Wir sind PROC LOGISTIC!

Georg Heinze
Med. Univ. Wien

- Mit Version 9.2 wurden in SAS/STAT zwei Methoden implementiert, die auf Forschungsarbeit unserer Arbeitsgruppe zurückgehen
- Wir konnten dadurch zur langen Liste der Publikationen beitragen, die in validierter weltweit verwendeter Software implementiert wurden

Beispiel: Frühgeburten

- Berger et al, J Perinat Med, 2003

Fruchtwasserkultur	**CLD-**	**CLD+**
Negativ	40	0
Ureaplasma urealyticum	17	4

Beispiel: Frühgeburten

- Schätzung der Odds Ratio (OR) mit PROC LOGISTIC:

Faktor	**OR (95% KI)**	**p-Wert**
FW-Kultur	>999 (0 - >999)	0.936

Maximum Likelihood

- *Modell*:

  $Y = f(X, \beta) + \varepsilon$

- *Parameter* β:

  log OR (Logistische Regression)

- *Likelihood*:

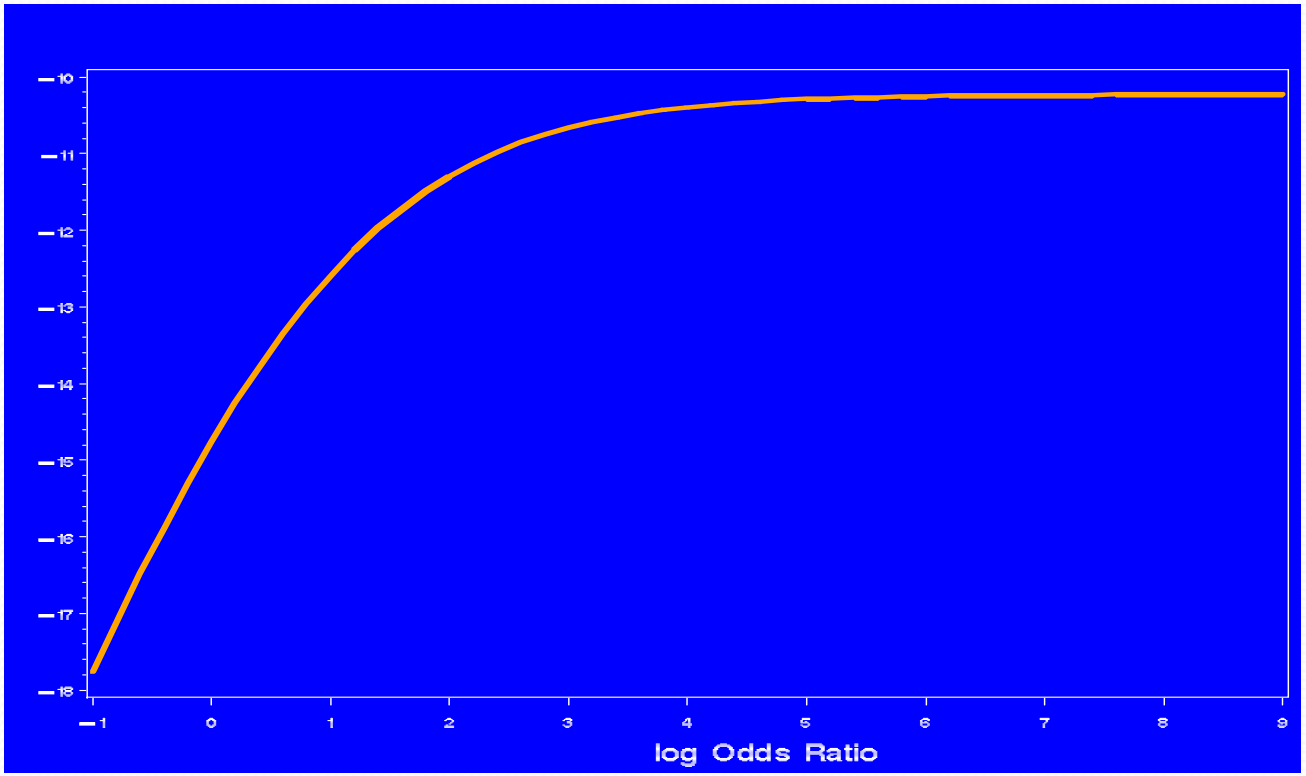
  Plausibilität der Daten unter Modellannahme β

