Bedingte logistische Regression mit PROC LOGISTIC: Möglichkeiten seit SAS V9 und Vergleich zu PROC PHREG

Rainer Muche, Bettina Danner
Institut für Biometrie
Schwabstraße 13
89075 Ulm
rainer.muche@uni-ulm.de

Zusammenfassung

Die bedingte logistische Regression ist die Standardmethode zur Auswertung von 1:m gematchten Fall-Kontrollstudien u. a. zur Identifizierung von Risikofaktoren für relativ seltene Erkrankungen. Bis SAS 8 konnte man die bedingte logistische Regression für das 1:m - Matching nur in der Prozedur PHREG ausführen. Dabei musste man in dieser für die Proportional Hazard (Cox) Regression vorgehaltenen Prozedur einige Tricks anwenden, um eine ML-Schätzung für die bedingte logistische Regression zu erhalten. Seit SAS Version 9 gibt es die bedingte logistische Regression auch in der Prozedur PROC LOGISTIC. Durch die Angabe des neuen Statements STRATA in PROC LOGISTIC lässt sich nun eine bedingte logistische Regression ausführen, wobei in diesem Strata-Statement die Variable angegeben werden muss, die die jeweils (m+1) zusammengehörigen Kontrollen und Fälle kennzeichnet. Durch diese Einbindung in PROC LOGISTIC lassen sich nun viele aus der Analyse unbedingter logistischer Regressionen bekannten Möglichkeiten nutzen, z.B. Class-Statement, Hierarchie-Prinzip, exakte logistische Regression.

Schlüsselwörter: bedingte logistische Regression, PROC LOGISTIC, PROC PHREG

1 Einleitung

Die bedingte logistische Regression ist die Standardmethode zur Auswertung von 1:m gematchten Fall-Kontrollstudien u. a. zur Identifizierung von Risikofaktoren für relativ seltene Erkrankungen (Hosmer/Lemeshow 2000). In diesen Studien werden einem Fall jeweils bis zu m Kontrollen anhand verschiedener Variablen zugeordnet, so dass der Effekt dieser Variablen in Bezug auf den Vergleich von Fällen und Kontrollen eliminiert wird. Die Zuordnung von m > 1 Kontrollen kann hierbei effizient sein, wenn die Anzahl der Fälle relativ gering und die Rekrutierung von Kontrollpersonen relativ leicht ist.

Die logistische Regression ist eine Regression einer binären Zielgröße ***Y***. Die Wahrscheinlichkeit zu erkranken wird dabei abhängig von verschiedenen Einflussgrößen ***Xi***, den so genannten Risikofaktoren, modelliert. Die Parameterschätzung erfolgt mit der Maximum-Likelihood-Methode (ML).
Die Likelihoodfunktion kann folgendermaßen angegeben werden:

$$L(\beta) = \prod_{k=1}^{K} \frac{P(x_{1k} \mid y_{1k} = 1) \prod_{i=2}^{n_k} P(x_{ik} \mid y_{ik} = 0)}{\sum_{j=1}^{n_k} \left(P(x_{1_j k} \mid y_{1_j k} = 1) \prod_{i=2}^{n_k} P(x_{ijk} \mid y_{ijk} = 0) \right)}$$

Da es sich bei der gematchten Fall-Kontroll-Studie um eine geschichtete Zufallsstichprobe mit ***K*** gematchten Paaren (1 Fall und n_k-1 Kontrollen im k-ten Paar) handelt, sieht die Likelihood-Funktion wie angegeben aus. Diese Likelihoodfunktion kann nicht in eine Likelihood der üblichen unbedingten logistischen Regression umgeformt werden, so dass die Durchführung einer bedingten logistischen Regression zwingend notwendig ist.

