Logistische Regression in SAS®

Oliver Kuß
Medizinische Universitätsklinik,
Abt. Klinische Sozialmedizin,
Bergheimer Str. 58, 69115 Heidelberg,
email: okuss@med.uni-heidelberg.de

3. Konferenz für SAS®-Anwender
in Forschung und Entwicklung (KSFE)
25. - 26. Februar 1999
Ruprecht Karls-Universität Heidelberg

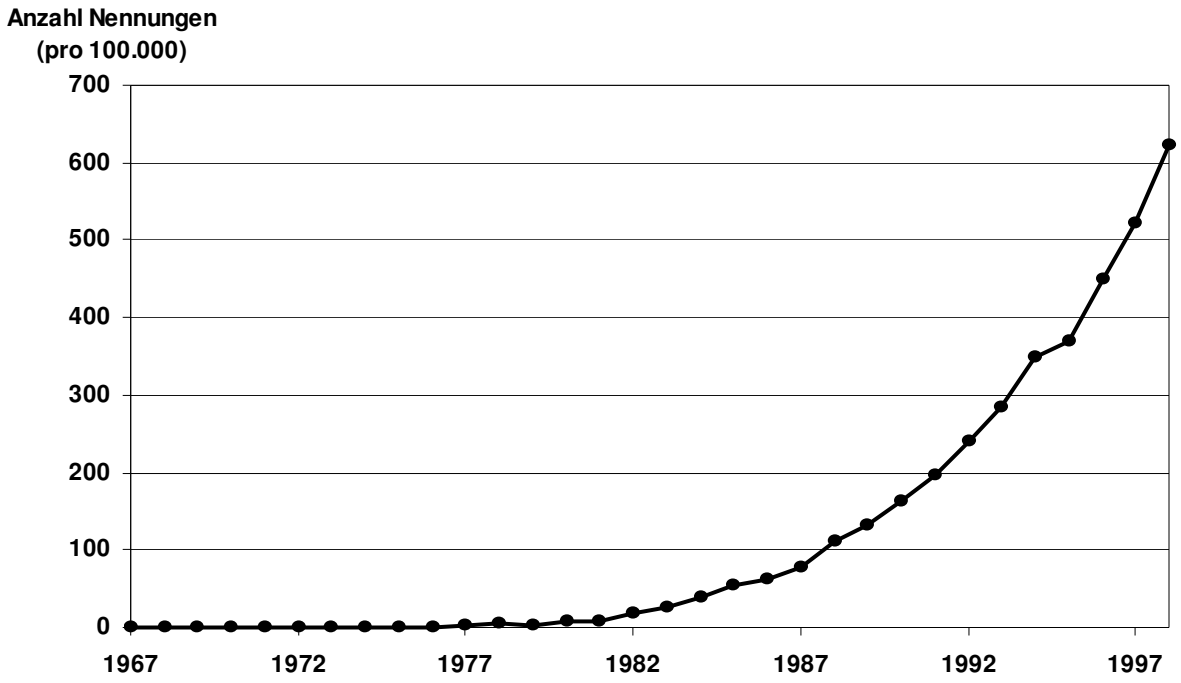
1. Die Einleitung

Standardsatz:

„Das logistische Regressionsmodell hat sich seit seiner Einführung in den siebziger Jahren zu einer Standardmethode in der Biometrie und Epidemiologie entwickelt, wenn es um die Auswertung von binären Zielgrößen geht.“

Beweis:

Resultate einer MEDLINE-Suche nach „Logistic Regression“ in Abstract oder Keyword (adjustiert nach der Gesamtzahl der publizierten Artikel)



Aber:
Logistische Regression wird nicht nur in der Biometrie und der Epidemiologie verwendet.

Andere Disziplinen:
Ökonomie, Informationstechnik, Biologie, Linguistik, Psychologie, Ökologie, Soziologie, Geowissenschaften, Bevölkerungswissenschaft, Politische Wissenschaft

Gründe für die wachsende Beliebtheit:
- Interpretierbarkeit der geschätzten Parameter als Odds Ratios
- Wahrscheinlichkeiten für das Eintreten des Zielereignisses können geschätzt werden
- Anwendung in prospektiven und retrospektiven Designs
- Verfügbarkeit von geeigneter Software

2. Das Modell

Logistische Regression beschreibt den Zusammenhang zwischen einer kategoriellen Zielgröße und einer Menge von erklärenden Variablen.

