Generalisierte lineare Modelle mit SAS 8e

Andreas Christmann
Universität Dortmund
Hochschulrechenzentrum
44221 Dortmund
a.christmann@hrz.uni-dortmund.de

Zusammenfassung

Generalisierte lineare Modelle sind für die Praxis von großer Bedeutung. Es wird dargestellt, wie diese Modelle innerhalb der Software SAS 8e verwendet werden können. Verschiedene SAS-Prozeduren können zur Anpassung generalisierter linearer Modelle verwendet werden, z.B. PROC GENMOD, PROC GAM, PROC LOGISTIC, PROC PROBIT und PROC INSIGHT. Vor- und Nachteile dieser Prozeduren werden dargestellt. Die Möglichkeiten, die SAS zur Anpassung generalisierter linearer Modelle bietet, werden verglichen mit der Funktionalität von S-PLUS, SPSS und R. Im Tutorium wurde anhand zahlreicher Beispiele die Verwendung generalisierter linearer Modelle in SAS demonstriert.

Keywords: Generalisierte lineare Modelle, GAM, GENMOD, INSIGHT, LOGISTIC, PROBIT, SAS.

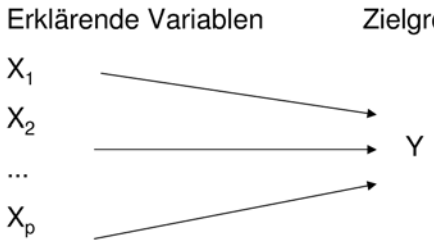
1. Einführung

Generalisierte lineare Modelle (kurz: GLIMs) werden in den unterschiedlichsten Bereichen in Forschung und Entwicklung an Hochschulen und in der Industrie eingesetzt. In fast allen Anwendungsbereichen wird selbstverständlich die klassische Regression und die Varianzanalyse eingesetzt, wenn die Zielgröße stetig, symmetrisch und approximativ normalverteilt ist.

Die größere Modellklasse der verallgemeinerten linearen Modelle ermöglicht es, auch Zielgrößen zu modellieren, deren Verteilungen diskret oder asymmetrisch sind. Bei biometrischen Studien modelliert man mittels GLIMs die Erkrankungswahrscheinlichkeiten in Abhängigkeit von potenziellen Einflußgrößen. Bei Versicherungen setzt man GLIMs zur Modellierung der Schadenhäufigkeit und des Schadenbedarfs ein. Als Stichwort sei hier das Marginalsummenmodell genannt. Zur Identifikation von Risikofaktoren bei der Kreditvergabe durch Banken werden ebenfalls GLIMs eingesetzt. Weitere Beispiele für den erfolgreichen Einsatz von generalisierten linearen Modellen sind die Identifikation potentieller Neukunden beim Direct Marketing sowie die Identifikation zufriedener bzw. unzufriedener Kunden im Customer Relationship Management (CRM).

Generalisierten linearen Modellen liegt die in Abbildung 1 dargestellte Unterscheidung zwischen potenziellen Einflußgrößen X_1, ..., X_p und einer Zielgröße Y (oder mehreren Zielgrößen) zugrunde.

Abbildung 1: Schema für generalisierte lineare Modelle



Generalisierte lineare Modelle unterscheiden sich bezüglich der getroffenen Annahmen über das Skalenniveau und die Verteilung der Zielgröße, sowie über die Art des Einflußes der erklärenden Variablen auf die Zielgröße. GLIMs sind u.a. für die Praxis deshalb so wichtig, weil viele unterschiedliche Verteilungstypen für die Zielgröße modelliert werden können, z.B. die Normalverteilung für stetige und symmetrisch verteilte Zielgrößen, die Binomialverteilung zur Modellierung von Ereigniswahrscheinlichkeiten, die Poisson-Verteilung zur Modellierung der Anzahl seltener Ereignisse und die Gamma-Verteilung für stetige und positive Zielgrößen mit einer asymmetrischen Verteilung.

2. Generalisierte lineare Modelle

Verallgemeinerte lineare Modelle (GLIMs) bilden eine große und flexible Modellklasse und spielen in vielen Anwendungsbereichen eine wichtige Rolle. Standard-Lehrbücher zu GLIMs sind z.B. Cox und Snell (1989), McCullagh und Nelder (1989) und Agresti (1996). Ein verallgemeinertes lineares Modell ist charakterisiert durch die folgenden Strukturen.

- Die Zielvariablen $Y_1,...,Y_n$ sind stochastisch unabhängig und besitzen eine Wahrscheinlichkeitsverteilung aus einer Exponentialfamilie. Die Varianz von Y_i hängt vom Erwartungswert von Y_i über eine sogenannte Varianzfunktion V ab, so daß gilt: $Var(Y_i)=\phi V(\mu_i)/w_i$, wobei der Dispersionsparameter ϕ eine Konstante ist und w_i ein bekanntes Gewicht für die i-te Zufallsvariable ist.
- Es liegt eine feste oder eine stochastische (n,p)-dimensionale Designmatrix X vor, wobei x_i^T die i-te Zeile von X ist.
- Es existiert eine lineare Komponente (linearer Prädiktor) $\eta_i=x_i^T\beta$, wobei $\beta \in R^p$ unbekannt ist.
- Es existiert eine monotone differenzierbare Link-Funktion g mit der Eigenschaft, daß zwischen dem Erwartungswert μ_i der Zielvariablen Y_i und dem linearen Prädiktor die Beziehung $g(\mu_i) = x_i^T\beta$ gilt.

