Logistische Regression in SAS®

Oliver Kuß

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

Kurzfassung

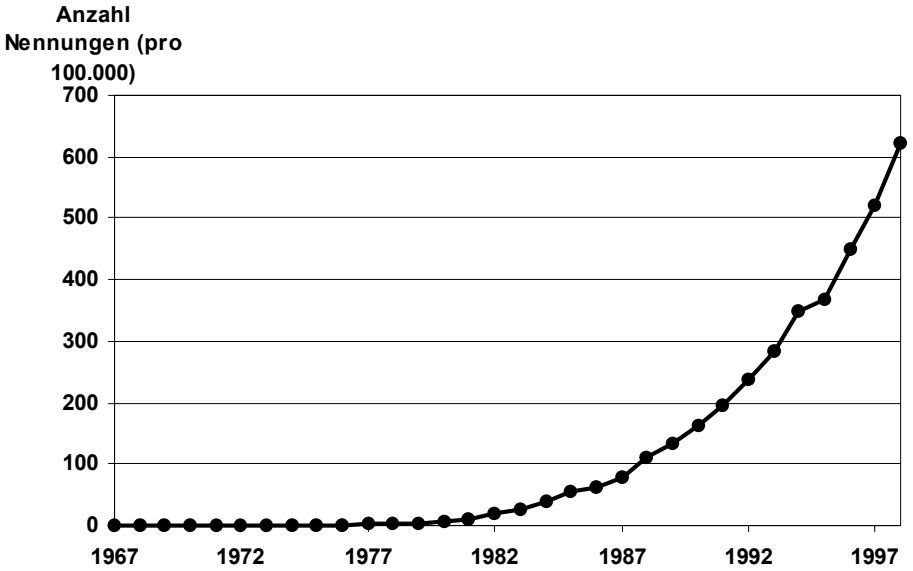
Das logistische Regressionsmodell hat sich seit seiner Einführung in den siebziger Jahren zu einer Standardmethode nicht nur, aber vor allem, in der Biometrie und Epidemiologie entwickelt, wenn es um die Auswertung von kategoriellen Zielgrößen geht. Die Gründe dafür sind vielfältig, exemplarisch seien genannt die leichte Interpretierbarkeit der geschätzten Parameter als Odds-Ratios, die Möglichkeit der Anwendung sowohl in prospektiven als auch in retrospektiven Designs und nicht zuletzt, die Verfügbarkeit von geeigneter Software. SAS stellt vier Prozeduren (CATMOD, GENMOD, LOGISTIC und PROBIT) zur logistischen Regressionsanalyse zur Verfügung, ferner sind Lösungen in PROC IML und PROC NLIN möglich. Dieser Beitrag stellt anhand eines Beispieldatensatzes aus der Medizin die einzelnen Prozeduren kurz vor und zeigt einige ihrer Stärken und Schwächen auf.

Einleitung

„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.“ Mit einem Standardsatz wie diesem beginnt eine Reihe von Artikeln, die sich methodisch mit der logistischen Regression befassen. Daß diese Behauptung auch zu belegen ist, zeigt das Ergebnis einer Ad-hoc-Literaturrecherche, die vom Autor in der Vorbereitung dieses Beitrags durchgeführt wurde. Dabei wurde in MEDLINE(Ovid) nach Veröffentlichungen gesucht, die das Stichwort „Logistic Regression“ im Abstract oder als Keyword enthalten und die gefundene Anzahl nach Jahren getrennt aufgezeichnet. Da die absolute Anzahl der medizinischen Veröffentlichungen über die Jahre ansteigt, wurde danach adjustiert. Es zeigt sich tatsächlich seit Anfang der siebziger Jahre ein nahezu exponentielles Anwachsen der Anzahl der Veröffentlichungen, die sich mit logistischer Regression beschäftigen oder in denen Daten mit Hilfe dieser Methode ausgewertet werden.

