DONALD“ steht hier für „**DO**rtmunder **N**utritional and **A**nthropometric **L**ongitudinally **D**esigned Study“. Ein Termin besteht in der Regel aus einem Interview, einer medizinischen und einer anthropometrischen Untersuchung. Im Anschluss an die Untersuchung wird zu Hause ein Drei-Tage-Ernährungsprotokoll erstellt. Bei Probanden ab einem Alter von rund drei Jahren wird am dritten Protokolltag ein 24-Stunden-Urin gesammelt.

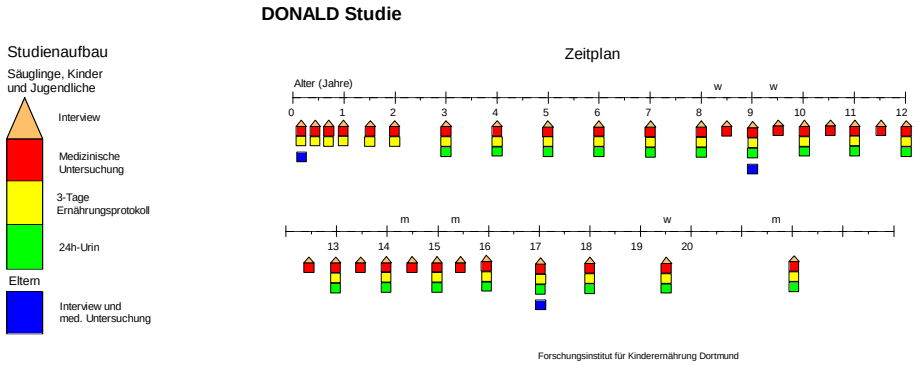

Abbildung 1: Studiendesign der DONALD-Studie: Komponenten der Untersuchung bei Säuglingen, Kindern, Jugendlichen und Eltern sowie Zeitplan der altersabhängigen Untersuchungen

Ziel der Untersuchungen, der Befragungen und der Protokolle ist es unter anderem, langfristige Trends in den Lebens- und Ernährungsgewohnheiten von Kindern zu erkennen.

2 Statistische Problemstellung

Das klassische statistische Modell zur Analyse von Einflussfaktoren auf eine Zielvariable ist das allgemeine Lineare Modell mit festen Effekten. In dieser Studie treten jedoch zwei Probleme auf, die es unmöglich machen, mit den

klassischen Verfahren zu arbeiten, da die einzelnen Beobachtungen voneinander abhängig sein können.

Zum einen erfolgen Wiederholungsmessungen zu nicht äquidistanten Zeitpunkten, zum anderen existieren stochastisch abhängige Messverläufe, da häufig mehrere Kinder einer Familie an der Studie teilnehmen. Wird jeder Familie ein (zufälliger) Familieneffekt und jedem Kind innerhalb der Familie ein (zufälliger) Kindeffekt zugeordnet, so ergibt sich ein Gemischtes Lineares Modell mit festen und zufälligen Effekten. Diese Modelle können mit der SAS-Prozedur PROC MIXED befriedigend umgesetzt werden, sofern die Beobachtungen als normalverteilt angesehen werden können. Das Problem bei vielen Trendauswertungen in der DONALD-Studie ist jedoch, dass die Zielvariable y kein normalverteiltes Merkmal, in vielen Fällen noch nicht einmal stetig ist. In diesen Fällen kann mit PROC MIXED nicht gearbeitet werden. Zwei Beispiele mit nicht stetigen Zielvariablen werden im folgenden betrachtet.

3 Formulierung eines adäquaten Modellansatzes für binäre Zielvariablen am Beispiel der Zufuhr von Kalzium

Im Bereich der Ernährungsmedizin wird das Problem der sogenannten Nährstoffverdünnung seit einigen Jahren heftig diskutiert. Damit ist die Frage gemeint, ob ein Proband obwohl er mehr isst, das heißt mehr Energie/Kalorien aufnimmt, möglicherweise relativ weniger Nährstoffe erhält, weil er zum Beispiel Nahrungsmittel mit „leeren Kalorien“ wie etwa Zucker bevorzugt. Aufgrund dieser Beobachtung wurde in der DONALD-Studie die Frage untersucht, ob es eine Abhängigkeit zwischen dem Zuckerverzehr und der Einhaltung der Empfehlungen für die Aufnahme von Kalzium gibt. Des weiteren wurde analysiert, ob es bezüglich dieser Einhaltung Trends über die Zeit hinweg gibt und ob das Geschlecht einen Einfluss hat. Gesucht wird also ein Modell, das die Wahrscheinlichkeit p genug Kalzium aufgenommen zu haben, in Abhängigkeit vom Zuckerkonsum (P_ZUCK), vom Untersuchungsjahr (JAHR) und vom Geschlecht (SEX) beschreibt. P_ZUCK steht hierbei für den prozentualen Anteil des Zuckers an der Gesamtenergiezufuhr.