Für multivariate GLIMs vergleiche man Fahrmeir und Kaufmann (1985) und Liang und Zeger (1986).

Aus mathematischer Sicht unterscheidet man kanonische und nicht-kanonische Link-Funktionen. In der folgenden Tabelle sind wichtige Beispiele für GLIMs aufgeführt.

Tabelle 1: Wichtige Spezialfälle generalisierter linearer Modelle

Modell	**Skala der Zielvariablen**	**Verteilung**	**Link-Funktion**	**Varianz-Funktion**
Lineare Regression und ANOVA	Stetig	Normal	Identität; $\eta_i=\mu_i$	$V(\mu_i)= 1$
Logistische Regression	Diskret; 0/1 oder 0/1/…/m	Binomial	LOGIT; $\eta_i=\Lambda^{-1}(\mu_i)=\log(\mu_i /(1-\mu_i))$ Λ: Verteilungsfunktion der logistischen Verteilung	$V(\mu_i)=\mu_i(1-\mu_i)$
Generalisierte logistische Regression	Diskret; Nominal	Multinomial	GLOGIT; $\eta_{i,j} = \log(\frac{P(Y_i = j \mid x_i)}{P(Y_i = k_{ref} \mid x_i)}) = x_i^T \beta_j$	
Proportional Odds Regression	Diskret; Ordinal	Multinomial	CUMLOGIT ; $\eta_{i,j} = \log(\frac{P(Y_i \le j \mid x_i)}{1 - P(Y_i \le j \mid x_i)}) = \gamma_j + x_i^T \beta$	
Probit-Regression	Diskret; 0/1 oder 0/1/.../m	Binomial	PROBIT; $\eta_i=\Phi^{-1}(\mu_i)$; Φ : Verteilungsfunktion der Standardnormal-verteilung	$V(\mu_i)=\mu_i(1-\mu_i)$
Poisson-Regression	Anzahl; ≥ 0	Poisson	$\eta_i=\log(\mu_i)$	$V(\mu_i)=\mu_i$
Gamma-Regression mit LOG-Linkfunktion	Positiv; stetig	Gamma	$\eta_i=\log(\mu_i)$	$V(\mu_i)=\mu_i^2$
Gamma-Regression mit inverser Linkfunktion	Positiv; stetig	Gamma	$\eta_i=1/\mu_i$	$V(\mu_i)=\mu_i^2$
Inverse Gauß-Regression	Positiv; stetig	Inverse Gauß	$\eta_i=1/\mu_i^2$	$V(\mu_i)=\mu_i^3$
Negative Binomial-Regression	Anzahl; ≥ 0	Negative Binomial	$\eta_i=\log(\mu_i)$	$V(\mu_i)= \mu_i +k\mu_i^2$

Zur Schätzung des unbekannten Parametervektors $\beta \in R^p$ wird derzeit sicher der Maximum-Likelihood-Schätzer (ML-Schätzer) am häufigsten verwendet. Die ML-Schätzung ist impliziert definiert als Lösung des folgenden Maximierungsproblems, wobei *L* die Likelihood-Funktion bezeichnet:

$$L(y, \beta, \phi) = \prod_{i=1}^{n} f(y_i, \mu_i(\beta), \phi, w_i) = \sup !$$

Numerisch ist es oft einfacher, die logarithmierte Likelihood-Funktion zu maximieren:

$$\log L(y, \beta, \phi) = \sum_{i=1}^{n} \log(f(y_i, \mu_i(\beta), \phi, w_i) = \sup !$$