- *Maximum-Likelihood-Prinzip:*
  - Schätzer ist jenes β, das Likelihood maximiert
  - Iteratives Schätzverfahren

Monotone Likelihood

- Im Beispiel:

  größte Likelihood wenn β unendlich ist



Monotone Likelihood

- Degenerierte Variation von *Y* in einer Subgruppe
  - kein CLD+ für negative FW-Kultur

- Schätzer für Modellparameter unendlich
- Standardfehler unendlich
- Konfidenzintervall unbestimmt
- Keine Signifikanz

Inzidenz

- Kleine Fallzahlen (hoher Zensierungsgrad), viele Faktoren
- Unbalanzierte Faktoren
- Hohe Effekte von Faktoren
- Hohe Korrelation von Faktoren

- Simulation (bootstrap) von Datensätzen

Der Weg...

- Schätzungen unverzerrt für große Fallzahl
- Verzerrungen bei kleiner Fallzahl:
  Schätzung für OR zu hoch

- Bisher:
  Korrektur *nach* Schätzung, setzt endliche Schätzer voraus
- Neu:
  Korrektur *während* der Schätzung (Firth, Biometrika 1993)

… zur Lösung

- Firths Idee von 1993 lange unbeachtet
- Anwendung auf Logistische Regression
  (von uns publiziert in *Statistics in Medicine* 2002)

- OR-Schätzer werden endlich
- Verzerrung wird eliminiert oder weitgehend reduziert
  (durch Simulation gezeigt)

Pönalisierte Schätzung

- Maximiere *pönalisierte* Likelihood (PML)
- Pönalisierung: wie *k* Pseudobeobachtungen
- Balance der Pseudobeobachtungen garantiert endliche Schätzungen
- Gewichtung der Pseudobeobachtungen erfolgt iterativ während der Schätzung

Konfidenzintervalle

- Üblich: symmetrisch

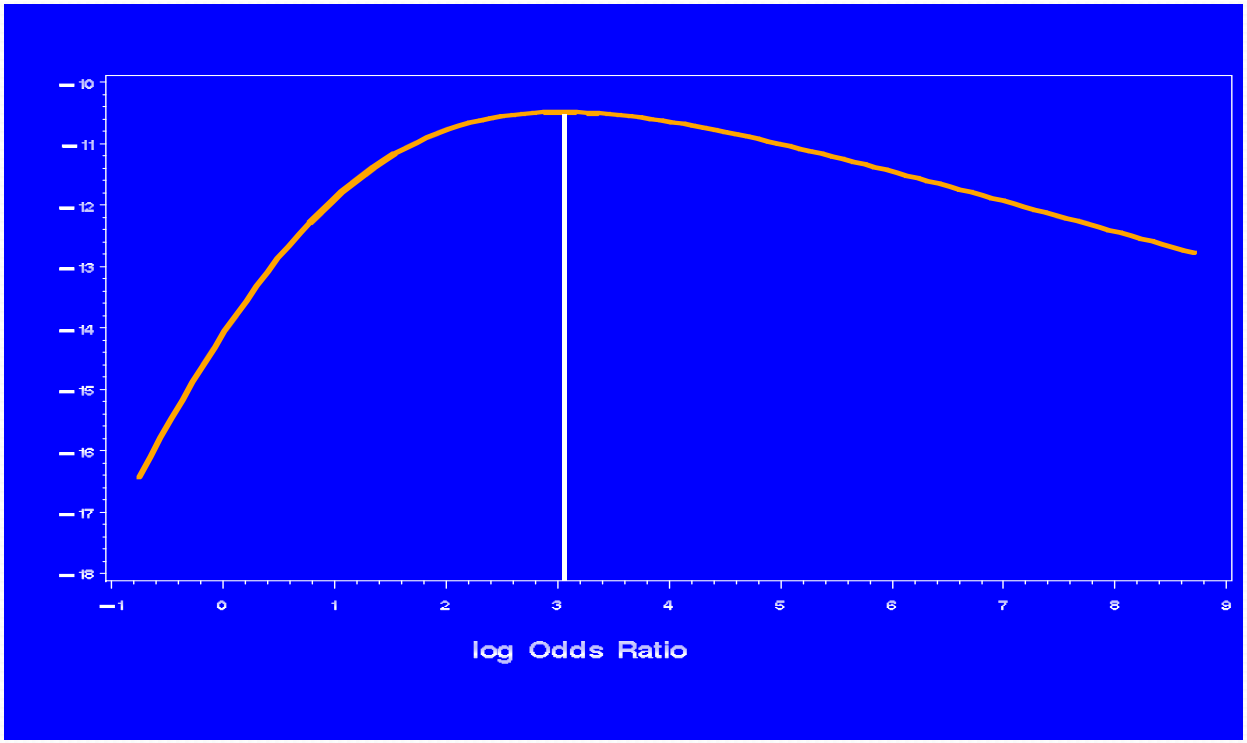
  log OR/log RR +/- 2 StdErr
- Bei monotoner Likelihood keine Symmetrie:

  95%-KI überdecken bis zu 100%

- KI auf Basis einer *pönalisierten Likelihood-Ratio* Statistik erlaubt Assymmetrie
- 95% weitgehend eingehalten (Simulation)

Beispiel: Frühgeburten

- OR (95% KI): 20.8 (2.1 – 2017), p=0.007
- Profil der pönalisierten Likelihood:



Praktische Relevanz

- Problem der monotonen Likelihood eliminiert
- Umstieg auf anderen Modelltyp nicht mehr notwendig
- Routineanwendung einfach
- Automatisierte Anwendung
  - Simulationsverfahren
  - Auswertung von hochdimensionalen Microarray-Daten
  - schrittweise Variablenselektion

Wir sind auch PROC PHREG!

- Gleiches Problem in Cox Regression für Lebensdaueranalysen
- Lösungsansatz funktioniert analog (implementiert in PROC PHREG)
- Von uns publiziert in *Biometrics* 2001

Software

- Wir haben diese neuen Methoden in SAS Makros implementiert, die seit einigen Jahren auf unserer Institutshomepage frei angeboten werden

- Nachdem unsere Methoden häufig verwendet und zitiert wurden, erfolgte die Implementierung in SAS 9.2:
- PROC LOGISTIC und PROC PHREG

Zur Geschichte der Implementation: 2002

Georg Heinze

From: **Georg Heinze <Georg.Heinze@akh-wien.ac.at>**
To: **gerhard.svolba@aut.sas.com**
Subject: **Suggestion**
Copies to: **michael.schemper@AKH-WIEN.AC.AT**
Send reply to: **Georg.Heinze@univie.ac.at**
Date sent: **Tue, 3 Sep 2002 09:02:24 +0200**

Dear Gerhard,

When applying logistic and Cox regressions to small samples occasionally a breakdown of standard maximum likelihood estimation is observed which results in infinite estimates of the parameters though the likelihood converges. This phenomenon has been termed 'separation' or 'monotone likelihood' and can now be totally avoided by means of a special correction originally suggested by David Firth (1993) to reduce the bias of small sample estimates.
In two recent papers (2001, 2002) we have further developed this correction and have demonstrated that the suggested method substantially improves the statistician's analysis options for small samples.

Corresponding programs have been developed and made available (see web resources below). We think that the options provided by our programs are of sufficient general relevance to applied statisticians to be included in a leading standard package like SAS. Therefore we wanted to draw your attention to these recent developments and are also willing to supply you with further information if needed.

References:

Firth D. (1993). Bias reduction of maximum likelihood estimates. Biometrika 80, pp. 27-38

Heinze and Schemper (2001). A solution to the problem of monotone likelihood in Cox regression. Biometrics 57, pp. 114-119.

Heinze and Schemper (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine 21, pp. 2409-2419.

Georg Heinze

From: **Gerhard Svolba <Gerhard.Svolba@aut.sas.com>**
To: **"'Georg.Heinze@univie.ac.at'" <Georg.Heinze@univie.ac.at>**
Subject: **AW: Suggestion**
Date sent: **Thu, 5 Sep 2002 13:19:16 +0200**

Hallo Georg!