Die Lösung für dieses Problem ist der Einsatz eines Generalisierten Linearen Gemischten Modells. „Generalisiert“ heißt hier, dass nicht p selbst zur Modellbildung herangezogen, sondern mit Linkfunktionen gearbeitet wird. Eine bei binären Zielgrößen standardmäßig verwandte Linkfunktion ist die Logit-Funktion, die folgendermaßen definiert ist:

$$\text{logit}(p) = \ln\left[\frac{p}{1-p}\right] \ .$$

Die Logit-Funktion logit: $(0,1) \to \mathbb{R}$ ist eine monoton wachsende stetige Funktion, die das Intervall $(0,1)$ in den Bereich der gesamten reellen Zahlen abbildet. Die Funktion ln: $(0,\infty) \to \mathbb{R}$ ist die natürliche Logarithmusfunktion.

Der Einsatz von Linkfunktionen (in diesem Fall der Logit-Funktion) führt zu folgender Modellbildung:

$$\text{logit}(p) = ax + bt + cz. \tag{3.1}$$

Dabei sind x, t (t gemessen in Jahren) und z die bereits erwähnten Einflussfaktoren, das heißt a beschreibt den Zucker-, b sei der Zeit- und c der Geschlechtseffekt.

4 Formulierung eines adäquaten Modellansatzes für Zählvariablen am Beispiel des Mahlzeitenverhaltens

Eine weitere in der DONALD-Studie interessierende Zielvariable ist die Anzahl der von einem Kind pro Tag gegessenen Mahlzeiten. Es soll untersucht werden, ob die „klassische Mahlzeitenstruktur“ mit drei Haupt- und zwei Zwischenmahlzeiten heute noch Gültigkeit hat oder durch mehrere kleine Mahlzeiten mit „snacks“ oder „fastfood“ abgelöst worden ist.

Die Anzahl der Mahlzeiten soll im folgenden Modell in Abhängigkeit vom Alter des Kindes, vom Untersuchungsjahr und vom Geschlecht dargestellt werden. Die Zielgröße kann als Zählvariable angesehen werden, ist also ebenfalls kein stetiges Merkmal. Deshalb wird auch hier wieder mit Linkfunktionen gearbeitet.

Üblicherweise wird eine Zählvariable y als poissonverteilt vorausgesetzt. Sei μ der Erwartungswert dieser Zufallsvariable. Dann wird der Einfluss vom Alters des Kindes, der Zeit und des Geschlechts durch folgende Beziehung modelliert:

$$\ln(\mu) = ax + bt + cz. \tag{4.1}$$

In diesem Modell beschreibt a den Alterseffekt, b und c sind die auch in Modell (3.1) auftretenden Effekte von Zeit und Geschlecht.

5 Das GLIMMIX-Makro

Das Problem bei „Generalisierten Gemischten Linearen Modellen“ liegt darin, dass es soweit bekannt keine eigenständige SAS-Prozedur gibt, die diese Modelle korrekt behandeln kann. Lägen nur Messwiederholungen vor und gäbe es

das Problem der Geschwisterkinder nicht, könnte das Modell zufriedenstellend mit PROC GENMOD bearbeitet werden, da es sich hierbei um ein einfaches Generalisiertes Lineares Modell handeln würde.

Würden zwar auch Geschwisterkinder an der Studie teilnehmen (also zufällige Effekte auftreten), gäbe es aber keine unregelmäßigen Messwiederholungen, so könnte PROC NLMIXED erfolgreich eingesetzt werden. Im Fall der DONALD-Studie sind jedoch weder PROC GENMOD noch PROC NLMIXED für eine korrekte Datenanalyse geeignet.

Generalisierte Lineare Gemischte Modelle können jedoch mit dem SAS Makro GLIMMIX bearbeitet werden, welches SAS 8.0 in den SAS Online Samples zur Verfügung stellt. Eine ältere Version ist auch in [1], S. 505 abgedruckt.