Die Beliebtheit der ML-Methode hat neben theoretischen Überlegungen wie z.B. guten asymptotischen Eigenschaften auch den Grund, dass dieses Schätzverfahren in den meisten gängigen Software-Produkten für statistische Fragestellungen implementiert ist. Im Gegensatz zum Spezialfall der klassischen linearen Regression und der Varianzanalyse läßt sich der ML-Schätzwert in GLIMs in der Regel nicht in geschlossener Form angeben. Er wird deshalb mit numerischen Iterationsverfahren bestimmt. Hier seien nur modifizierte Newton-Raphson-Verfahren und Fisher's Scoring-Verfahren erwähnt.
Basierend auf der ML-Schätzmethode lassen sich approximative bzw. asymptotische Profile-Likelihood-Konfidenzintervalle sowie Konfidenzintervalle nach Wald für die zu schätzenden Parameter angeben. Klassische Tests zur Prüfung, ob und wenn ja, welche potenziellen Einflußgrößen einen signifikanten Einfluß auf die Zielgröße haben, sind der Likelihood-Ratio-Test, der Score-Test und der Wald-Test. Hauck und Donner (1977) zeigten, daß der Wald-Test bei logistischen Regressionsmodellen zu irreführenden Ergebnissen führen kann, so daß sich für derartige Modelle insbesondere der Likelihood-Ratio-Test anbietet. LR-Tests sind jedoch nicht in allen SAS-Prozeduren für GLIMs implementiert, vgl. Abschnitt 3.
Es sei allerdings erwähnt, dass diesen unbestreitbaren Vorteilen des Maximum-Likelihood-Schätzers auch gravierende Nachteile gegenüberstehen. So ist der ML-Schätzer in der Regel nicht robust gegenüber einer Verletzung der strikten parametrischen Modellannahmen und des Auftretens von Ausreißern. Oft genügt schon ein einziger extremer Wert, um den Wert der ML-Schätzung beliebig stark zu beeinflussen. In diesem Fall spricht man von einem Bruchpunkt von 0. Robuste Verfahren basierend auf Schätzern vom M-Typ wurden u.a. vorgeschlagen von Pregibon (1982), Künsch, Stefanski und Carroll (1989) sowie Cantoni und Ronchetti (2001). Robuste Schätzverfahren mit einem hohen Bruchpunkt wurden u.a. von Christmann (1994, 1998) vorgestellt.
In mancher Hinsicht ist die binäre Regression mit der logistischen Regression als wichtigstem Vertreter ein wichtiger Spezialfall. Albert und Anderson (1984) und Santner und Duffy (1986) zeigten, daß der Maximum-Likelihood-Schätzwert bei der logistischen Regression mit einem Intercept-Term *nicht* für alle möglichen Datensätze existiert. Der ML-Schätzwert existiert genau dann, wenn der Datensatz einen Overlap besitzt, d.h. falls es *keine* affine Hyperebene

gibt, die die Erfolge (y_i=1) von den Mißerfolgen (y_i=0) trennt. Der ML-Schätzwert bei logistischer Regression mit einem Intercept-Term existiert *nicht*, falls der Datensatz vollständig bzw. quasi-vollständig separierbar ist, d.h. wenn eine derartige affine Hyperebene existiert. Nun ist jedoch das Problem, die minimale Anzahl von Fehlklassifikationen zu ermitteln, die bezüglich einer beliebigen affinen Hyperebene möglich ist, NP-hart, so daß kein effizienter Algorithmus existiert, um diese Zahl für beliebige Datensätze exakt zu berechnen, vgl. Höffgen, Simon, und van Horn (1995). Christmann und Rousseeuw (2001) schlugen daher approximative Verfahren basiert auf der regression depth Methode zur Bestimmung dieser Zahl vor. Christmann, Fischer und Joachims (2002) modifizierten diese Verfahren und verglichen sie mit der insbesondere bei Informatikern beliebten Support Vector Machine (SVM), vgl. Vapnik (1995), Schölkopf und Smola (2002) und Hastie, Tibshirani und Friedman (2001).
In manchen Anwendungen liegen mehrere voneinander abhängige Zielgrößen (Y_1,...,$Y_{q)}$ vor. Als Beispiel sei hier nur eine Poisson-Regression genannt, bei der die beiden Zielgrößen (Y_1 , Y_2) die Anzahl der Fehler angeben, die dieselbe Person in einem großen Warenlager bei der Kommissionierung von Aufträgen unter zwei unterschiedlichen Versuchsbedingungen begeht. Zur Modellierung derartiger Fragestellungen verwendet man oft den von Liang und Zeger (1986) vorgeschlagenen GEE-Ansatz (Generalized Estimating Equation), bei der die unbekannten Parameter geschätzt werden als implizite Lösung des folgenden Nullstellenproblems:

$$S(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta}^T V_i^{-1}(\beta)(Y_i - \mu_i(\beta)) = 0 .$$

3. SAS-Prozeduren für GLIMs

In SAS bestehen viele Möglichkeiten, generalisierte lineare Modelle zu verwenden. Hierzu zählen insbesondere die Prozedur PROC GENMOD sowie SAS/INSIGHT. Aber auch die Prozeduren PROC GAM, PROC LOGISTIC, PROC PROBIT, PROC CATMOD und PROC PHREG können zur Anpassung mancher generalisierter linearer Modelle eingesetzt werden. Für manche Auswertungen im Rahmen vom GLIMs eignen sich auch SAS/ANALYST, SAS/ASSIST, SAS/Enterprise Miner, SAS/IML sowie die Prozeduren PROC NLMIXED und PROC NLIN. Zudem besteht eine Vielzahl von SAS/Macros, die in Verbindung mit GLIMs eingesetzt werden können. Das Macro %ROC stellt z.B. eine Implementation der von deLong, deLong und Clarke-Pearson (1988) vorgeschlagenen nichtparametrischen Methode zum Vergleich mehrerer voneinander abhängiger ROC-Kurven dar. Ein anderes Beispiel ist %GLIMMIX für generalized linear mixed models.
Im Einzelfall ist nicht immer unmittelbar klar, welche der im vorigen Abschnitt genannten Prozeduren für die konkrete Fragestellung am geeignesten ist. Um die Entscheidung für eine spezielle Prozedur bzw. Modul zu erleichtern, enthält die Tabelle 2 spezifische Informationen zu den einzelnen SAS-Produkten ohne

