Logistische Regression in SAS einmal anders

Oliver Kuß

Institut für Medizinische Epidemiologie,
Biometrie und Informatik,
Universität Halle-Wittenberg

KSFE Potsdam, 21. Februar 2003

Gliederung

- Das logistische Regressionsmodell
- Der Beispieldatensatz
- Die Standardprozeduren
- Die Nicht-Standardprozeduren
- Fazit

Das logistische Regressionsmodell I

= das Standardmodell zur Regressionsanalyse binärer Zielgrößen

Gründe:

- Anschauliche Interpretation der Parameter
- Prognosen für das Eintreten des Zielereignisses sind möglich
- Gültig unter prospektivem und retrospektivem Sampling
- Kein Problem mit der Software (PROC LOGISTIC, PROC GENMOD, PROC PROBIT, PROC CATMOD)
- **Mitglied von zahlreichen Modellklassen**

Das logistische Regressionsmodell II

Modellgleichung:

$$\text{logit}(\pi_i) \;=\; \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

mit

$$\begin{array}{rcl} \pi_i & = & \text{p}(Y_i = 1), \\ \text{logit}(\pi_i) & = & log\frac{\pi_i}{1-\pi_i}, \\ Y_i & = & \text{Zielgröße}, \\ x_1, \ldots, x_p & = & \text{Kovariablen}, \\ i & = & 1, \ldots, N, \end{array}$$

$$Y_i \;\sim\; \text{Binomial}(1, \pi_i)$$

Schätzung der Parameter i.a. durch ML (IRLS-Algorithmus)

Der Beispieldatensatz

aus Zusammenarbeit mit Dr. J. Rimbach, Universitätsfrauenklinik Heidelberg

- Stichprobe: 162 Frauen mit unerfülltem Kinderwunsch

- Zielgröße: Schwangerschaft (innerhalb der ersten drei Jahre)

- Kovariablen: Alter (in Jahren), Dauer der Infertilität (in Jahren), Eileiterdefekt (binär)

Ergebnisse:

Variable	$\widehat{\beta}$	p-Wert	Odds Ratio
Intercept	2.012	0.143	–
AGE	-0.051	0.226	0.950
INFER	-0.141	0.075	0.869
TUBPHYSD	-0.888	0.038	0.411

Die Standardprozeduren

Die Daten:

```
data pregnant;
     input age infer tubphysd pregnant;
     cards;
     25      2          0          1
     27      2          1          0
     27      3          1          0
     27      4          1          0
     28      2          1          1
            ...
;run;
```

Die Prozeduren:

```
proc logistic data=pregnant descending;
     model pregnant=age infer tubphysd;
run;

proc genmod data=pregnant descending;
     model pregnant=age infer tubphysd
                    / link=logit d=bin;
run;
```

Die Nicht-Standardprozeduren I

PROC QLIM

QLIM = Qualitative and LIMited Dependent Variable Model (Experimental Procedure seit Version 8.0)

```
proc qlim data=pregnant;
     model pregnant=age infer tubphysd /
                    type=blogit
                    covest=hessian;
     endogenous discrete=(pregnant 0 1);
   * hetero age / link=exp square;
run;
```

Eigenschaften:

Verschiedene Schätzer der Kovarianzmatrix der Parameter möglich (COVEST= HESSIAN/ OP/ QML)

Mit Hilfe des HETERO-Statements können heteroskedastische Residuen modelliert werden

Eine Fülle (8) von R^2-Maßen wird ausgegeben:

Likelihood Ratio (R)	12.278	2 * (LogL - LogL0)
Upper Bound of R (U)	195.13	- 2 * LogL0
Aldrich-Nelson	0.0704	R / (R+N)
Cragg-Uhler 1	0.0730	1 - exp(-R/N)
Cragg-Uhler 2	0.1042	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.0753	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.0263	1 - ((LogL-K)/LogL0)^ (-2/N*LogL0)
McFadden's LRI	0.0629	R / U
Veall-Zimmermann	0.1289	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.3283	

Die Nicht-Standardprozeduren II

PROC MODEL

Das logistische Regressionsmodell kann auch als nicht-lineares Modell interpretiert werden:

Modellgleichung:

$$Y_i \;=\; \text{expit}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) + \epsilon$$

mit expit(a)=$\frac{exp(a)}{1+exp(a)}$

```
proc model data=pregnant;
     pregnant=   exp(b0 + bage*age + binfer*infer
                       + btubphys*tubphysd)/
               (1+exp(b0 + bage*age + binfer*infer
                       + btubphys*tubphysd));
     parms b0 2 bage 0 binfer 0 btubphys -0.5;
     fit pregnant / GMM KERNEL=(PARZEN,0,0)
                    COLLIN
                    DW GODFREY=3
                    WHITE BREUSCH=(infer tubphysd);
run;
```

Eigenschaften:

- Fülle von Schätzverfahren möglich (OLS/ GMM/ FIML/ SUR/ N2SLS/ N3SLS)

- Ausgabe von Kollinearitäts-Statistiken (COLLIN, bisher nur in SAS/INSIGHT implementiert)

- Tests für Autokorrelation in den Residuen (DW, GODFREY)

- Tests für Heteroskedastizität in den Residuen (WHITE, BREUSCH=(var))

Die Nicht-Standardprozeduren III

PROC GAM

Das logistische Regressionsmodell ist auch ein Spezialfall der generalisierten additiven Modelle

```
proc gam data=pregnant;
     model pregnant= param(age infer tubphysd)
                    / dist = logist;
     score data=neu out=scoreout;
run;
```

Eigenschaften:

- SCORE-Statement erlaubt Prognosen für neue Beobachtungen

- Eigentlich erst interessant für nicht-parametrische Modelle

Die Nicht-Standardprozeduren IV

Sonstige

Das logistische Regressionsmodell als Spezialfall der gemischten Modelle (PROC NLMIXED, %GLIMMIX, %NLINMIX)

Das logistische Regressionsmodell als Spezialfall des bedingten logistischen Regressionsmodells (PROC PHREG, PROC MDC)

Fazit

- SAS bietet eine Vielzahl von Möglichkeiten (auch jenseits der Standardprozeduren), logistische Regression zu rechnen

- Anwendung von Nicht-Standard-Prozeduren liefert eine Reihe von zusätzlichen Erkenntnissen (sowohl praktischer als auch theoretischer Natur)

- Urteilen Sie selbst: Echter Nutzen oder Akademische Spinnerei?