Dabei ist zu bemerken, dass GLIMMIX auf PROC MIXED aufsetzt. Dieses bringt für den Anwender den Vorteil, dass alle program statements, die in PROC MIXED verfügbar sind (wie z. B. CLASS, MODEL, RANDOM, REPEATED), auch innerhalb dieses Makros benutzt werden können. Naheliegenderweise ist deshalb die Form des Outputs identisch mit dem Output von PROC MIXED. Nachteile von GLIMMIX sind ein oft sehr großer Zeitaufwand bei der Auswertung und die Tatsache, dass bisher wenig Forschungsarbeit hinsichtlich der in GLIMMIX zur Anwendung kommenden statistischen Verfahren, das heißt Schätzern die mit Quasi-Likelihood-Verfahren berechnet werden s. [2], in kleinen Stichproben geleistet worden ist. Die mit Hilfe von GLIMMIX berechneten Teststatistiken (und p-Werte), für die eine asymptotische Normalverteilung unterstellt wird, scheinen zwar zu vernünftigen Ergebnissen zu führen, dennoch ist hier noch ein größerer Forschungsbedarf notwendig insbesondere über die für eine zufriedenstellende Asymptotik erforderliche Stichprobengröße.

6 Anwendung von GLIMMIX bei binären Daten

In dem hier untersuchten Datensatz beschreibt die Variable REACH_CA, ob die empfohlene Kalziumzufuhr erreicht wurde oder nicht. Dabei steht, wie allgemein üblich, „0“ für „Empfehlung nicht erfüllt“ und „1“ für „Empfehlung erfüllt“. Es ist aber, ebenso wie in PROC GENMOD oder PROC LOGISTIC, auch möglich, Datensätze mit Variablen auszuwerten, die die Anzahl der Kinder mit ausreichender Kalziumzufuhr der Gesamtanzahl der Kinder gegenüberstellen. In diesem Fall stehen die entsprechenden Variablen als Quotient im MODEL statement. Es können allerdings in diesen Fällen Konvergenzprobleme bei der Schätzung der Modellparameter auftreten.

Obige Fragestellung kann mit folgendem SAS-Programm bearbeitet werden:

```
%GLIMMIX(DATA=CALCIUM,
  PROCOPT=METHOD=ML,
  STMTS=%STR(CLASS FAMILY  NRKIND  SEX;
             MODEL REACH_CA = P_ZUCK  JAHR  SEX / S ;
             RANDOM FAMILY NRKIND(FAMILY);
             REPEATED / SUB=FAMILY*NRKIND TYPE=SP(EXP)(ALTER) ),
  ERROR=BINOMIAL,
  LINK=LOGIT);
RUN;
```

Mit %GLIMMIX wird das Makro aufgerufen, DATA=CALCIUM gibt die benötigte Datei an. Hier wurden 846 Probanden untersucht und 4993 Beobachtungen gemacht. PROCOPT=METHOD=ML beschreibt, dass mit dem (quasi) Maximum-Likelihood-Schätzer gearbeitet wurde. Statements aus PROC MIXED stehen innerhalb der Klammern von STMTS=%(...). Dabei beschreibt FAMILY die Familiennummer und NRKIND die Kindnummer innerhalb einer Familie. Diese sind neben dem Geschlecht SEX, codiert mit „1“ für „Junge“ und „2“ für „Mädchen“, die Klassifikationsvariablen und müssen in dem CLASS statement vor dem MODEL statement aufgeführt werden. Das MODEL statement beschreibt das zugrunde gelegte Modell s. (3.1), das RANDOM statement berücksichtigt die zufälligen Effekte (Familien- und Kindeffekt), wobei hier der Kindeffekt als zufälliger Effekt innerhalb einer Familie angesehen wird und das REPEATED statement berücksichtigt die Korrelation zwischen zwei Messungen bei demselben Kind. Hier wird ein exponentieller Abfall der Korrelation mit zunehmendem zeitlichen Abstand angenommen.
Durch ERROR=BINOMIAL wird die Verteilung der Zielvariable beschrieben und LINK=LOGIT gibt an, welche Linkfunktion benutzt worden ist.

Folgendes Ergebnis wird mit obigem Programm erzeugt:

Untersuchung der Erreichung der empfohlenen Kalziumzufuhr – LOGIT Link

Effekt	Geschlecht	Schätzer	Standard-fehler	FG	Test-statistik	p-Wert
Intercept		1.2056	0.1424	223	8.46	<.0001
P_ZUCK		−0.03213	0.006999	4141	−4.59	<.0001
JAHR		−0.03837	0.009115	4141	−4.21	<.0001
SEX	1	0.7672	0.1029	4141	7.46	<.0001
SEX	2	0	.	.	.	.