selbstverständlich alle Möglichkeiten und Optionen aufzulisten. Zu beachten ist auch, dass verschiedene SAS-Prozeduren mitunter andere Parametrisierungen für diskrete Einflußgrößen bzw. für die Zielgröße verwenden. Gegebenfalls müssen spezielle Optionen wie etwa DESCENDING, REF=1 oder PARAM=REF bei PROC LOGISTIC verwendet werden, um die in der statistischen Literatur üblichen Parametrisierungen umzusetzen.

Tabelle 2: Kurzübersicht über verschiedene Prozeduren/Module für GLIMs

PROC/Modul	**Kurzbeschreibung**
GENMOD	Allround-Prozedur, flexibel, schnell, LR-Tests, GEE
GAM	semiparametrisches Modell, Prüfung von GLIM-Modellannahmen
LOGISTIC	kategoriale Regression, Ausreißer-Erkennung, PL-Konfidenzintervalle, Check auf Separation (Nichtexistenz des ML-Schätzwertes, automatische Modellwahl, exakte bedingte logistische Regression
PROBIT	Zielgröße diskret, Inverse Regression und Konfidenzgrenzen (EDα), Schätzung einer spontanen Ereignis-Wahrscheinlichkeit
CATMOD	viele Response-Funktionen, log-lineare Modelle, Profil-Analyse mit GSK-Ansatz
PHREG	Fall-Kontroll-Studien bei m:n Matching, Proportional Hazards Modell
NLMIXED	Nichtlineare gemischte Modelle (feste und zufällige Effekte); Verteilungstypen: Normal, Binomial, Gamma, Negbin, Poisson, General
NLIN	Nichtlineare Modelle; Schätzmethoden: Least Squares, Weighted Least Squares
SAS/INSIGHT	Interaktiv, dynamische Graphiken, aber relativ langsam für große Dateien, LR-Tests
SAS/Analyst	Verwendung einer graphischen Oberfläche; lineare oder logistische Regression
SAS/ASSIST	Verwendung einer graphischen Oberfläche, lineare oder binäre Regression (logit, probit, cloglog)
SAS/Enterprise Miner	Data Mining, sehr große Dateien
SAS/IML	Entwicklung eigener Verfahren bzw. Modifikation bestehender Verfahren

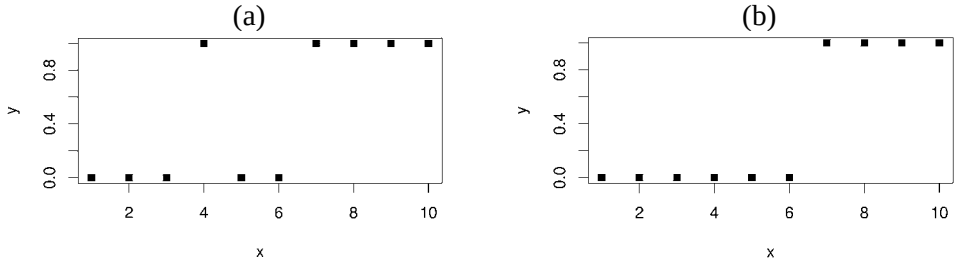
Für einen gegebenen Datensatz, ein spezifiziertes generalisiertes lineares Modell und identischer Parametrisierung stehen also in der Regel mehrere Prozeduren bzw. Module zur Verfügung, um dieses Modell anzupassen. In diesem Fall führen verschiedene Prozeduren bzw. Module oft - aber durchaus nicht immer - zu den gleichen Ergebnissen, wenn man von minimalen numerischen Unterschieden absieht, die u.a. in der Wahl unterschiedlicher Abbruchkriterien der verwendeten Iterationsverfahren begründet sind. Dies soll an einem sehr

einfachen Beispiel demonstriert werden, obwohl das Problem bei komplexen Datensätzen ebenfalls auftreten kann und dort viel schwieriger zu entdecken ist. Betrachtet wird ein einfaches logistisches Regressionsmodell mit einem Intercept-Term, einer stetigen erklärenden Variablen x_i , einer binären Zielgröße y_i und 10 Beobachtungen. Der Datensatz in Situation (a) ist gegeben durch :

x_i : 1, 2, 3, 4, 5, 6, 7, 8, 9,10 und y_i : 0, 0, 0, 1, 0, 0, 1, 1, 1, 1.

Der Datensatz in Situation (b) unterscheidet sich von dem in der vorigen Situation nur darin, dass y_4=0 statt y_4=1 ist. Wie die Abbildung 2 zeigt, hat der Datensatz (a) offenbar overlap und die ML-Schätzung für den Parametervektor existiert und ist eindeutig, während für den Datensatz (b) die ML-Schätzung nicht existiert, da mindestens ein geschätzer Koeffizient bei $+\infty$ oder $-\infty$ ist, wie man mit den allgemeinen Ergebnissen von Albert und Anderson (1984) und Santner und Duffy (1986) leicht beweist.

