GENERALISIERTE LINEARE MODELLE MIT SAS 8e

Andreas Christmann

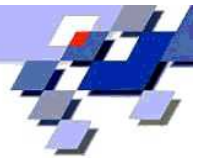
Universität Dortmund

A.Christmann@hrz.uni-dortmund.de

KSFE 2003, Potsdam

20.-21. Februar 2003

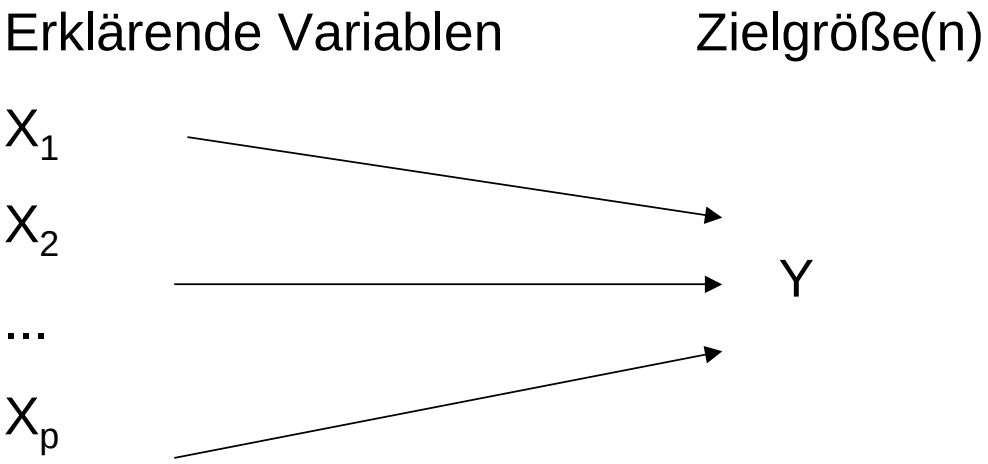
INHALT



1. ANWENDUNGEN

- in fast allen Bereichen: klassische Regression und Varianzanalyse
- Biometrie: Modellierung von Erkrankungswahrscheinlichkeiten bei klinischen Studien und Fall-Kontroll-Studien
- Versicherungen: Modellierung der Schadenhäufigkeit und des Schadenbedarfs
  $\longrightarrow$ Marginalsummenmodell
- Banken: Identifikation von Risikofaktoren bei Kreditvergabe
- IT-Branche: Identifikation potentieller Kunden beim Direct Marketing
- CRM: Identifikation zufriedener bzw. unzufriedener Kunden

ANWENDUNGEN (2)



- klassisch: Y stetig, reellwertig, Normalverteilung. Präziser: $Y_i, 1 \leq i \leq n$, st.u., normalverteilt mit: $\mathsf{E}(Y_i) = x_i'\beta$, $\mathsf{Var}(Y_i) = \sigma^2$, $\beta \in \mathbb{R}^p$, $\sigma^2 > 0$
- Y relative Häufigkeiten: Binomialverteilung
- Y Anzahlen: Poissonverteilung
- Y stetig und positiv: Gammaverteilung

2. Generalisierte Lineare Modelle

Generalisierte Lineare Modelle (GLIMs) bestehen aus 3 Komponenten:

1. zufällige Komponente:

   unabhängige Zielgrößen $Y_i, 1 \leq i \leq n$ aus Exponentialfamilie mit
   $\mathrm{E}(Y_i) = \mu_i$ und $\mathrm{Var}(Y_i) = \phi V(\mu_i)/w_i$, wobei
   ϕ =Dispersionsparameter, V = Varianzfunktion, w_i bekannte Gewichte

2. Link-Funktion:

   es exist. monotone differenzierbare Link-Funktion g: $g(\mu_i) = \eta_i$.
   g^{-1} heißt inverse Link-Funktion.

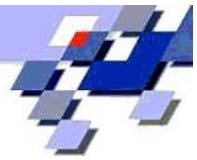
3. systematische Komponente (linearer Prädiktor):

   es exist. $\beta \in \mathbb{R}^p$: $\eta_i = x_i'\beta$

   Also: $\boxed{Y_i \text{ st.u., } \mathrm{E}(Y_i) = g^{-1}(x_i'\beta) \text{ und } \mathrm{Var}(Y_i) = \phi V(g^{-1}(x_i'\beta))/w_i}$