Allerdings wird die logistische Regression nicht nur in der Biometrie und der Epidemiologie verwendet. Eine weitere Ad-hoc-Recherche in Literaturdatenbanken anderer Disziplinen zeigt, daß sich dieses Verfahren auch in Gebieten wie der Ökonomie, der Informationstechnik, der Biologie, der Linguistik, den Geowissenschaften und vielen anderen Bereichen wachsender Beliebtheit erfreut.

Die Gründe dafür sind vielfältig. Exemplarisch seien genannt die leichte Interpretierbarkeit der geschätzten Parameter als Odds-Ratios, die Möglichkeit zur Schätzung von Wahrscheinlichkeiten für das Eintreten des Zielereignisses und nicht zuletzt, die Verfügbarkeit von geeigneter Software.



Das logistische Regressionsmodell beschreibt, ganz allgemein, den Zusammenhang zwischen einer kategoriellen Zielgröße und einer Menge von erklärenden Variablen. In seiner einfachsten Form, mit einer binären Zielgröße Y und einer Menge von Kovariablen X, hat das Modell die Form

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

mit

p_i Wahrscheinlichkeit für das Eintreten des Zielereignisses, gegeben die Werte der Kovariablen ($p_i = p(\text{Y} = 1 \mid \mathbf{x_i})$),

α: konstanter Term (Intercept),

$\boldsymbol{\beta}$: Vektor von Steigungsparametern und

$\mathbf{x_i}$: Vektor von Kovariablen.

Mögliche Erweiterungen des Modells sind andere Linkfunktionen (neben dem Logit-Link), Zielgrößen mit mehr als zwei Ausprägungen (nominal oder ordinal) oder auch korrelierte Beobachtungen.

Der Beispieldatensatz

Zur Veranschaulichung der verschiedenen Möglichkeiten, die SAS® bietet, logistische Regressionsmodelle zu schätzen, wird ein medizinischer Datensatz herangezogen. Dabei handelt es sich um ein Kollektiv von 162 Frauen mit unerfülltem Kinderwunsch, die in einer internationalen Multicenter-Studie unter der Federführung von Herrn Dr. Rimbach von der Heidelberger Frauenklinik erhoben worden waren. Motiv für die Durchführung dieser Studie war die Überprüfung einer neuen Methode zur Diagnose von empfängnisverhindernden Eileiterdefekten. Die erhobene Zielgröße war dabei das Eintreten einer Schwangerschaft im Verlauf der Beobachtungszeit von drei Jahren, als wichtige Kovariablen ins Modell aufgenommen wurden das Alter der Frau (in Jahren, AGE), die Dauer der Infertilität (in Jahren, INFER) und das Vorliegen eines Eileiterdefektes (ja/nein, TUBPHYSD). Eine klinische Wertung der gewonnenen Erkenntnisse kann in diesem Rahmen natürlich nicht stattfinden und wird späteren Veröffentlichungen vorbehalten sein, der Datensatz dient, wie gesagt, nur zur Veranschaulichung der eingesetzten SAS®-Prozeduren.
Der unten stehende SAS®-Output zeigt die geschätzten Parameter des Modells mitsamt ihren Standard-Fehlern, den Ergebnissen eines Wald-Signifikanztests auf Null-Einfluß der

jeweiligen Variable und das durch Exponierung aus dem Parameterschätzer hervorgegangene Odds-Ratio. Die negativen Vorzeichen der zu den Kovariablen gehörenden Parameter-Schätzer weisen darauf hin, daß mit steigendem Lebensalter, steigender Andauer der Infertilität und beim Vorliegen eines Eileiterdefektes (wenn auch nicht in allen Fällen statistisch signifikant) die Chance sinkt, noch schwanger zu werden. Eine Präzisierung dieses Sachverhaltes erlauben die geschätzten Odds-Ratios: Zum Beispiel sinkt mit jedem zusätzlichen Lebensjahr die Chance, noch schwanger zu werden, auf 95% im Vergleich zum Vorjahr.

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

Die einzelnen Prozeduren

Im folgenden sollen nun die diversen Möglichkeiten, die SAS® bietet, den beschriebenen Datensatz im Kontext der logistischen Regression zu analysieren, vorgestellt werden. Dazu wird zuerst der benötigte SAS®-Code abgebildet, der den oben stehenden SAS®-Output reproduziert. Im Anschluß daran steht eine, naturgemäß subjektive, Auswahl der Eigenschaften der jeweiligen Prozedur.