Abbildung 2: Darstellung der Daten des Beispiels zur logistischen Regression



Die Tabelle 3 zeigt, dass für den Datensatz (a) viele verschiedene SAS-Prozeduren zu den gleichen numerischen Resultaten führen. Lediglich PROC GAM berechnet die p-Werte des Wald-Tests zur Prüfung, ob x ein signifikanten Einfluß auf die Zielgröße hat, basierend auf einer Approximation durch eine t-Verteilung statt einer Normalverteilung.

Tabelle 3: Logistische Regression für den Datensatz (a)

PROC	**Schätzung für**		**p-Wert für x**		**Bemerkung**
	Intercept	**x**			
GENMOD	-4.84	0.88	LR	0.010	
			Wald	0.088	
INSIGHT	-4.84	0.88	LR	0.010	
			Wald	0.088	
LOGISTIC	-4.84	0.88	LR	0.010	
			Score	0.021	
			Wald	0.088	
	-1.77	0.75	Exact	0.032	
GAM	-4.84	0.88	Wald	0.126	p-value aus t-Verteilung
PROBIT	-4.84	0.88	Wald	0.088	
CATMOD	-4.84	0.88	Wald	0.088	

Ein völlig anderes Bild zeigt sich jedoch für Datensatz (b), obwohl sich dieser nur in einem einzigen Datenpunkt von Datensatz (a) unterscheidet, vgl. Tabelle 4. Einige SAS-Prozeduren, u.a. auch GENMOD, geben die nicht korrekte Information, dass der Iterationsalgorithmus konvergiert habe. PROC LOGISTIC erkennt korrekt, dass der Datensatz vollständig separierbar ist und somit die ML-Schätzung nicht existiert. Trotzdem gibt auch aber PROC LOGISTIC Schätzwerte an. Diese unterscheiden sich offenbar für diesen Datensatz von Prozedur zur Prozedur erheblich. Gleiches gilt für die p-Werte selbst gleicher Tests. Bei den Wald-Tests ergeben sich sowohl p-Werte, die kleiner als 0.001 sind, als auch p-Werte nahe bei 1.0. Bemerkenswert ist auch, dass PROC GENMOD und SAS/INSIGHT für diesen Datensatz zwar die gleichen numerischen Werte als ML-Schätzwerte angeben, die Wald-Tests jedoch zu völlig anderen Ergebnissen führen: während der p-Wert bei PROC GENMOD bei 1.0 liegt und somit x nicht als signifikante Einflußgröße auf dem 5% Niveau bezeichnet wird, gibt SAS/INSIGHT einen p-Wert von kleiner als 0.001 an, der zu allen gebräuchlichen Testniveaus die Variable x als signifikante Einflußgröße ausweist. Wie Tabelle 5 zeigt, liegt der Grund für diese unterschiedlichen Ergebnisse der WALD-Tests in PROC GENMOD und SAS/INSIGHT darin, dass diese beiden SAS-Prozeduren extrem unterschiedlich Werte für die approximativen Standardfehler der ML-Schätzwerte für diesen Datensatz angeben (unterschiedliche Vorzeichen im Exponent)!

Tabelle 4: Logistische Regression für den Datensatz (b)

PROC	**Schätzung für**		**p-Wert für x**		**Bemerkung**
	Intercept	**x**			
GENMOD	-498.3	76.7	LR	< 0.001	NOTE: Algorithm converged.
			Wald	1.000	
INSIGHT	-498.3	76.7	LR	< 0.001	Convergence not attained in 40 iterations.
			Wald	< 0.001	The validity of the model fit is questionable.
LOGISTIC	-65.1	10.0	LR	< 0.001	WARNING: There is a complete separation of data points.
			Score	0.007	The maximum likelihood estimate does not exist.
			Wald	0.566	WARNING: The LOGISTIC procedure continues in spite of
	-2.4	1.0	Exact	0.0095	the above warning.
GAM	-169.7	26.1	Wald	0.816	WARNING: The local score algorithm did not converge.
PROBIT	-584.6	89.9	Wald	1.000	NOTE: Algorithm converged.
CATMOD	-116.6	17.9	Wald	-	NOTE: Maximum likelihood computations converged.

