Relative Risiken aus einem log-Binomial-Modell unter Berücksichtigung der Randbedingungen durch die SAS-Prozedur NLP

Jürgen Wellmann
Institut für Epidemiologie und Sozialmedizin, Universität Münster
Domagkstr. 3
48149 Münster
wellmann@nwz.uni-muenster.de

Dirk Taeger
Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin (BGFA),
Institut der Ruhr-Universität Bochum
Bürkle-de-la-Camp-Platz
44789 Bochum
taeger@bgfa.de

Thomas Behrens
Bremer Institut für Präventionsforschung und Sozialmedizin (BIPS), Universität Bremen
Linzer Straße 10
28359 Bremen
behrens@bips.uni-bremen.de

Robert Kim
Helmut Maurer
Institut für Numerische und Angewandte Mathematik, Universität Münster
Einsteinstraße 62
48149 Münster
maurer@math.uni-muenster.de

Zusammenfassung

Bei der statistischen Analyse von epidemiologischen Daten werden häufig Regressionsmodelle eingesetzt, um dem Problem des „Confounding“ Rechnung zu tragen. Bei der Auswertung von Querschnittsstudien ist dabei ein Generalisiertes Lineares Modell (GLM) mit binomialverteilter Zielvariable und log-Link-Funktion besondern attraktiv, da es das von vielen Epidemiologen für diesen Studientyp bevorzugte relative Risiko als Ergebnis liefert. Das oben genannte GLM ist zum Beispiel in der SAS-Prozedur GENMOD implementiert. Bei der Berechnung der Maximum-Likelihood-Schätzer muss allerdings berücksichtigt werden, dass der Parameterraum des Modells durch lineare Nebenbedingungen beschränkt wird. PROC GENMOD und vergleichbare Prozeduren in anderen statistischen Programmpaketen berücksichtigen diese Nebenbedingungen nicht adäquat. Das führt gelegentlich zu Konvergenzproblemen. Als Reaktion darauf wurden in den letzten 10 Jahren in der epidemiologischen Literatur alternative Ansätze vorgeschlagen, die das eigentliche Problem aber nicht lösen, sondern nur umgehen
Wir wollen daher darauf aufmerksam machen, dass man die Nebenbedingungen bei diesem Modell mit bekannten numerischen Ansätzen durchaus berücksichtigen kann. In SAS finden sich entsprechende Algorithmen in der Prozedur NLP aus dem Bereich

Operations Research. Wir erläutern die Anwendung dieser Prozedur an einem Beispiel-Datensatz.

Schlüsselworte: Constrained Maximum Likelihood, Optimierung unter Nebenbedingungen, Generalisierte Lineare Modelle

1 Einleitung

1.1 Hintergrund

Die vorliegende Arbeit beschäftigt sich mit der Schätzung von relativen Risiken aus epidemiologischen Querschnittsstudien. Dieses Problem ist in den letzten Jahren in der Epidemiologie teilweise recht kontrovers diskutiert worden, ohne dass eine Lösung gefunden wurde, die in allen Belangen befriedigend wäre. Wir wollen nun einen Weg aufzeigen, wie zumindest die numerische Seite des Problems, die dabei eine große Rolle spielt, auch für den angewandt arbeitenden Epidemiologen befriedigend gelöst werden könnte.

Bei einer Querschnittsstudie wird zu einem bestimmten Zeitpunkt ein (repräsentativer) Querschnitt aus einer definierten Bevölkerung erhoben. Dieser Studientyp erlaubt daher Aussagen zur Häufigkeit (Prävalenz) von Erkrankungen in der Bevölkerung zu diesem Zeitpunkt. Genauso kann auch die Verteilung von Risikofaktoren untersucht werden. Hier geht es jetzt um den Vergleich der Prävalenzen einer Krankheit bei unterschiedlich exponierten Probanden. Aus diesem Vergleich kann man unter Umständen einen Rückschluss auf Ursache-Wirkungs-Beziehungen zwischen den Risikofaktoren und der betrachteten Krankheit ziehen.

Die Prävalenz entspricht einer Wahrscheinlichkeit, ebenso wie die kumulative Inzidenz, die in prospektiven Studien erhoben werden kann. Prospektive Studien beobachten Probanden über einen bestimmten Zeitraum. Die kumulative Inzidenz ist dann die Häufigkeit des Auftretens einer Krankheit in diesem Zeitraum. Alles Folgende gilt also auch für kumulative Inzidenzen aus prospektiven Studien. Allerdings wird in prospektiven Studien häufiger die Inzidenzrate anstelle der kumulativen Inzidenz berechnet, so dass sich das im Folgenden betrachtete Problem dann nicht stellt.

1.2 Risikomaße in Querschnittsstudien

Ein attraktives Risikomaß zum Vergleich zweier Prävalenzen ist das relative Risiko, das im Zusammenhang mit Querschnittsstudien auch als Prävalenz-Ratio (PR) bezeichnet wird. Wir unterscheiden zunächst der Einfachheit halber nur exponierte und nicht-exponierte Probanden mit den Krankheits-Prävalenzen p_1 und p_0. Dann ist das Prävalenz-Ratio PR = p_1/p_0.