Beispiele für Generalisierte lineare Modelle

Modell	Skala von Y	Verteilung von Y	Link-Funktion	Varianz-Funktion
Klassische Regression und ANOVA	stetig	Normal	Identität: $\eta_i = \mu_i$	$V(\mu_i) = 1$
Logistische Regression	Anteil; 0/1	Binomial	Logit: $\eta_i = \Lambda^{-1}(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$	$V(\mu_i) = \mu_i(1-\mu_i)$
Generalisierte logist. Regression	nominal	Multinomial	GLOGIT $\eta_{i,j} = \log(\frac{\mathrm{P}(Y_i=j\|x_i)}{\mathrm{P}(Y_i=k_{ref}\|x_i)}) = x_i'\beta_j$	
Proportional Odds Regression	ordinal	Multinomial	CUMLOGIT $\eta_{i,j} = \log(\frac{\mathrm{P}(Y_i\leq j\|x_i)}{1-\mathrm{P}(Y_i\leq j\|x_i)}) = \gamma_j + x_i'\beta$	
Probit-Regression	Anteil; 0/1	Binomial	Probit: $\eta_i = \Phi^{-1}(\mu_i)$	$V(\mu_i) = \mu_i(1-\mu_i)$
Poisson-Regression	Anzahl	Poisson	$\eta_i = \log(\mu_i)$	$V(\mu_i) = \mu_i$
Gamma-Regression	positiv, stetig	Gamma	$\eta_i = 1/\mu_i$	$V(\mu_i) = \mu_i^2$
Gamma-Regression	positiv, stetig	Gamma	$\eta_i = \log(\mu_i)$	$V(\mu_i) = \mu_i^2$
Inverse Gauß-Regression	positiv, stetig	Inverse Gauß	$\eta_i = \mu_i^{-2}$	$V(\mu_i) = \mu_i^3$
Negative Binomial-Regression	Anzahl	Negative Binomial	$\eta_i = \log(\mu_i)$	$V(\mu_i) = \mu_i + k\mu_i^2$



Schätzen und Testen in GLIMs

- Klassischer Ansatz: McCullagh & Nelder '89
  ML Schätzer, Quasi-Likelihood-Schätzer
  Konfidenzintervalle nach Wald, Profile Likelihood. Tests: LR, Wald, Score

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

- exakte bedingte logistische Regression: bedingt bzgl. suffizienter Statistik
  Cox & Snell '89, Cytel: LogXact for Windows
- Robust:
  M-Schätzer: Pregibon '82, Künsch, Stefanski & Carroll '89, Cantoni & Ronchetti '01, ...
  mit hohem Bruchpunkt: Christmann '94, '98
- Generalized Estimating Equation (GEE): mehrere Messungen pro Einheit

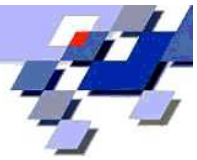
$$S(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i(\beta)'}{\partial \beta} \mathbf{V}_i^{-1}(\beta)(\mathbf{Y}_i - \mu_i(\beta)) = 0 \qquad \text{(Liang \& Zeger '86)}$$

3. GLIMs IN SAS

- PROC GENMOD
- PROC GAM
- PROC LOGISTIC, PROC PROBIT, PROC CATMOD, PROC PHREG
- SAS/INSIGHT

außerdem:
SAS/ANALYST, SAS/ASSIST, SAS/Enterprise Miner, SAS/IML,
PROC NLMIXED, PROC NLIN, ...

SAS/Macros: %ROC, %GLIMMIX, ...



4. VERGLEICH DER SAS-PROZEDUREN FÜR GLIMs

GENMOD	Allround-Prozedur, flexibel, schnell, LR-Tests, GEE
GAM	semiparametrisches Modell, Prüfung von GLIM-Modellannahmen
LOGISTIC	kategoriale Regression, Ausreißer-Erkennung, PL-Konfidenz-intervalle, Check auf Separation (Nichtexistenz des MLE), automatische Modellwahl, exakte bedingte logist. 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
SAS/INSIGHT	interaktiv, dynamische Graphiken, aber langsam

$\longrightarrow$ Beispiel-Programm: vergleich.sas

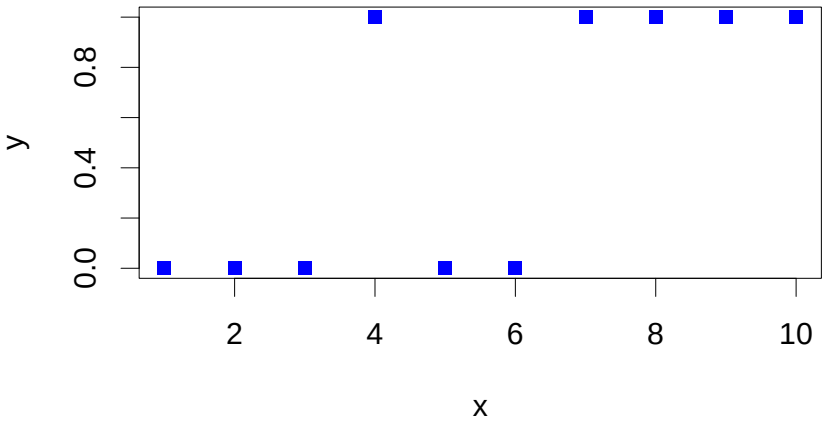
VERGLEICH DER SAS-PROZEDUREN (2)

1. Beispiel: logistische Regression mit Intercept

```
x:   1 2 3 4 5 6 7 8 9 10
y:   0 0 0 1 0 0 1 1 1  1
```

Overlap: ML-Schätzung existiert.