Tabelle 5: Logistische Regression für den Datensatz (b)

PROC	Schätzung für		Standardfehler für		p-Wert für x	
	Intercept	x	Intercept	x		
GENMOD	-498.3	76.7	1.96 E+9	3.08 E+8	LR-Test:	< 0.001
					Wald-Test:	1.000
INSIGHT	-498.3	76.7	6.76 E-9	4.32 E-8	LR-Test:	< 0.001
					Wald-Test:	< 0.001

Mit PROC LOGISTIC läßt sich eine exakte bedingte logistische Regression anpassen, die obiges Problem weitgehend umgeht. Hierbei bedingt man bezüglich einer unter den parametrischen Modellannahmen des logistischen Regressionsmodells gegebenen suffizienten Statistik. Diese Methode hat jedoch zwei wesentliche Nachteile. Zum einen ist sie sehr rechenintensiv, so daß sie bisher (selbst mit der Software LogXact von Cytel) nicht für große komplexe Datensätze anwendbar ist. Zum anderen ist in der konkreten Anwendung i.a. unbekannt, ob die parametrischen Modellannahmen des logistischen Regressionsmodells *exakt* zutreffen. So ist oft nicht bekannt, ob die Daten einer logistischen Regression, einer Probit-Regression oder einem anderen binären Regressionsmodell folgen. Das Konzept der suffizienten Statistik beruht aber wesentlich auf der Annahme, dass die Modellannahme exakt erfüllt ist, da sonst entweder gar keine sinnvolle suffiziente Statistik existiert oder keine solche Statistik bekannt ist.
Einen alternativen Ansatz schlagen Rousseeuw und Christmann (2002) vor, um das Existenz-Problem des ML-Schätzers bei logistischen Regressionsmodellen zu umgehen.

4. Überblick über die SAS-Prozedur PROC GENMOD

Die Syntax-Struktur zu der vielseitigen Prozedur PROC GENMOD lautet:

PROC GENMOD ... ;
BY variables;
CLASS variables; ←*diskrete Einflußgrößen*
CONTRAST 'label' effect values ... ; ← *LR-Test:* $H_0 : L'\beta = 0$
ESTIMATE 'label' effect values ... ; ← *Wald-Test:* $H_0 : l'\beta = 0$
FREQ variable;
LSMEANS effects ...; ← *LS-Mean* $L'\hat{\beta}$
MAKE 'table' OUT=SAS-data-set; ← *oder ODS !*
OUTPUT < OUT=SAS-data-set > <keyword=name ...>;
MODEL response = <effects> ... ; ← *Modell-Spezifikation*
REPEATED SUBJECT= subject-effect ... ; ← *Kovarianz für GEE-Modelle*
WEIGHT variable; ← *Gewichte* w_i

Die Verteilung der Zielgröße, die Default Link-Funktion, und weitere Link-Funktionen, die in PROC GENMOD verfügbar sind, sind in den Tabellen 6 und 7 aufgelistet.

Tabelle 6: Verteilung der Zielgröße und Default Link-Funktion (GENMOD)

Spezifikation der Verteilung für Y		**Default Link-Funktion**
NORMAL	Normal-Verteilung	μ
BINOMIAL	Binomial-Verteilung	Logit; $\log(\mu/(1-\mu))$
MULTINOMIAL	Multinomial-Verteilung	Cumulative logit
POISSON	Poisson-Verteilung	$\log(\mu)$
GAMMA	Gamma-Verteilung	$1/\mu$
IGAUSSIAN	Inverse Gau{ss}-Verteilung	$1/\mu^2$
NEGBIN	Negative Binomial-Verteilung	$\log(\mu)$

Tabelle 7: Weitere Link-Funktionen für PROC GENMOD

LINK=	**Link-Funktion**
IDENTITY	identity
LOG	log
LOGIT	logit
PROBIT	probit
CUMCLL	cumulative complementary log-log
CUMLOGIT	cumulative logit
CUMPROBIT	cumulative probit
CLOGLOG	complementary log-log
POWER(number)	power with number

Mit dem folgenden SAS-Programm kann zum Beispiel eine Poisson-Regression mit LOG-Link-Funktion für eine Zielgröße (Anzahl) mit drei diskreten Einflußgrößen (Klasse, Gruppe und Fahr) und einer stetigen Einflußgröße (Alter) angepaßt werden. Zusätzlich werden Konfidenzintervalle zum Niveau 95% berechnet sowie LR-Tests und eine Residualanalyse durchgeführt. Abweichend von den Default-Spezifikationen wird die gewünschte Genauigkeit mit 10^{-10} bei einer maximalen Iterationsanzahl von 200 angegeben.

```
PROC GENMOD DATA=a ORDER=INTERNAL;
  CLASS Klasse Gruppe Fahr;
  MODEL Anzahl = Alter Klasse Gruppe Fahr /
          DIST=POISSON LINK=LOG NOSCALE LRCI OBSTATS
          TYPE3 ALPHA=0.05 CONVERGE=1.E-10 MAXIT=200 ITPRINT;
RUN;
```

PROC GENMOD kann flexibel zur Anpassung von GLIMs eingesetzt werden, die nicht standardmäßig in SAS implementiert sind. Hierzu dienen die

Statements FWDLINK, INVLINK und DEVIANCE. Als Beispiel sei hier ein binäres Regressionsmodell für die Zielgröße y mit den erklärenden Variablen Gruppe (diskret) und x (stetig) genannt, bei dem als Link-Funktion statt der kanonischen Link-Funktion (Inverse der logistischen Verteilungsfunktion) die inverse Verteilungsfunktion der Student'schen t-Verteilung mit 3 Freiheitsgraden verwendet wird, die über dickere Tails als die logistische Verteilung verfügt.

```
PROC GENMOD DATA=a DESCENDING ORDER=DATA;
 CLASS Gruppe;
 FWDLINK LINK=TINV(_MEAN_,3);
 INVLINK ILINK=PROBT(_XBETA_,3);
 MODEL y= Gruppe x / DIST=BINOMIAL TYPE3 LRCI MAXIT=200; RUN;
```