Das Ergebnis zeigt, dass hier alle Effekte zum 5%-Niveau signifikant sind. Da die Schätzer für den Zuckerkonsum (−0.03213) und das Untersuchungsjahr (−0.03837) negativ sind, bedeutet das, dass die Wahrscheinlichkeit, die Empfehlungen einzuhalten, mit zunehmendem Zuckerkonsum wie erwartet abnimmt und auch ein zeitlicher Trend zu sehen ist, wenngleich diese Trends sehr

schwach sind. Der signifikante Geschlechtseffekt zeigt, dass Jungen die Empfehlungen sehr viel besser einhalten als Mädchen.

Nachteil des oben aufgeführten Programms ist jedoch ein sehr großer Zeitaufwand bei seiner Durchführung. Dieser kann vermieden werden, wenn das RANDOM statement durch folgende gleichwertige Zeile ersetzt wird

```
RANDOM INT NRKIND /SUB=FAMILY;
```

die denselben Ouput produziert.

Neben der Logit-Funktion ist es aber auch möglich, mit anderen Linkfunktionen bei Problemstellungen mit binären Zielvariablen zu arbeiten. So ist auch die Bearbeitung von Probit-Modellen ohne Probleme möglich. Diese Modelle sind mit GLIMMIX umsetzbar, sofern das statement LINK=LOGIT durch LINK=PROBIT ersetzt wird.

7 Anwendung von GLIMMIX bei Zählvariablen

Gegeben sei die Situation aus Kapitel 4. Das heißt, es soll untersucht werden, ob die Anzahl der pro Tag eingenommenen Mahlzeiten abhängig vom Alter und Geschlecht des Kindes ist. Außerdem interessiert, ob die Daten der DONALD-Studie einen zeitlichen Trend beinhalten.
Die Variable N_MAHLZ zähle die pro Tag gegessenen Mahlzeiten, unabhängig davon ob es sich um große oder kleine Mahlzeiten handelt.

Diese Fragestellung kann dann mit folgendem SAS-Programm analysiert werden:

```
%GLIMMIX(DATA=MAHLZ,
  PROCOPT=METHOD=ML,
  STMTS=%STR(CLASS  FAMILY  NRKIND  SEX;
             MODEL N_MAHLZ = ALTER  JAHR  SEX / S ;
             RANDOM  INT  NRKIND / SUB=FAMILY;
             REPEATED / SUB=FAMILY*NRKIND TYPE=SP(EXP)(ALTER) ),
  ERROR=POISSON,
  LINK=LOG);
RUN;
```

Das Programm ist also genauso aufgebaut wie das erste Programmbeispiel zur Analyse der binären Zielvariablen. Einzig die Verteilung der Zielvariablen und die Linkfunktion müssen entsprechend angepasst werden. Für die Verteilung der Zielvariablen wird eine Poissonverteilung angenommen und die Linkfunktion ist hier, siehe Modell (4.1), die natürliche Logarithmusfunktion $\ln(x)$.

Hier zeigt sich folgendes Ergebnis:

Untersuchung des Mahlzeitenverhaltens – LOG Link

Effekt	Geschlecht	Schätzer	Standard-fehler	FG	Test-statistik	p-Wert
Intercept		1.8933	0.01385	224	136.66	<.0001
ALTER		−0.00903	0.001276	4141	−7.07	<.0001
JAHR		−0.00046	0.001237	4141	−0.37	<.7077
SEX	1	0.003730	0.01115	4141	0.33	<.7380
SEX	2	0	.	.	.	.

Es zeigt sich, dass nur der Alterseffekt zum 5%-Niveau signifikant ist. Der Schätzer (−0.00903) selbst spricht für einen schwach abnehmenden Trend. Die Vermutung, dass sich in den letzten 18 Jahren eine Veränderung im Mahlzeitenverhalten ergeben hat, konnte also aufgrund der Datenlage hier, zumindestens was die Anzahl der Mahlzeiten betrifft, nicht bestätigt werden.

Zusammenfassend lässt sich folgendes über das SAS Makro GLIMMIX feststellen. Generalisierte Lineare Gemischte Modelle lassen sich mit GLIMMIX umsetzen. Zur Programmierung nötig sind Kenntnisse und Erfahrungen mit PROC MIXED, da statements aus dieser SAS Prozedur innerhalb von GLIMMIX benutzt werden müssen.

Neben den hier in den Beispielen gezeigten Linkfunktionen sind noch weitere auch selbst definierte möglich.

Nachteile des Makros sind ein nicht unbeträchtlicher Zeitaufwand bei der Anwendung und die kritisch zu beurteilenden Ergebnisse in kleinen Stichproben.

Literatur

[1] SAS Institute Inc. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

[2] McCullagh, P. und Nelder, J. A. (1989). Generalized Linear Models, 2nd Edition. New York: Chapman and Hall.