PROC LOGISTIC

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

PROC LOGISTIC ist die Standard-Prozedur zur Schätzung logistischer Regressionsmodelle in SAS®. Umso schwerer wiegen zwei Nachteile dieser Prozedur im Vergleich zu ihren „Konkurrenz"-Prozeduren: Das fehlende CLASS-Statement und die fehlende Möglichkeit, Interaktionsterme im MODEL-Statement direkt anzugeben.
Mittels des CLASS-Statements wird einer Prozedur das Skalenniveau der Kovariablen mitgeteilt. Dies ist unerläßlich für die korrekte Definition der Designmatrix und damit der zu schätzenden Parameter. Während es nun alle anderen Prozeduren erlauben, kategorielle Kovariablen als solche im CLASS-Statement anzugeben, wird man von PROC LOGISTIC dazu gezwungen, diese Kovariablen in einem vorgeschalteten DATA-Step zu Dummy-Variablen umzucodieren.
Ähnlich umständlich müssen zur Modellierung von Interaktionstermen von Variablen auch diese vorher außerhalb der Prozedur definiert werden.
Entschädigt für die beschriebenen Schwächen wird man von PROC LOGISTIC durch die umfangreichen Möglichkeiten zur Residuenanalyse, die Implementation des z.Zt. geläufigsten und in vielen Situationen zuverlässigsten Goodness-of-Fit-Tests, des Hosmer-Lemeshow-Tests und durch eine reiche Auswahl von Variablenselektionsmethoden, auch wenn vor deren unbedachter und unkritischer Anwendung gewarnt werden muß.

Desweiteren erlaubt PROC LOGISTIC eine Adjustierung der geschätzten Standardfehler der Parameterschätzer nach Overdispersion. Von Overdispersion spricht man, wenn, etwas lax gesprochen, die beobachtete Varianz der Zielgröße größer ist als die, die man ihr als binomial-verteilte Zufallsvariable eigentlich zugesteht. In einer Interpretation des logistischen Modells, wie sie in Ökonometrie gebräuchlich ist, bedeutet Overdispersion Heterogenität der wahren Modellparameter über die Beobachtungen.
Die geschätzten Wahrscheinlichkeiten für das Eintreten der Zielgröße für jede Beobachtung, die standardmäßig von PROC LOGISTIC mit dem OUTPUT-Statement in Datensätze herausgeschrieben werden können, können mithilfe der CTABLE-Option noch bias-adjustiert werden. Dieser Bias entsteht, weil zur Schätzung der Wahrscheinlichkeit für eine bestimmte Beobachtung das Modell verwendet wird, dessen Parameter mithilfe dieser Beobachtung geschätzt wurden. Daraus resultiert eine Verzerrung der geschätzten Wahrscheinlichkeit für diese Beobachtung in Richtung zum tatsächlich beobachteten Ereignis.
Schließlich können mit PROC LOGISTIC noch eine Reihe von anderen logistischen Regressionsmodellen geschätzt werden, als Beispiele seien hier nur genannt das Proportional Odds Modell für ordinale Zielgrößen oder das bedingte logistische Regressionsmodell zur Analyse von 1:1-gematchten Fall-Kontroll-Studien.