Die Prozedur PROC GENMOD ist flexibel und relativ schnell. Likelihood-Ratio-Tests sind implementiert, was aufgrund des in Hauck und Donner (1977) beschriebenen Phänomens insbesondere bei binären Regressionsmodellen wichtig ist, da Wald-Test hier zu irreführenden Ergebnissen führen können.
In PROC GENMOD ist der wichtige GEE-Ansatz nach Liang und Zeger (1986) zur Modellierung mehrerer abhängiger Zielgrößen implementiert.
Im Gegensatz zu PROC LOGISTIC führt PROC GENMOD bei logistischer Regression leider keine Prüfung auf Existenz der ML-Schätzung durch und eine exakte bedingte logistische Regression ist nicht implementiert. In Version 8.2 ist zwar eine tabellarische Residualanalyse basierend auf dem nicht-robusten ML-Schätzer implementiert, es fehlen jedoch Regression Diagnostics. Moderne robuste Schätzmethoden wie M-Schätzer oder S-Schätzer sind in PROC GENMOD in der Version 8.2 nicht implementiert.

5. PROC GAM

Zur Prüfung, ob die Modellannahme eines linearen Prädiktors

$$E(Y \mid x_1,...,x_p) = g^{-1}(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)$$

in generalisierten linearen Modellen gerechtfertigt ist, eignet sich u.a. die von Hastie und Tibshirani (1990) vorgeschlagene größere Klasse der generalisierten additiven Modelle (GAM), bei denen statt eines linearen Zusammenhangs ein additiver Zusammenhang gemäß

$$E(Y \mid x_1,...,x_p) = g^{-1}(\beta_0 + f_1(x_1) + ... + f_p(x_p)),$$

angenommen wird, wobei die Funktionen f_j unbekannte Funktionen sind.

Mit PROC GAM können derartige Modelle unter Verwendung von smoothing splines (SPLINE), local regression (LOESS) und bivariaten Splines (SPLINE2) angepaßt werden. Die Glättungsverfahren SPLINE2 und LOESS sind besonders rechenintensiv. Unter SAS 8.2 TS2M0 scheint die LOESS-Methode nicht immer stabil zu sein. Andere Glättungsmethoden sind in der Version 8.2 nicht implementiert. Für die Praxis wäre es beispielsweise wünschenswert, wenn der

Nutzer sein Vorwissen für die Art der Funktionen f_j angeben könnte, wie z.B. isoton, antiton, konkav oder konvex.
Als Verteilungen für die Zielgröße kann angegeben werden: NORMAL, BINOMIAL, POISSON und GAMMA. In der Online-Help wird zusätzlich noch als Verteilung IGAMMA aufgeführt. Vermutlich ist hier IGAUSS für die inverse Gaußverteilung gemeint.

6. Vergleich mit anderen Software-Produkten

Bei einem Vergleich unterschiedlicher Software-Produkte spielen selbstverständlich viele Faktoren wie Betriebssystem, Leistungsumfang, Flexibilität, Erweiterbarkeit, Dokumentation, Größe der zu analysierenden Dateien und nicht zuletzt der Preis eine Rolle. Statt eines allgemeinen Vergleichs werden hier einige Gemeinsamkeiten und Unterschiede der folgenden vier Software-Produkte unter Windows *ausschließlich unter dem Aspekt generalisierter linearer Modelle* aufgeführt:

- SAS Version 8.2
- SPSS Version 11
- S-PLUS Version 6
- R Version 1.5.

Von diesen vier Software-Produkten bietet SPSS eindeutig den geringsten Leistungsumfang (binär logistisch, multinomial logistisch, ordinal, probit mit Schätzung der EDα), da selbst die wichtigen Spezialfälle der Poisson- und der Gamma-Regression nicht standardmäßig verfügbar sind.
SAS bietet – wie oben beschrieben – ein erheblich umfangreicheres Spektrum als SPSS an. Im konkreten Anwendungsfall ist nicht immer unmittelbar klar, welche der vielen SAS-Prozeduren für GLIMs eingesetzt werden sollte. Die Implementation robuster Schätz- und Testmethoden sind wünschenswert, wie sie inzwischen für lineare Regressionsmodelle innerhalb von SAS/IML mit LTS, LMS, MCD und MVE verfügbar ist, vgl. auch Chen (2002). Mit der Prozeduer GAM können generalisierte additive Modelle angepaßt werden, um die Modellannahmen eines GLIMs zu prüfen.
S-PLUS bietet mit der Funktion glm die Möglichkeit, generalisierte lineare Modelle für unterschiedliche Response-Verteilungen mit einer Vielzahl von Link-Funktionen anzupassen. S-PLUS bietet neben der klassischen Maximum-Likelihood-Methode auch die Möglichkeit, GLIMs mit robusten Schätz-verfahren anzuwenden und die Ergebnisse in übersichtlicher tabellarischer oder graphischer Form gegenüberzustellen. Die S-PLUS Funktion gam erlaubt die Anpassung generalisierter additiver Modelle, die u.a. hilfreich sind, um Modellannahmen von GLIMs zu prüfen.
Die leistungsfähige GNU-Software R (http://www.r-project.org/) weist viele Ähnlichkeiten zu S-PLUS auf. Wie S-PLUS bietet auch R mit der Funktion glm die Möglichkeit, generalisierte lineare Modelle flexibel anzupassen. Mithilfe ergänzender functions und packages, die im Internet verfügbar sind, läßt sich die Leistungsfähigkeit von R weiter erhöhen. Hier seien nur car, lqs, gee, lqs, mgcv und Mass genannt.