Für eine **binäre** Zielgröße hat das Modell die Form

$$\text{logit}(p_i) = \log\left(p_i \big/ 1 - p_i\right) = \alpha + \beta' x_i$$

mit

p_i: Wahrscheinlichkeit für das Eintreten des Zielereignisses $p_i = p(y_i = 1 \mid x_i)$,

α: Intercept-Parameter,

β: Vektor von Steigungsparametern,

x_i: Vektor von Kovariablen

Mögliche Erweiterungen:

- Andere Linkfunktion (Probit, Gompit)
- Zielgröße mehrkategoriell (nominal, ordinal)
- Beobachtungen nicht mehr unabhängig

3. Der Beispieldatensatz

Stichprobe: 162 Frauen mit unerfülltem Kinderwunsch
Zielgröße: Schwangerschaft
Erklärende Variablen:

- Alter (in Jahren),
- Dauer der Infertilität (in Jahren),
- Eileiterdefekt

Ergebnis:

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCEPT	2.0117	1.3734	2.1456	0.1430	.
AGE	-0.0510	0.0422	1.4647	0.2262	0.950
INFER	-0.1409	0.0791	3.1735	0.0748	0.869
TUBPHYSD	-0.8880	0.4284	4.2973	0.0382	0.411

4. Die einzelnen Prozeduren

4.1 PROC LOGISTIC

```
proc logistic data=pregnanc;
       model nready/ntotal=age infer tubphysd;
run;
```

- Kein CLASS-Statement
- Interaktionsterme können nicht im MODEL-Statement angegeben werden
- Umfangreiche Residuen-Analyse
- Hosmer-Lemeshow-Test
- ROC-Analyse
- Variablen-Selektionsmethoden
- Adjustierung nach Overdispersion
- Bias-adjustierte geschätzte Wahrscheinlichkeiten (CTABLE-Option)
- Andere Modelle können gefittet werden (Proportional Odds Modell, Bedingte logistische Regression, Bradley-Terry-Modell)

4.2 PROC GENMOD

```
proc genmod data=pregnanc;
       class tubphysd;
       model nready/ntotal=age infer tubphysd / dist=bin link=logit;
run;
```

- CLASS-Statement
- Adjustierung nach Overdispersion
- MAKE-Statement und Output Delivery System
- GEE-Methode implementiert (REPEATED-Statement) für korrelierte Beobachtungen

4.3 PROC PROBIT

```
proc probit data=pregnanc;
       class tubphysd;
       model nready/ntotal=age infer tubphysd / d=logistic;
  run;
```

- CLASS-Statement
- Standard-Linkfunktion: Probit
- Ordinale Zielgrößen

4.4 PROC CATMOD

```
proc catmod data=pregnant order=data;
       direct age infer;
       model pregnant=age infer tubphyst;
run;
```

- Stetige Kovariablen müssen explizit angegeben werden (DIRECT-Statement)
- Andere Parametrisierung, deshalb Odds Ratios für kategorielle Kovariablen nur auf Umwegen
- Multinomiale logistische Regression (nominale und ordinale Zielgrößen)
- Bedingte logistische Regression
- Korrelierte Beobachtungen
- WLS-Methode