PROC GENMOD

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

Bei der Analyse mit PROC GENMOD macht man es sich zunutze, daß das logistische Regressionsmodell neben der linearen Regression, der Poisson-Regression und einigen anderen Regressionsmodellen zur Familie der generalisierten linearen Modelle gehört [vgl. [15]). Um die Parameter des logistischen Modelles korrekt zu schätzen, gibt man lediglich die Verteilung der Zielgröße (DIST-Option im MODEL-Statement) und die Linkfunktion (LINK-Option im MODEL-Statement) an.
PROC GENMOD erlaubt mithilfe des MAKE-Statements (als Erweiterung des OUTPUT- und des OUTEST-Statements der anderen Prozeduren) einen komfortablen Zugriff auf nahezu alle Berechnungen, die die Prozedur durchführt und die Ablage dieser Berechnungen in SAS-Dateien.
Seit Version 6.12 ist in PROC GENMOD (über das REPEATED-Statement) die GEE-Methode (Generalized Estimating Equations) zur Schätzung von logistischen Regressionsmodellen mit korrelierten Beobachtungen implementiert.
Es scheint, als sei PROC GENMOD diejenige Prozedur, die von SAS® Institute dazu auserkoren ist, *die* Prozedur der Zukunft für die logistische Regression zu sein, da laut SAS® - Homepage (http://www.sas.com/rnd/app/da/new/dastat.htm#GLM) PROC GENMOD die einzige Prozedur unter den hier vorgestellten ist, die mit Einführung von Version 7 umfangreiche Verbesserungen erfährt. In Aussicht gestellt werden unter anderem die Möglichkeiten, alternierende logistische Regressionmodelle zu fitten, ordinale Zielgrößen zu verarbeiten und Score-Tests für die Parameter-Schätzer zu berechnen.

PROC PROBIT

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

PROC PROBIT ist vor allem dazu vorgesehen, logistische Modelle im Rahmen von Toxizitäts-Studien zu berechnen. Daraus resultieren Sonderfunktionen wie die Schätzung eines medianen Wertes der Kovariablen (LD50), ab dem laut gefittetem Modell bei der Hälfte der Beobachtungen das Ereignis eintreten wird oder die Option, eine natürliche Response-Rate beim Wert 0 aller Kovariablen anzugeben, die in die Modell-Gleichung aufgenommen wird (C-Option im PROC-Statement). Die Link-Funktion, die standardmäßig in PROC PROBIT eingestellt ist, ist der Probit-Link, d.h. um ein logistisches Regressionsmodell mit Logit-Link zu schätzen, muß man diese Link-Funktion mit Hilfe der D-Option im MODEL-Statement explizit angeben. PROC PROBIT erlaubt ferner die Berechnung eines Proportional Odds Modells für ordinale Zielgrößen.

PROC CATMOD

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

PROC CATMOD ist die Standard-Prozedur in SAS® zur Schätzung von Modellen mit kategoriellen Variablen (**cat**egorical **mod**els). Deshalb müssen bei PROC CATMOD nicht die kategoriellen Kovariablen im CLASS-Statement angegeben werden, sondern die stetigen Kovariablen explizit im DIRECT-Statement. Bei der Interpretation der Parameter ist zu beachten, daß das zugrundeliegende Modell in PROC CATMOD im Vergleich zu den anderen Prozeduren anders parametrisiert wird, so daß nur die Parameterschätzer für die stetigen Kovariablen exakt reproduziert werden, die Parameterschätzer für kategorielle Kovariablen jedoch nur halb so groß erscheinen. PROC CATMOD erlaubt die Schätzung von logistischen Modellen mit mehrkategorieller Zielgröße, wobei nicht wie in PROC LOGISTIC oder PROC PROBIT nur ordinale Zielgrößen, sondern auch echt nominale Zielgrößen geschätzt werden können. Allerdings soll nicht verschwiegen werden, daß die am ehesten geeignete Prozedur zur Berechnung von multinominalen Modellen in SAS® PROC PHREG ist. PROC CATMOD erlaubt als einzige Prozedur die Schätzung der Parameter, zusätzlich zur Standardmethode mit Hilfe des Maximum Likelihood-Prinzips mit der Weighted Least Squares-Methode, wobei diese allerdings nur zu verläßlichen Ergebnissen führt, wenn es hinreichend viele Meßwiederholungen gibt, d.h. wenn für jede Wertemenge der Kovariablen genügend Beobachtungen vorliegen.