→ Beispiel-Programm: vergleich2.sas



PROC	Intercept $\hat{\beta}_0$	x $\hat{\beta}_1$	p-value für x		Bemerkung
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	

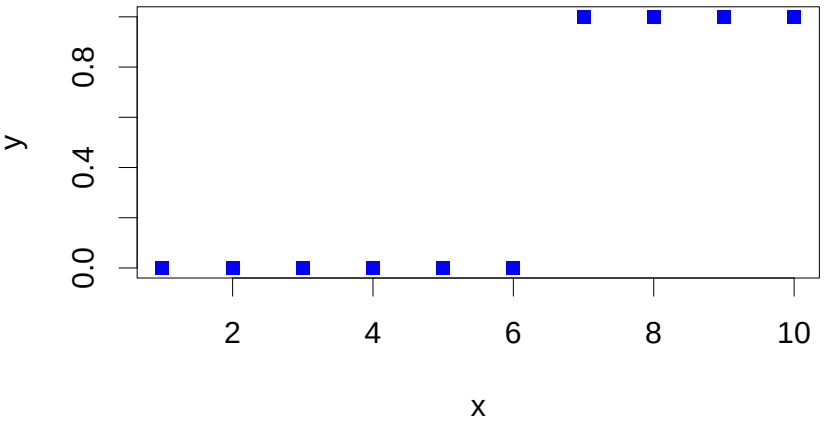
VERGLEICH DER SAS-PROZEDUREN (3)

2. Beispiel: logistische Regression mit Intercept

```
x:   1 2 3 4 5 6 7 8 9 10
y:   0 0 0 0 0 0 1 1 1  1
```

ML-Schätzung existiert NICHT !

→ Beispiel-Programm: vergleich2.sas



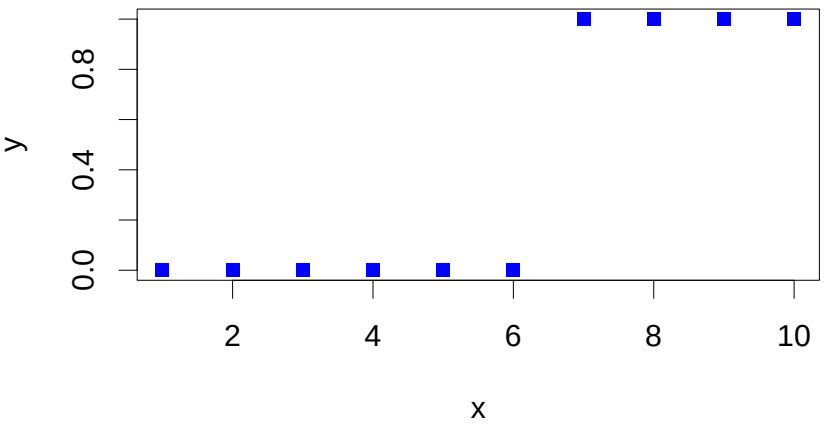
PROC	$\hat{\beta}_0$	$\hat{\beta}_1$	p-value für x		Bemerkung
GENMOD	-498.3	76.7	LR	<0.001	NOTE: Algorithm converged.
			Wald	1.000	
INSIGHT	-498.3	76.7	LR	<0.001	Convergence not attained in 40 iterations.
			Wald	<0.001	The validity of the model fit is questionable.
LOGISTIC	-65.1	10.0	LR	<0.001	WARNING: 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.
					NOTE: Parameters marked with '#' are regarded to be infinite.

VERGLEICH DER SAS-PROZEDUREN (4)

Fortsetzung des 2. Beispiels:

```
x:   1 2 3 4 5 6 7 8 9 10
y:   0 0 0 0 0 0 1 1 1  1
```

ML-Schätzung existiert NICHT !