4.5 PROC NLIN

```
proc nlin nohalve sigsq=1 data=pregnanc (rename=(age=_old1 infer=_old2
tubphysd=_old3));

      parms intercpt=0 age=0 infer=0 tubphysd=0;

      _y_=intercpt + age*_old1 + infer*_old2 + tubphysd*_old3 ;
            if _iter_=-1 then do;
                  _mu_=0;
                  _loss_ = 0;
                  if nready=0 then nready=0.1;
                  if nready=ntotal then nready=ntotal-0.1;
                  _weight_=nready*(ntotal-nready)/ntotal;
                  nready=log(nready/(ntotal-nready));
            end;

            else do;
                  _mu_=exp(_y_);
                  _der_=_mu_/(_mu_+1)**2;
                  _mu_=_mu_/(1+_mu_);
                  _der_=_der_*ntotal;
                  _y_ = _mu_;
                  _mu_=ntotal*_y_;
                  _weight_=1/(ntotal*_y_*(1-_y_));
                  _loss_=(-nready*log(_y_) -(ntotal-nready)*log(1-
                  _y_))/_weight_;
            end;

      model nready=_mu_;

      der.intercpt=_der_;
      der.age=_der_*_old1;
      der.infer=_der_*_old2;
      der.tubphysd=_der_*_old3;

run;
```

- Etwas speziell, eher als Ausgangspunkt zur Berechnung von komplizierteren Modellen

4.6 PROC IML

```
* IRLS-Algorithmus zur Berechnung der Parameter-Schätzer;
b = repeat(0,ncol(x),1); oldb=b+1;
do iter=1 to 20 while(max(abs(b-oldb))>1e-8);
      oldb=b;
      p=1/(1+exp(-(x*b)));
      f=p#p#exp(-(x*b));
      loglik =sum( ((y=1)#log(p) + (y=0)#log(1-p))#wgt);
      btransp = b`;
      w = wgt/(p#(1-p));
      xx = f # x;
      xpxi = inv(xx`*(w#xx));
      step = xpxi*(xx`*(w#(y-p)));
      b = b + step;
end;

* Berechnung des Deviance-Tests auf Null-Einfluß aller Kovariablen
gemeinsam;
     p0 = sum((y=1)#wgt)/sum(wgt);  /* average response */
     loglik0 =sum( ((y=1)#log(p0) + (y=0)#log(1-p0))#wgt);
     chisq =  ( 2 # (loglik-loglik0));
     df    = ncol(x)-1;
     prob  = 1-probchi(chisq,df);
     print ,'Likelihood Ratio with Intercept-only Model' chisq df prob,;

* Wald-Test auf Null-Einfluß der Kovariablen separat;
     stderr = sqrt(vecdiag(xpxi));
     tratio = b/stderr;
     print ,'Wald-Tests fuer die Parameter' parm b stderr tratio,;
```

- Voller Zugriff auf alle berechneten Größen, ideal zur Weiterverarbeitung

5. Die Bugs?

5.1 Prüfung auf Existenz der ML-Schätzer

Separation im Raum der Kovariablen ⇔
Nichtexistenz der Parameter-Schätzer

Separation: Existenz einer Hyperebene im Raum der Kovariablen, so daß diese die Beobachtungen mit Y=0 von denen mit Y=1 trennt.

Beispiel: 1 Kovariable → Hyperebene ist ein Punkt

	Kovariable	Zielgröße
1	-0.95570	0
2	-0.36591	0
3	-0.27472	0
4	-0.27403	0
5	0.24945	1
6	0.39787	1
7	0.62435	1
8	0.70329	1
9	0.71933	1
10	0.90918	1

Keine einzige Prozedur diagnostiziert die Separation und liefert eine Warnung!!

5.2 GOF-Tests in PROC GENMOD

- EVENT/TRIAL-Syntax

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	104	113.1736	1.0882
Scaled Deviance	104	113.1736	1.0882
Pearson Chi-Square	104	97.5790	0.9383
Scaled Pearson X2	104	97.5790	0.9383
Log Likelihood	.	-91.4276	.

- ACTUAL-TRIAL-Syntax

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	158	182.8552	1.1573
Scaled Deviance	158	182.8552	1.1573
Pearson Chi-Square	158	159.0097	1.0064
Scaled Pearson X2	158	159.0097	1.0064
Log Likelihood	.	-91.4276	.

6. Das Fazit

- SAS® bietet eine Vielzahl von Möglichkeiten, logistische Regressionsmodelle zu fitten
- Auswahl der Prozeduren ist abhängig vom Modell, für Standardanwendungen sind aber PROC LOGISTIC und PROC GENMOD die Methoden der Wahl, alle anderen vorgestellten Prozeduren sind in andere Richtungen spezialisiert