Bis SAS 8 konnte man die bedingte logistische Regression für das 1:m - Matching nur in der Prozedur PHREG ausführen (Stokes 1995). Dabei musste man in dieser für die Proportional Hazard (Cox) Regression vorgehaltenen Prozedur einige wenige Tricks anwenden, um eine ML-Schätzung für die bedingte logistische Regression zu erhalten (Sander 2004). Eine spezielle Aufbereitung der Daten, die nicht unbedingt für den Anwender logisch nachzuvollziehen ist, muss dazu durchgeführt werden. Dann läßt sich die gewünschte Analyse allerdings problemlos durchführen. Seit SAS Version 9 gibt es die bedingte logistische Regression auch in der Prozedur PROC LOGISTIC.

2 Bedingte logistische Regression in PROC PHREG

Zuerst möchten wir kurz auf die bisherige Durchführung der bedingten logistischen Regression in PROC PHREG eingehen.

2.1 Idee, Durchführung

Die ML-Schätzung der bedingten logistischen Regression im Kontext der Proportional Hazard Regression erhält man durch die Nutzung einer stratifizierten Überlebenszeitanalyse. Dazu muss die Zensierungsinformation speziell so angegeben werden, dass die Kontrollen zensiert sind, also „länger leben“ als die Fälle. Technisch lässt sich dies in PROC PHREG realisieren durch die folgende Definition der notwendigen Variablen:

- **Identifikationsvariable** *gruppe* für jeden Fall und seine Kontrollen
- **Zensierungsvariable** *outcome*: für Fälle=1, für Kontrollen=0
- **Zielgröße** *fall_kontrolle*: Fall=1, Kontrolle=2

Durch die Angabe des Statements STRATA in PROC PHREG wird stratifizierte Analyse durchgeführt

2.2 SAS Syntax

Die zugehörige Syntax für den Aufruf in PROC PHREG sieht dann folgendermaßen aus:

```
PROC PHREG DATA=temp;
     STRATA gruppe;
     MODEL fall_kontrolle*outcome(0) = einflussgroessen / <options>;
```

so dass die Kontrollen entsprechend zensiert werden.

2.3 Probleme / Möglichkeiten

Mit diesem Vorgehen gibt es einige Probleme, die hier kurz angesprochen werden:

- Zuerst ist das Auffinden der für eine bedingte logistische Regression passenden SAS-Prozedur zu nennen. Das sich unter der für die Überlebenszeitanalyse vorgehaltenen Prozedur PROC PHREG die bedingte logistische Regression verbirgt, kann als ein typischer SAS-Trick (gewusst wo!) bezeichnet werden und kann viele Anwender verwirren.
- Außerdem ist die Ausgabe angepasst an die Begrifflichkeiten der Survival Analyse, nicht an die der logistischen Regression. Zum Beispiel wird im Output das Hazard ratio ausgegeben, was dem Odds ratio in der bedingten logistischen Regression entspricht.
- Ein weiterer wichtiger Nachteil ist das Fehlen einer automatischen Dummy-Codierung von kategoriellen Variablen in PROC PHREG durch ein CLASS-Statement. Dadurch lässt sich die Prüfung (statistischer Test) einer Dummy-codierten klassierten Variable nur über die Definition der Dummy-Variablen im Data-Step und die Nutzung des TEST-Statements durchführen. In der Konsequenz können so aber die automatischen Variablenselektionsmethoden (Forward, Backward, Stepwise) nicht eingesetzt werden, jeder Schritt muss „per Hand“ durchgeführt werden.

Hinweis für weitergenutzte, ältere Programmversionen / Makros: Falls die bedingte logistische Regression über PROC PHREG eingebunden ist in größere Programmmodule / Makros oder standardmäßig genutzt wird, ist allerdings kein Handlungsbedarf gegeben. Falls der Leistungsumfang der Auswertung mit PROC PHREG ausreicht, kann mit dem Programm weiter gearbeitet werden, da die bisherige Lösung über PROC PHREG ohne Fehler auch in SAS Version 9 weiterläuft. Ein Umstieg auf die neuen Möglichkeiten in PROC LOGISTIC ist somit nicht notwendig.