Habe Dein Email direkt nach Cary weitergeleitet. Halte Dich auf dem laufenden sobald ich etwas weiss.

lg
Gerhard

-----Ursprüngliche Nachricht-----
Von: Georg Heinze [mailto:Georg.Heinze@akh-wien.ac.at]
Gesendet: Dienstag, 3. September 2002 09:02
An: Gerhard Svolba
Cc: michael.schemper@akh-wien.ac.at
Betreff: Suggestion

Dear Gerhard,

When applying logistic and Cox regressions to small samples occasionally a breakdown of standard maximum likelihood estimation is observed which results in infinite estimates of the parameters though the likelihood converges. This phenomenon has been termed 'separation' or 'monotone likelihood' and can now be totally avoided by means of a special correction originally suggested by David Firth (1993) to reduce the bias of small sample estimates.
In two recent papers (2001, 2002) we have further developed this correction and have demonstrated that the suggested method substantially improves the statistician's analysis options for small samples.

Corresponding programs have been developed and made available (see web resources below). We think that the options provided by our programs are of sufficient general relevance to applied statisticians to be included

Zur Geschichte der Implementation: 2006

- 16. Jan. 2006:
  Hello Dr. Heinze,

  I read your 2002 Stmed paper on separation and I'm very interested in pursuing the topic. I know that you created the algorithm in R, but you also mentioned that it is available in SAS. Is it possible to obtain it, as well as the 2004 report? My interest is solely for the purpose of research.

  Thank you very much.

  Leonardo Auslender, Member Appl. Staff
  SAS Institute. 908 470 0080 x 8217

Zur Geschichte der Implementation: 2006

- Dear Mr Auslender,

  you can find all relevant material on our website
  www.meduniwien.ac.at/msi/biometrie/programme/fl

  Years ago, I have suggested to SAS to implement the method we suggested in the SiM paper, however, with no success so far. As you probably know, the method has alread been implemented in LogXact 7. The example Cytel provides on the website uses confidence intervals based on normal approximation, but we prefer those based on profile penalized likelihood.

  Best regards,
  Georg

Zur Geschichte der Implementation: 2006

- Hello Dr. Heinze,

  Thank you very much for your e-mail. We are discussing in SAS all these issues, but I can't promise that we'll be implementing this or that tool.

  Leonardo.

Implementation in SAS 9.2

Eine (kleine) Simulationsstudie

Korrespondenz mit Usern

- Viele User unserer Makros haben mit mir Kontakt aufgenommen, um verschiedene Probleme zu diskutieren, oder auch um mich auf neue Ideen zu bringen

- Hello Georg Heinze,

  This is ..., a statistician working at Google.

  Could I have a soft copy of your paper "A solution to the problem of separation in logistic regression"? I think I might find some relevant information from your paper, but I haven't been able to find a copy.

  Thank you.

  ...

- Dear ...,

  attached please find the paper and a second upcoming Statistics in Medicine paper you might also be interested in.

  By the way, you're working at google? How come you could not find the paper? I typed in "a solution to the problem of separation georg heinze" and got the results as attached.

  Best regards,

  Georg

- Dear Georg,

  Thank you very much for the paper, and even more for the additional paper.
  You had a copy online, which google.com found. :)
  Long way to go! I just started at Google a month ago.
  Thanks again.

  Thanks,
  …

Wir sind noch nicht fertig!

- In intensiver Korrespondenz mit Usern werden wir immer wieder auf Situationen hingewiesen, in denen numerische Probleme auftreten
- Kürzlich haben wir den Ansatz erweitert auf eine Anwendung in der bedingten Logistischen Regression (zur Analyse von Fall-Kontroll Studien)
- Weitere Forschungsprojekte im laufen

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2409–2419 (DOI: 10.1002/sim.1047)

A solution to the problem of separation in logistic regression

Georg Heinze*,† and Michael Schemper

Department of Medical Computer Sciences, University of Vien
Vienna, Austria

BIOMETRICS 57, 114–119
March 2001

A Solution to

Section o

SUMMARY

the likeli
likelihood
predictiv
tion we
to redu
means
are ava
interva
manc
is fin

KEY
Pro

1. Introd

Statistici
icine or w
by simul
lihood, i
converg
mate d
param
monot
mum
For
the
all
alw
a l