PROC	Intercept $\hat{\beta}_0$	x $\hat{\beta}_1$	Intercept S.E.($\hat{\beta}_0$)	x S.E.($\hat{\beta}_1$)	p-value für x	
GENMOD	-498.3	76.7	$1.96 \cdot 10^{+9}$	$3.08 \cdot 10^{+8}$	LR	<0.001
					Wald	1.000
INSIGHT	-498.3	76.7	$6.76 \cdot 10^{-9}$	$4.32 \cdot 10^{-8}$	LR	<0.001
					Wald	<0.001

5. DETAILS ZU SAS-PROZEDUREN FÜR GLIMs

5.1 PROC GENMOD

```
PROC GENMOD ... ;
BY variables;
CLASS variables; ← diskrete Einflußgrößen
CONTRAST 'label' effect values ... ; ← LR-Test: H_0 : L'β = 0
ESTIMATE 'label' effect values ... ; ← Wald-Test: H_0 : l'β = 0
FREQ variable;
LSMEANS effects ...; ← LS-Mean L'β̂
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
```

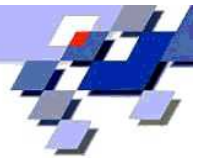
PROC GENMOD (2)

⚠ Default ab 8.1: bei DIST=BINOMIAL wird $P(Y = 0)$ modelliert (falls 0 kleinster Wert).

Beispiel: logistische Regression

```
PROC GENMOD DATA=a DESCENDING;
  MODEL y=x1 / DIST=BINOMIAL LINK=LOGIT; RUN;
```

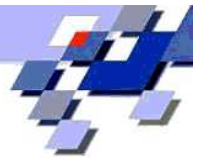
- Notwendig: PROC GENMOD, MODEL
- CLASS vor MODEL
- MODEL vor CONTRAST



PROC GENMOD (3):

Spezifikation der Verteilung im MODEL-Statement

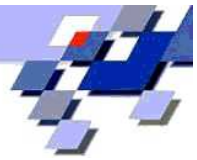
Spezifikation der Verteilung für Y DIST=		Default für Linkfunktion
NORMAL	Normal-Verteilung	μ
BINOMIAL	Binomial-Verteilung	logit, $\log(\mu/(1-\mu))$
MULTINOMIAL	Multinomial-Verteilung	cumulative logit
POISSON	Poisson-Verteilung	$\log(\mu)$
GAMMA	Gamma-Verteilung	$1/\mu$
IGAUSSIAN	Inverse Gauß-Verteilung	$1/\mu^2$
NEGBIN	Negative Binomial-Verteilung	$\log(\mu)$



PROC GENMOD (4):

Spezifikation der Link-Funktion im MODEL-Statement

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



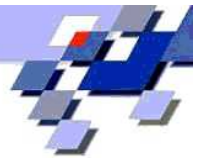
PROC GENMOD (5)

Flexibilität zur Anpassung zusätzlicher Modellklassen:

```
FWDLINK variable = expression;
INVLINK variable = expression;
DEVIANCE variable = expression;
VARIANCE variable = expression;
```

Beispiel: binäre Regression mit Linkfunktion g =Inverse der T(3)

```
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;
```



PROC GENMOD (6)

- \+ flexibel, relativ schnell
- \+ LR-Tests (Hauck-Donner-Phänomen !)
- \+ Generalized Estimating Equations (GEE) nach Liang & Zeger (1986): Modellierung mehrerer abhängiger Zielgrößen
- − bei logistischer Regression: kein Check auf Separation (vgl. PROC LOGISTIC)
- − keine automatische Modellwahl
- − kein EXACT-Statement
- − tabellarische Residualanalyse, aber keine Regression Diagnostics
- − keine robusten Schätzer (M-, S-, HBDP-, ...)

→ Beispiel-Programme: genmod.sas, gee.sas

5.2 PROC GAM : Generalisierte Additive Modelle (Hastie & Tibshirani, 1990)

- größere Modellklasse als GLIMs: Prüfung der GLIM-Modellannahmen
- GLIM: $\mathrm{E}(Y|X_1,\ldots,X_p) = g^{-1}(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)$
- GAM: $\mathrm{E}(Y|X_1,\ldots,X_p) = g^{-1}(\beta_0 + f_1(X_1) + \ldots + f_p(X_p))$

f_j unbekannte Funktionen, nichtparametrische Schätzung

```
PROC GAM <option>;
CLASS variables;
MODEL dependent = <PARAM(effects)> smoothing effects ... ;
SCORE data=SAS-data-set out=SAS-data-set; ← Prognosen
OUTPUT <out=SAS-data-set> keyword ... ;
BY variables;
ID variables;
FREQ variable;
```

PROC GAM (2)

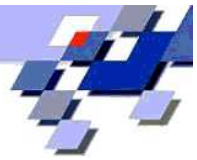
- Notwendig: PROC GAM und MODEL
- SCORE-Statement darf mehrmals vorhanden sein
- Alle anderen Statements maximal einmal.