Hinweis zu PROC TPHREG: In der experimentellen Prozedur PROC TPHREG gibt es im Unterschied zu PROC PHREG u.a. das CLASS-Statement, so dass eine automatische Dummy-Codierung eingebunden ist. Allerdings haben Vergleichsberechnungen ergeben, dass eine andere Koeffizientenschätzung als in PROC LOGISTIC ausgegeben wird. Die Erklärung dafür konnte aus Zeitgründen noch nicht untersucht werden, wir vermuten eine andere Definition der Dummy-Codierung der kategoriellen Variable als Default gegenüber PROC LOGISTIC.

3 Bedingte logistische Regression in PROC LOGISTIC

In der Beschreibung der Änderungen zwischen Version 8 und 9 in SAS („What´s new in data Analysis“) findet man in der SAS Online Doc die folgende Passage:

“You can perform a conditional logistic regression on binary response data by specifying the STRATA statement. This enables you to perform matched-set and case-control analyses. The number of events and nonevents can vary across the strata. Many of the features available with the unconditional analysis are also available with a conditional analysis.” (SAS Online Doc 9.1.3)

Somit ist ab Version 9 eine bedingte logistische Regression mit der PROC LOGISTIC möglich. Durch die Angabe des neuen Statements STRATA in PROC LOGISTIC lässt sich nun die bedingte logistische Regression ausführen, wobei in diesem Strata-Statement die Variable angegeben werden muss, die die jeweils (m+1) zusammengehörigen Kontrollen und Fälle kennzeichnet. Die Definition einer künstlichen Zensierungsvariable und die Festlegung der Zielgröße auf die Codierung 1 bzw. 2, wie in PROC PHREG gefordert, ist nicht mehr notwendig. Hier kann, wie üblich, der Fall mit 1 und die Kontrolle mit 0 codiert sein.

3.1 SAS Syntax

Die folgende Syntax ermöglicht dann die Durchführung einer bedingten logistischen Regression:

```
PROC LOGISTIC DATA=temp;
     CLASS class;
     STRATA gruppe;
     MODEL fall_kontrolle(event='1') = class / <options>;
```

3.2 Möglichkeiten / Probleme

Durch die Einbindung in PROC LOGISTIC können wichtige Auswertungsroutinen genutzt werden, die in PROC PHREG nicht vorhanden sind. Die wichtigsten sind im Abschnitt 2.3 schon angesprochen worden:

- Der Auffindeort im SAS-System und die Ausgabe sind angepasst an die bedingte logistische Regression.
- Durch das in dieser Prozedur vorhandene CLASS-Statement lassen sich für kategorielle Variablen automatisch Dummy-Variablen generieren, die zusammengehörend betrachtet werden in der weiteren Auswertung. Dadurch werden hier automatisch Regressionskoeffizienten und entsprechende Auswertungen für jede einzelne Dummy-Variable bestimmt sowie in der Typ-3-Statistik ein zusammenfassender Test des Einflusses auf die Zielgröße angegeben.

- Dadurch lassen sich u.a. die automatischen Variablenselektionsmethoden (Forward, Stepwise, Backward) mit eingebundenen, dummy-codierten kategoriellen Variablen nutzen, die im Zusammenhang logistischer Regressionen bei Auswertungen im epidemiologischen Umfeld häufig genutzt werden.
- Ein wichtiges Prinzip bei den Variablenselektionen ist das Hierarchieprinzip, wenn Wechselwirkungen untersucht werden. Dies besagt, dass Wechselwirkungen niedrigerer Ordnung und die Haupteffekte im Modell verbleiben, wenn eine Wechselwirkung höherer Ordnung ins Modell aufgenommen wird. Diese wichtige Auswertungsvariante ist in PROC LOGISTIC für die unbedingte logistische Regression vorhanden und kann nun auch für die bedingte logistische Regression genutzt werden.
- Ein weiteres Feature der PROC LOGISTIC ist die exakte logistische Regression. Diese ist angezeigt, wenn wenige Beobachtungen in Bezug auf die Anzahl Variablen ins Modell aufgenommen werden sollen und wenn wegen einer, bei stetigen Einflussgrößen häufig vorkommenden „Datenseparation" der Schätzalgorithmus der Regressionskoeffizienten nicht konvergiert. Allerdings ist die Nutzung des EXACT-Statements nur bei kleinen Datensätzen sinnvoll, da die Auswertung sonst sehr lange läuft oder abbricht.