ELSEVIER

Fixing

Departme

Re

Abstract

When analyzing clinical
estimates of a logistic mode
in a data set and known a
likelihood method cannot a
proposed by Heinze and Sch
maximum likelihood method,
an SPLUS library to make this
programs are also capable of
plotting the PPL function as wa
© 2002 Elsevier Science Ireland

Keywords: Monotone likelihood; No

1. Introduction

For analyzing clinical stud
comes, the logistic regression
used. The straightforward i
estimated parameters as log o
popularity in medical researc
of allowing models with more

* Corresponding author. Tel.: +43
1-40400-6687
E-mail address: georg.heinze@akh

0169-2607/02/$ - see front matter © 2
PII: S0169-2607(02)00088-3

STATISTICS IN MEDICINE
Statist. Med. 2008; 27:6455–6469
Published online 24 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3418

Avoiding infinite estimates of time-dependent effects
in small-sample survival studies

Heinze*,† and Daniela Dunkler

dical Statistics and Informatics, Medical University of Vienna,
nna A-1090, Austria

e-dependent effects.
ameter estimate
he inclusion
dictive

Research Article

Received 1 November 2008,
Accepted 16 October 2009
Published online in Wiley Interscience
(www.interscience.wiley.com) DOI: 10.1002/sim.3794

Statistics in Medicine

Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets

Georg Heinze*† and Rainer Puhr

Conditional logistic regression is used for the analysis of binary outcomes when subjects are stratified into several subsets, e.g. matched pairs or blocks. Log odds ratio estimates are usually found by maximizing the conditional likelihood. This approach eliminates all strata-specific parameters by conditioning on the number of events within each stratum. However, in the analyses of both an animal experiment and a lung cancer case–control study, conditional maximum likelihood (CML) resulted in infinite odds ratio estimates and monotone likelihood. Estimation can be improved by using Cytel Inc.'s well-known LogXact software, which provides a median unbiased estimate and exact or mid-*p* confidence intervals. Here, we suggest and outline point and interval estimation based on maximization of a penalized conditional likelihood in the spirit of Firth's (*Biometrika* 1993; 80:27–38) bias correction method (CFL). We present comparative analyses of both studies, demonstrating some advantages of CFL over competitors. We report on a small-sample simulation study where CFL log odds ratio estimates were almost unbiased, whereas LogXact estimates showed some bias and CML estimates exhibited serious bias. Confidence intervals and tests based on the penalized conditional likelihood had close-to-nominal coverage rates and yielded highest power among all methods compared, respectively. Therefore, we propose CFL as an attractive solution to the stratified analysis of binary data, irrespective of the occurrence of monotone likelihood. A SAS program implementing CFL is available at: http://www.muw.ac.at/msi/biometrie/programs. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: bias reduction; case-control studies; infinite estimates; modified score function; monotone likelihood; penalized likelihood

1. Introduction

Conditional logistic regression [1] is the standard tool for the analysis of matched case–control studies in which cases, i.e. patients, are each matched to one or several healthy controls. This type of analysis supplies parameter estimates which can be interpreted as log odds ratios. By means of multivariable modeling, odds ratio estimates can be adjusted for risk factors that were not accounted for in the matching procedure. Usually, parameter estimates are found by maximizing a conditional log likelihood, and inference is based on the estimated covariance matrix of the parameter estimates.

However, maximum likelihood analysis may lead to unreliable results in studies with a small number of cases or with several unbalanced risk factors. Possible shortcomings are a substantial small-sample bias [2, 3], an invalid inference [4] or a divergence of the iterative maximization caused by separation [5]. Exact conditional analysis [4, 6–8] addresses some of these problems. Exact point estimation and inference can be hampered by degenerate, nearly degenerate or highly discrete sufficient statistics [5, 9].

For unconditional logistic regression analysis, a bias-reduction method introduced by Firth [10] can be used to obtain unbiased estimates. In a former paper [5], we applied this method to the 18 female cases by age and developing lung cancer following radiation treatment for and particular interest lies in

el, which allows one become a standard for l's flexibility. Although

, Medical University of Vienna,

3-N13

Received 21 July 2008
Accepted 22 July 2008

likelihood;

Dank

- Michael Schemper
- Meinhard Ploner
- Daniela Dunkler
- Rainer Puhr