Glättungsmethoden

`PARAM(variable)`	parametrisch
`SPLINE(variable <, df=number>)`	smoothing spline
`LOESS(variable <, df=number>)`	local regression
`SPLINE2(variable, variable <,df=number>)`	bivariate thin-plate smoothing spline

⚠ Evtl. notwendig: DATA-Step zur Berechnung der geschätzten Werte (falls Link-Funktion $\neq$ Identität)
Grund: Geschätzer Wert `P_Y` ist $x_i'\hat{\beta}$, nicht $g^{-1}(x_i'\hat{\beta})$!
Dies ist anders als in PROC GENMOD !

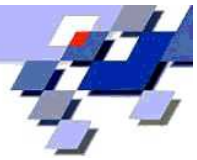
⚠ Default: Bei DIST=BINOMIAL wird $P(Y = 1)$ modelliert.



PROC GAM (3)

- \+ flexibel durch smoothing
- \+ Bivariate Spline-Methode implementiert, aber sehr langsam, relativ alt
- − Smoothing mit `LOESS` ist sehr langsam; z.T. instabil unter SAS 8.2 TS2M0
- − nur Standard-Glättungsmethoden, nicht: isoton, antiton, konkav, konvex
- − Verteilungen für Y laut Handbuch + PDF-Dokumentation: NORMAL, BINOMIAL, POISSON, GAMMA; laut Online-Help auch IGAMMA (???)
- − nur kanonische Link-Funktionen
- − keine automatische Modellwahl

→ Beispiel-Programm: gam.sas



5.3 PROC LOGISTIC

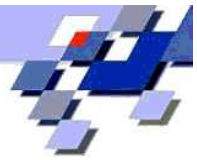
```
PROC LOGISTIC <options>;
BY variables;
CLASS variable <(v-options)> ...;
CONTRAST 'label' effect values ,... ;
EXACT <'label'> <Intercept> <effects> ...;
FREQ variable;
MODEL events/trials = <effects> ... ;
MODEL variable <(variable_options)> = <effects> ...;
OUTPUT <OUT=SAS-data-set> <keyword=name ...> ;
<label:> TEST equation1 , ...  ;
UNITS independent1 = list1 ... ;
WEIGHT variable ...;
```

← Wald-Test: $H_0 : L'\beta = 0$

← exakte bed. log. Regr.

← Wald-Test: $H_0 : L'\beta = c$

← Einheiten für Odds Ratios



PROC LOGISTIC (2)

- CLASS vor MODEL, MODEL vor CONTRAST und EXACT
- ⚠ Default: es wird $P(Y = 0)$ modelliert.

  `PROC LOGISTIC DATA=a DESCENDING; MODEL y=x1/LINK=LOGIT; RUN;`
- ⚠ ML-Schätzung existiert nicht für alle Datensätze bei logist. Regression:

  Albert & Anderson '86, Christmann & Roussseeuw '01, Christmann, Fischer, & Joachims '02
- Modellwahl: SELECTION=NONE, BACKWARD, FORWARD, STEPWISE, SCORE

  Score-Tests, nicht LR-Tests
- Ausreißererkennung & Regression diagnostics (leider) basierend auf ML-Schätzer: MODEL-Options INFLUENCE und IPLOTS (Pregibon '81)

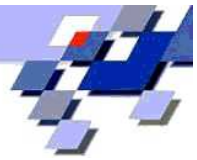
  Bessere Alternativen: robuste Schätzer

  Künsch, Stefanski & Carroll '89, Christmann '98, Rousseeuw & Christmann '02, ...

PROC LOGISTIC (3)

- Exakte bedingte logistische Regression: Verteilung von $Y|(X'Y = X'y)$ (bedingt bzgl. suffiziente Statistik) hängt unter H_0 *nicht* von $\beta \in \mathbb{R}^p$ ab.
  Nicht verwendet: CONTRAST, TEST, UNITS
  Keine Berechnungen: WEIGHT, LINK≠LOGIT, OFFSET, NOFIT, SELECTION

  `ALPHA=0.05` $\leftarrow 100(1-\alpha)\%$ Konfidenzintervalle
  `ESTIMATE=PARM oder ODDS oder BOTH.` ⚠ $p_{two-sided} = 2p_{one-sided}$
  `JOINT` $\leftarrow H_0 : \beta = 0$ und für alle x_i: $H_0^i : \beta_{i \in I} = 0$
  `JOINTONLY` $\leftarrow H_0 : \beta = 0$
  `ONESIDED` $\leftarrow p_{one-sided} = \min\{p_{left}, p_{right}\}$
  `OUTDIST=SAS-data-set` $\leftarrow$ Abspeichern der bedingten Verteilung



PROC LOGISTIC (4)

- Exakte bedingte logistische Regression: oft langsam, nur für kleine Datensätze !