Weiterhin werden folgende Einschränkungen bei der Nutzung von PROC LOGISTIC für die bedingte logistische Regression von SAS in der SAS Online Doc angegeben:
„The SCORE" and „WEIGHT" statements are not available with a STRATA statement. The following MODEL options are also not supported: CLPARM=PL, CLODDS=PL, CTABLE, LACKFIT, LINK=, NOFIT, OUTMODEL=, OUTROC=, SCALE= (SAS Online DOC)

3.3 Beispiel

Am Beispiel einer 1:3 gematchten Fall-Kontroll-Studie zur Untersuchung des Risikos von Parametern der Mundhygiene und des Zahnstatus auf Herz-Kreislauferkrankungen (Spahr 2006) werden die wesentlichen Aspekte des Aufrufs und des Outputs aufgezeigt.

```
proc logistic data=temp;
   class  bmiklass smk alkohol;
   strata gruppe;
   model  fall_kontrolle(event='1')=aalog cpitn_mean
                                    age sexm bmiklass smk alkohol
                                    diab bluthoch cholest_unb
                                    cholest_ja schule phyact statine
                                    / noint rl alpha=0.05 ;
```

PROC LOGISTIC is modeling the probability that Fall_Kontrolle='1'.
Convergence criterion (GCONV=1E-8) satisfied.
There were 756 observations read from the data set WORK.TEMP.

Conditional Analysis

Model Information

Data Set	WORK.TEMP
Response Variable	Fall_Kontrolle
Fall_Kontrolle	
Number of Response Levels	2
Number of Strata	**263**
Number of Uninformative Strata	20
Frequency Uninformative	39
Model	binary logit
Optimization Technique	Newton-Raphson ridge
Number of Observations Used	756

Response Profile

Ordered Value	Fall_Kontrolle	Total Frequency
1	1	243
2	2	513

Probability modeled is Fall_Kontrolle='1'.

Conditional Analysis

Class Level Information

Class	Value	Design Variables		
bmiklass	0	1	0	0
	1	0	1	0
	2	0	0	1
	3	-1	-1	-1
smk	0	1	0	
	1	0	1	
	2	-1	-1	
ALKOHOL	1	1	0	
	2	0	1	
	3	-1	-1	

Strata Summary

Response Pattern	Fall_Kontrolle 1	Fall_Kontrolle 2	Number of Strata	Frequency
1	**0**	**1**	**1**	**1**
2	**0**	**2**	**19**	**38**
3	**1**	**1**	**12**	**24**
4	**1**	**2**	**231**	**693**

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Aalog	1	19.9828	<.0001
cpitn_mean	1	0.0085	0.9264
age	1	0.0036	0.9524
sexm	1	0.0236	0.8779
bmiklass	**3**	**8.9323**	**0.0302**
smk	**2**	**7.0826**	**0.0290**
...			