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 abseitig, vor allem angesichts des ausführlichen Programmcodes, erscheint die Schätzung des Modells mit PROC NLIN. Da aber das logistische Regressionsmodells auch ein nichtlineares Modell ist, ist dies durchaus möglich. Zur korrekten Berechnung der Parameterschätzer und ihrer Standardfehler führt PROC NLIN nur bei Einstellung der Optionen NOHALVE und SIGSQ=1 im PROC-Statement.
Einen Gewinn wird man von dieser Art der Modellschätzung wahrscheinlich nur haben, wenn man kompliziertere logistische Modelle (z.B. mit zusätzlichen Parametern in der Modellgleichung) an seine Daten anpassen will und den hier abgebildeten Code als Ausgangspunkt nimmt. Das hier abgebildete Programm ist eine abgespeckte Version des %GLIM-Makros ([2], S. 1168-1175), einem Vorläufer von PROC GENMOD.

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,;

                                        ⋮
```

Ähnlich umständlich wie die Benutzung von PROC NLIN erscheint die Verwendung von PROC IML, wobei zu bedenken ist, daß hier nur ein kleiner Teil des notwendigen Programmcodes abgebildet ist, der nötig wäre, um z.B. alle Berechnungen nachzuvollziehen, die PROC LOGISTIC liefert. Wenn jedoch neue statistische Verfahren programmiert werden sollen oder z.B. im Rahmen von Simulationsuntersuchungen der Zugriff auf alle Zwischenergebnisse ohne umständliches Hantieren mit OUTPUT-Statements machbar sein soll, empfiehlt sich die Umsetzung dieser Aufgaben mit PROC IML. Die vollständige Version des hier abgebildeten SAS/IML®-Programms findet sich in [1], S.135-138.

Fazit

SAS® bietet eine Vielzahl von Möglichkeiten, logistische Regressionsmodelle zu schätzen. Die Auswahl der richtigen Prozedur ist abhängig vom Zweck, d.h. vom konkreten Modell, das geschätzt werden soll. Für Standardanwendungen sind aber sicherlich PROC LOGISTIC und PROC GENMOD die Prozeduren der Wahl, alle anderen vorgestellten Prozeduren sind in andere Richtungen spezialisiert.

Literatur

Ausführliche Beschreibungen der hier nur kurz vorgestellten Prozeduren findet man in folgenden Handbüchern:

[1] SAS Institute Inc. (1989), SAS/IML® Software: Usage and Reference, Version 6, Fourth Edition, Cary, NC.

[2] SAS Institute Inc. (1990), SAS/STAT User´s Guide, Vol.1 &2, Version 6, Fourth Edition, Cary, NC. (PROC CATMOD, LOGISTIC, NLIN, PROBIT)

[3] SAS Institute Inc. (1992), SAS Technical Report P-229. SAS/STAT Software: Changes and Enhancements. Cary, NC. (PROC PHREG)

[4] SAS Institute Inc. (1993), SAS Technical Report P-243. SAS/STAT Software: The GENMOD Procedure. Cary, NC.

Folgende Bücher/Artikel sind für den Einstieg in die logistische Regression mit SAS® zu empfehlen:

[5] SAS Institute Inc. (1995), Logistic Regression Examples Using the SAS® System, Version 6, First Edition, Cary, NC: SAS Institute Inc.

[6] Stokes, Maura E., Davis, Charles S., and Koch, Gary G.(1995), Categorical Data Analysis Using the SAS System, Cary, NC: SAS Institute Inc., 1995.

[7] So, Y. (1993), A Tutorial on Logistic Regression, Proceedings of the Eighteenth Annual SAS Users Group International, 1290-1295.

Allgemeine Lehrbücher zur logistischen Regression und verwandten Modellen:

[8] Agresti, A. (1984), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

[9] Aldrich, J.H. and Nelson, F.D. (1984), *Linear Probability, Logit, and Probit Models*, 07-045, Thousand Oaks, CA: Sage Publications Inc.

[10] Collett, D. (1991), *Modeling Binary Data*, London: Chapman and Hall.

[11] Demaris, A. (1992), *Logit Modeling: Practical Applications*, 07-086, Thousand Oaks, CA: Sage Publications Inc.

[12] Hosmer Jr., D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons Inc.

[13] Liao, T.F. (1994), *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*, 07-101, Thousand Oaks, CA: Sage Publications Inc.

[14] Long, J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage.

[15] McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models, Second Edition*, London: Chapman & Hall.