  Simulierter Datensatz

  mit $n = 121$, $p = 3$, $\beta = (-5, 1, 1)'$, Intercept und $x_1, x_2 \in \{-5, \ldots, 5\}$:

SAS 8.2/PROC LOGISTIC	30 Sekunden
Cytel/LogXact 2.0	17 Sekunden

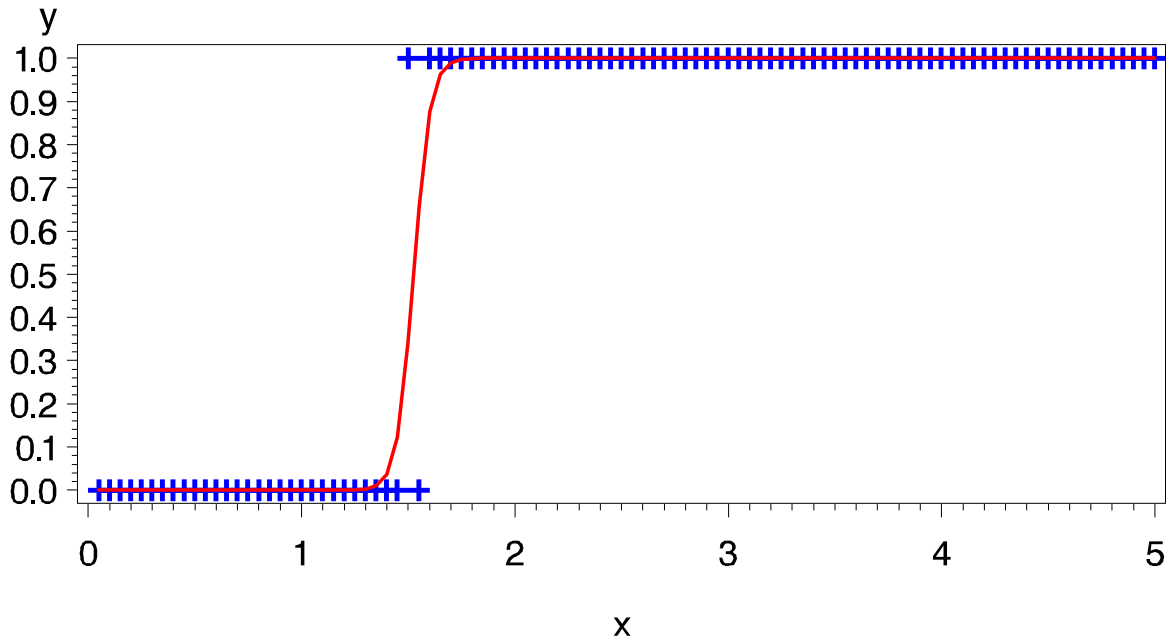
→ Beispiel-Programm: logistic.sas

PROC LOGISTIC (5)

⚠ Wald-Tests können bei binärer Regression irreführend sein !

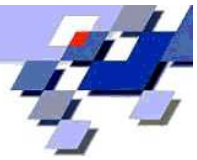
Hauck & Donner (1977, JASA): Erklärende Variablen ohne oder mit Einfluß auf die Zielvariable können zu großen p-Werten führen !

Beispiel: logist. Regr. $\mathrm{P}(Y_i = 1|x_i) = \Lambda(-25 + 17x_i)$, $1 \leq i \leq 100$



Wald-Test: $p = 0.113$, aber LR-Test: $p < 0.001$ (PROC GENMOD)

→ Beispiel-Programm: HauckDonner.sas



PROC LOGISTIC (6)

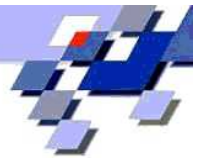
- ROC-Analyse: MODEL-Option OUTROC=data set
- Vergleich mehrerer ROC-Kurven: SAS-Macro %ROC von D.M. DeLong

  DeLong, DeLong, Clarke-Pearson (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics, 44, 837-845.

  `http://ftp.sas.com/techsup/download/stat/roc.html`

  Anwendung: Projekt mit den Städt. Kliniken Dortmund.

  Vergleich der Güte digitaler und traditionelle Roentgenbilder des Brustbereichs mit Goldstandard Computer-Tomographie



PROC LOGISTIC (7)

- \+ Existenz-Prüfung der ML-Schätzung
- \+ automatische Modellwahl möglich
- \+ ROC-Analyse
- \+ exakte bedingte logistische Regression
- − keine LR-Tests. Wald-Tests können extrem irreführend sein !
  $\longrightarrow$ Hauck-Donner-Phänomen
- − ML-basierte Ausreißererkennung und Regression Diagnostics
- − keine robusten Methoden
- − kein Vergleich mehrerer ROC-Kurven

5.4 PROC PROBIT

Biometrie:

- Inverse Regression und Konfidenzgrenzen (EDγ) mit Option `INVERSECL`:
  Anwendung: Schätzung der Dosis x_0, für die gilt: $P(Y = 1|X = x_0) = \gamma$.
- Schätzung einer spontanen Ereignis-Wahrscheinlichkeit: `OPTC` und `C=`
  Anwendung: Baseline $\lim_{x \to -\infty} P(Y = 1|X = x) = c > 0$.

```
PROC PROBIT <options> ;
MODEL response=independents < / INVERSECL OPTC C= ... > ;
BY variables ;
CLASS variables ;
OUTPUT <OUT=SAS-data-set> <options> ;
WEIGHT variable ;
Plots mit: CDFPLOT, INSET, IPPPLOT, LPREDPLOT, PREDPPLOT
```

→ Beispiel-Programm: probit.sas

5.5 PROC CATMOD

```
PROC CATMOD <options> ;
DIRECT <variables> ;
MODEL response-effect=design-effects </options>;
CONTRAST 'label' row-description ... ;
BY variables;
FACTORS factor-description ... ;
LOGLIN effects;
POPULATION variables;
REPEATED factor-description ... ;
RESPONSE function ... ;
RESTRICT parameter=value ... ;
WEIGHT variable;
```

← stetige Einflußgrößen

← Test $H_0 : L\beta = 0$

← log-lineare Modelle

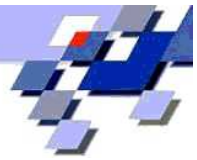
← repeated measurements

← Response Funktionen

← setzt Parameter auf Konstanten

PROC CATMOD (2)

- Notwendig: PROC CATMOD, MODEL
- DIRECT vor MODEL
- MODEL VOR CONRTRAST
- nur ein LOGLIN, REPEATED oder FACTOR zwischen 2 RUNs
- QUIT zum sicheren Beenden
- ⚠ interaktive Nutzung möglich: zwischen 2 RUN-Statements werden mehrere CONTRAST- und RESPONSE-Statements ausgeführt, aber nicht andere Statements.
  Beispiel: bei 2 LOGLIN-Statements wird das erste ignoriert.
- nur Wald-Tests für erklärende Variablen



PROC CATMOD (3): Response-Funktionen

1 erklärende Variable x mit Werten 1, 2 oder 3

Statement	q	Response-Funktion
Default (general. LOGITS)	2	$\log(p_1/p_3)$, $\log(p_2/p_3)$
`ALOGITS`	2	$\log(p_2/p_1)$, $\log(p_3/p_2)$
`CLOGITS`	2	$\log(\frac{1-p_1}{p_1}$, $\log(\frac{1-(p_1+p_2)}{p_1+p_2})$
`JOINT`	2	p_1, p_2
`LOGITS`	2	$\log(p_1/p_3)$, $\log(p_2/p_3)$
`MARGINAL`	2	p_1, p_2
`MEAN`	1	$1p_1 + 2p_2 + 3p_3$

Viel komplexer bei Wechselwirkungen, vgl. SAS-Manual SAS/STAT !

→ Beispiel-Programm: catmod.sas

5.6 PROC PHREG

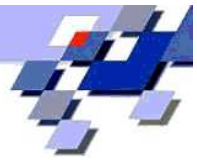
Biometrie: Fall-Kontroll-Studien bei m:n Matching

vgl. SAS/STAT, Example 23.3

```
DATA lbw; INPUT id age low lwt smoke ht ui @@; time=2-low; CARDS;
 25   16   1   130   0    0   0        143   16   0   110   0    0   0
...
;
PROC PHREG DATA=lbw;
MODEL time*low(0) = lwt smoke ht ui /  TIES=DISCRETE RISKLIMITS ALPHA=0.05;
STRATA age; RUN;
```

Kontrollen sind mit 0 kodiert, deshalb `low(0)`.

Bindungen: `TIES=EXACT, EFRON, DISCRETE, BRESLOW`



5.7 SAS/INSIGHT

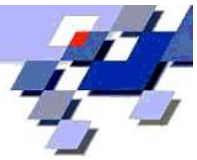
- \+ interaktive Datenanalyse, Syntax-Verwendung möglich
- \+ dynamische und verbundene Graphiken
- \+ QQ-Plots, Partial Leverage Plots
- \+ LR-Tests und LR-Konfidenzintervalle
- \+ übliche Link-Funktionen
- − langsam für große Datensätze
- − Verteilung der Response-Variablen Y:
  NORMAL, BINOMIAL, POISSON, GAMMA, INVGAUSSIAN
  nicht: MULTINOMIAL, kein Proportional Odds Modell, NEGBIN
- − kein GEE

→ Beispiel-Programm: insight.sas

SAS/INSIGHT (2)