Analysis of Maximum Likelihood Estimates / Odds Ratio Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	p	Point Estimate	95% Wald Confid. Limits	
Aalog		1	0.9874	0.2209	19.9828	<.0001	2.684	1.741	4.138
cpitn_mean		1	0.0238	0.2574	0.0085	0.9264	1.024	0.618	1.696
age		1	0.1181	1.9787	0.0036	0.9524	1.125	0.023	54.393
sexm		1	-1.5130	9.8449	0.0236	0.8779	0.220	<0.001	>999.999
bmiklass	**0**	1	-0.2043	0.4022	0.2580	0.6115	0.368	0.112	1.212
bmiklass	**1**	1	0.1260	0.3432	0.1348	0.7136	0.512	0.182	1.438
bmiklass	**2**	1	-0.7168	0.3434	4.3567	0.0369	0.220	0.080	0.606
smk	**0**	1	0.0436	0.2913	0.0224	0.8811	2.164	0.677	6.913
smk	**1**	1	0.6848	0.2654	6.6574	0.0099	4.109	1.385	12.192

4 Zusammenfassung, Vorteile der Durchführung bedingter logistischer Regressionen in PROC LOGISTIC

Durch die Angabe des neuen Statements STRATA in PROC LOGISTIC lässt sich nun eine bedingte logistische Regression ausführen, wobei in diesem STRATA-Statement die Variable angegeben werden muss, die die jeweils (m+1) zusammengehörigen Kontrollen und Fälle kennzeichnet. Diese Formulierung ist wesentlich nachvollziehbarer als die künstliche Definition einer „Zeitvariable“ für den Aufruf in PROC PHREG. Außerdem kann man nun die aus der Analyse unbedingter logistischer Regressionen bekannten Möglichkeiten von PROC LOGISTIC (Class-Statement, Hierarchie-Prinzip, exakte logistische Regression) nutzen. Zusammenfassend kann man die Vorteile der Nutzung von PROC LOGISTIC für eine bedingte logistische Regression gegenüber der Nutzung von PROC PHREG folgendermaßen aufzählen:

- Die Beschreibung zur Umsetzung einer bedingten logistischen Regression in SAS ist nun **einfacher zu finden**, da PROC LOGISTIC die logische Prozedur für die Durchführung einer logistischen Regression ist.
- Die Nutzung und die Lesbarkeit ist nachvollziehbarer, da eine „natürliche“ **Syntax** gefordert und eine entsprechend **logischere Output-Beschriftung** erzeugt wird.
- Die **Dummy-Codierung** kategorieller Variablen wird automatisiert erzeugt.

- Dadurch ist eine automatisierte **Variablenselektion** mit kategorialen Variablen und Dummy-Codierung möglich, sogar nach **Hierarchieprinzip**.
- Außerdem ist eine **exakte bedingte logistische Regression** bei kleinen Datensätzen möglich.

Literatur

[1] D.W. Hosmer, S. Lemeshow: Applied logistic regression (2nd Ed.). John Wiley & Sons Inc., New York, 2000

[2] SAS OnlineDoc (V. 9.1): SAS/STAT User´s Guide: The LOGISTIC Procedure. http://support.sas.com/91doc/docMainpage.jsp (Aufruf am 31.1.2007)

[3] S. Sander, B. Einsiedler, P. Kern, M. Kron: Auswertung von gematchten Fall-Kontroll-Studien mit der Software SAS. In: Schweizer, Großmann, Meule, Gaus (Hrsg.): Dokumentation-der Schritt ins 3. Jahrtausend, Proceedings der 8. DVMD-Tagung 2004 in Ulm. UniversitätsverlagUlm, 2004, 336-340

[4] A. Spahr, E. Klein, N. Khuseyinova, C. Boeckh, R. Muche, M. Kunze, D. Rothenbacher, G. Pezeshki, A. Hoffmeister, W. Koenig: Role of periodontal bacteria and importance of total pathogen burden – The coronary event and periodontal disease (CORODONT) study. Arch. Intern. Med. 166, 2006, 554-559

[5] M.E. Stokes, C.S. Davis, G.G. Koch: Categorical data Analysis Using the SAS System. SAS Institute Inc. Cary, NC, 1995