Im Vergleich zu anderen Software-Paketen treten bei SAS seltener Speicherplatz-Probleme beim Anpassen generalisierter linearer Modelle für größere Datensätze auf.

7. Zusammenfassung

SAS bietet mit der Version 8.2 vielfältige Möglichkeiten, univariate und multivariate generalisierte lineare Modelle mittels Programmierung oder per GUI anzupassen. Die Prozedur GENMOD deckt ein großes Spektrum an unterschiedlichen Response-Verteilungen und Link-Funktionen ab. Für spezielle Fragestellungen kann es jedoch im Einzelfall sinnvoll sein, eine andere SAS-Prozedur zur Anpassung generalisierter linearer Modelle zu verwenden. Da verschiedene SAS-Prozeduren mitunter unterschiedliche Parametrisierungen diskreter Einflußgrößen oder der Response-Variablen verwenden, ist bei der Interpretation der Ergebnisse ein sorgfältiges Lesen der Dokumentation ratsam. Dies trifft insbesondere auf die Schätzer der Parameter und der Odds Ratios zu. Eine Implementation moderner robuster Schätz- und Testverfahren ist wünschenswert.

Literatur

1. Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley, New York.
2. Albert, A., Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71, 1-10.
3. Cantoni, E., Ronchetti, E. (2001). Robust inference for generalized linear models. JASA, 96,1022-1030.
4. Chen, C. (2002). Robust tools in SAS. In: Dutter et al. Developments in robust statistics. Physika, Heidelberg, pp. 125-133.
5. Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. Biometrika, 81, 413-417.
6. Christmann, A. (1998). On positive breakdown point estimators in regression models with discrete response variables. Habilitationsschrift. Universität Dortmund.
7. Christmann, A., Fischer, P., Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. Computational Statistics, 17, 273-287.
8. Christmann, A., Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. Computational Statistics & Data Analysis, 37, 65-75.
9. Cox, D.R., Snell, E.J. (1989). Analysis of Binary Data. 2nd ed. Chapman & Hall, London
10. deLong, E.R., deLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics, 44, 837-845.

11. Fahrmeir, L., Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Annals of Statistics, 13, 342-368.
12. Hastie, T, Tibshirani, R. (1990). Generalized additive models. Chapman & Hall, London.
13. Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of statistical learning. Springer, New York.
14. Hauck, Jr., W.W., Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. JASA, 72, 851-853.
15. Höffgen, K.U., Simon, H.-U., van Horn, K.S. (1995). Robust Trainability of Single Neurons. J. Computer and System Sciences, 50, 114-125.
16. Liang, K.Y., Zeger, S.L. (1986). Longitudinal Data Analysis using generalized linear models. Biometrika, 73, 13-22.
17. Künsch, H.R., Stefanski, L.A., Carroll, R.J. (1989). Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models. JASA, 84, 460-466.
18. LogXact for Windows, User Manual, Cytel Software Corporation (1996)
19. McCullagh, P., Nelder, J.A. (1989). Generalized Linear Models. 2nd,ed. Chapman & Hall, London.
20. Pregibon, D. (1982). Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. Biometrics, 38, 485-498.
21. Rousseeuw, P.J., Christmann, A. (2002). Robustness against separation and outliers in logistic regression. Erscheint in: Computational Statistics & Data Analysis.
22. Santner, T.J., Duffy, D.E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika, 73, 755-758.
23. SAS/INSIGHT User's Guide (1993). SAS Institute Inc.
24. SAS/STAT Users Guide I und II (1990). SAS Institute Inc.
25. SAS/STAT Software: Changes and Enhancements, through Release 6.11 (1996). SAS Institute Inc.
26. SAS/STAT Software: Changes and Enhancements, Release 8.2 (2001). SAS Institute Inc.
27. Schölkopf, B, Smola, A.J. (2002). Learning with kernels. Support vector machines, regularization, optimization, and beyond. MIT-Press, Cambridge.
28. Stokes, M.E., Davis, C.S., Koch, G. (1995). Categorical data analysis using the SAS System, SAS Institute Inc.
29. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York.