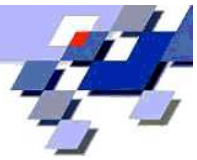
```
PROC INSIGHT < INFILE=fileref > < FILE<=fileref> >
  < DATA=SAS-data-set > < TOOLS >
  < NOMENU > < NOBUTTON > < NOCONFIRM >;

FIT variable-list < = effects-list >
  < / < FREQ=variable > < WEIGHT=variable >
  < LABEL=variable > < NOINT >
  < RESP=response > < BINOM=variable >
  < OFFSET=variable > < LINK=link >
  < POWER=value > < NOEXACT > < FISHER >
  < QUASI > < SCALE=scale > < CONSTANT=value > >;
```



WEITERE BEISPIELE

Bioassay	→ bioassay.sas
Proportional Odds Modell	→ propodds.sas
Gamma- u. Poisson-Regression	→ GammaPoisson.sas
GEE, bivariate Poisson-Regression	→ gee.sas
Macro für robuste Schätzung (positiver Bruchpunkt) bei großen Gruppen	→ hbdp.sas
Macro für robuste Schätzung (logist. Regression)	→ wemel.sas



6. VERGLEICH MIT ANDEREN SOFTWARE-PRODUKTEN (bzgl. GLIMs)

SPSS vs. SAS

- SAS bietet mehr Funktionalität und ist schneller für große Datensätze
- SPSS 11.0 nur für kategoriale Regression (logistisch etc.)

S-PLUS vs. SAS

- SAS gut für große Datensätze
- SAS verwendet leider nur nicht-robuste Schätzer (MLE, Quasi-MLE, WLS, ...)

  Problem: Masking und Swamping Effekte
- S-PLUS Library Robust: Methoden basierend auf ML-Schätzern, aber auch M-Schätzer

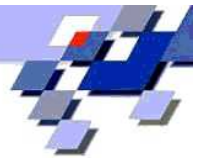
  (aber ohne hohen Bruchpunkt) für logistische und Poisson-Regression
- bzgl. GAMs: S-PLUS flexibler: andere nichtparametrische Glättungsverfahren möglich

LogXact vs. SAS

- SAS ist oft langsamer für exakte logistische Regression

7. ZUSAMMENFASSUNG

- \+ SAS bietet viele Prozeduren für generalisierte lineare Modelle
- \+ SAS Prozeduren oft relativ schnell bei großen Datensätzen
- \+ gute Dokumentation (im Vergleich zu Konkurrenzprodukten)
- \+ Prüfung der Modellannahmen durch PROC GAM möglich
- − oft unterschiedliche Parametrisierung kategorieller Einflußgrößen x_i
  $\Rightarrow$ Fehlinterpretation der Ergebnisse möglich; evtl. eigene Kodierung
- − Unterschiede bei Defaults für binäre Regressionsmodelle:
  Modellierung von $P(Y = 0)$ oder $P(Y = 1)$? $\Rightarrow$ Fehlinterpretation möglich
- − Ergebnisse verschiedener PROCs manchmal inkonsistent (bei gleicher Parametr.)
- − Plots für Regression Diagnostics nur z.T. implementiert, i.a. nicht automatisch
- − keine robusten Schätz- und Test-Verfahren



8. LITERATUR

Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley, New York.

Albert, A., Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71, 1-10.

Cantoni, E., Ronchetti, E. (2001). Robust inference for generalized linear models. JASA,96,1022-1030.

Chen, C. (2002). Robust tools in SAS. In: Dutter et al. Developments in robust statistics. Physika, 125-133.

Christmann, A. (1998). On positive breakdown point estimators in regression models with discrete response variables. Habilitationsschrift. Universität Dortmund.

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.

Christmann, A., Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. Computational Statistics and Data Analysis, 37, 65-75.

Cox, D.R., Snell, E.J. (1989). Analysis of Binary Data. 2nd ed. Chapman & Hall, London

Hauck, Jr., W.W., Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. JASA, 72, 851-853.

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.

LogXact for Windows, User Manual, Cytel Software Corporation (1996)

McCullagh, P., Nelder, J.A. (1989). Generalized Linear Models. 2nd ed. Chapman & Hall, London

Rousseeuw, P.J., Christmann, A. (2002). Robustness against separation and outliers in logistic regression. Erscheint in: Computational Statistics & Data Analysis.

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.

SAS/INSIGHT User's Guide (1993). SAS Institute Inc.

SAS/STAT Users Guide I und II (1990). SAS Institute Inc.

SAS/STAT Software: Changes and Enhancements, through Release 6.11 (1996). SAS Institute Inc.

SAS/STAT Software: Changes and Enhancements, Release 8.2 (2001). SAS Institute Inc.

Stokes, M.E., Davis, C.S., Koch, G. (1995). Categorical data analysis using the SAS System, SAS Institute Inc.