Alternativ kann man das (Prävalenz) Odds Ratio OR = $[p_1/(1-p_1)]/[p_0/(1-p_0)]$ berechnen. Bei beiden Risikomaßen bedeutet ein Wert von eins, dass das Risiko der Exponierten gleich dem der Nicht-Exponierten ist. Das OR ist immer weiter von eins entfernt als PR, außer wenn OR = PR = 1. Die Unterschiede zwischen PR und OR sind gering, wenn die Prävalenzen klein sind, etwa kleiner als 10% [1]. In anderen Fällen sollten die beiden Risikomaße nicht verwechselt werden. Eine gewisse Verwechselungsgefahr ist dadurch gegeben, dass in Fall-Kontroll-Studien mit „density sampling“ das Odds Ratio sehr wohl ein unverzerrter Schätzer für das relative Risiko ist [2,3]. Bei Querschnittsstudien und häufig auftretenden Krankheiten gilt diese Aussage nicht mehr, und einige Autoren argwöhnen sogar, dass dann das OR ein statistischer Trick sei, mit dem ein dramatischeres Ergebnis suggeriert werden soll [4].

Obwohl ein Odds Ratio, richtig interpretiert, zu einer validen Aussage auch in Querschnittsstudien führt, bleibt anzuerkennen, dass das PR einfacher zu interpretieren und zu vermitteln ist. Trotzdem wird bei der Analyse von Querschnittsstudien häufig das Odds Ratio eingesetzt. Ein Vorteil dieses Risikomaßes ist seine Äquivarianz bezüglich der Codierung der Zielgröße. Das „OR für gesund“ ist gleich dem Kehrwert von „OR für krank“. Die Interpretation der Ergebnisse bleibt also gleich, egal ob man die Zielvariable krank/gesund mit 1/0 oder 0/1 codiert. Das PR besitzt diese wünschenswerte Eigenschaft nicht. Der in der Praxis entscheidende Vorteil des Odds Ratio liegt aber wohl darin begründet, dass es im Rahmen einer logistischen Regression berechnet werden kann.

1.3 Regressionsmodelle für Querschnittsstudien

In epidemiologischen Beobachtungsstudien, wie zum Beispiel Querschnittsstudien, muss man oft bezüglich potentieller Confounder „adjustieren“. Darunter verstehen Epidemiologen den Ansatz, die interessierende Expositions-Variable zusammen mit den potentiellen „Confoundern“ als Einflussvariablen in einem Regressionsmodell zu

berücksichtigen. Ein geeignetes Modell dafür ist das logistische Regressionsmodell, das aber eben als Ergebnis Odds Ratios liefert.

Andererseits hat bereits 1986 Wacholder [5] in einem Artikel, der vor allem an eine epidemiologische Leserschaft gerichtet war, dargelegt, wie man analog ein adjustiertes PR schätzen kann. Dazu wird ein Generalisiertes Lineares Modell verwendet, bei dem man für die Zielvariable eine Binomialverteilung spezifiziert und den Logarithmus als Link-Funktion angibt. Wacholder verweist dazu auf das Programm GLIM. In SAS berechnet man dieses log-Binomial-Modell mit der Prozedur GENMOD:

```
PROC GENMOD ...;
     MODEL y = x1 x2 ... / DIST=BIN LINK=LOG;
RUN;
```

Die Schätzer werden dabei nach dem Maximum-Likelihood-Prinzip berechnet, also als Lösung des Problems, die log-Likelihood-Funktion als Funktion der Parameter zu maximieren. Hierbei kann allerdings ein Problem auftreten, auf das Wacholder ebenfalls aufmerksam gemacht hat. Das Modell sieht für die Wahrscheinlichkeit p_i des i-ten Probanden vor: $p_i = \exp(x_i'\beta)$, wobei x_i ein Vektor von Einflussvariablen und β einen Parametervektor bezeichnet, $i = 1,..., n$. Da die Wahrscheinlichkeit p_i zwischen 0 und 1 liegen muss, sind also nur Parametervektoren β zulässig, bei denen für alle x_i gilt $x_i'\beta \leq 0$. Im Hinblick auf die log-Likelihood-Funktion muss man sogar $x_i'\beta < 0$ fordern. Damit liegt ein Maximierungsproblem mit linearen Nebenbedingungen vor (constrained maximum likelihood). Standard-Statistik-Software, auch die SAS-Prozedur GENMOD, ignoriert diese Nebenbedingungen. Das ist kein Problem, solange die Lösung „mitten im Parameterraum“ liegt. Andernfalls kann das aber dazu führen, dass der Schätzalgorithmus nicht konvergiert.

1.4 Beispiel

Dieses Problem wurde von Deddens et al. [6,7] an folgendem Datensatz dargestellt:

Tabelle 1: Hypothetischer Datensatz

y_i	0	0	0	0	1	0	1	1	1	1
x_i	1	2	3	4	5	6	7	8	9	10

Als Modell wählen sie: $p_i = \exp(\beta_0 + x_i\beta_1)$. Die tatsächlichen Maximum-Likelihood-Schätzer für die beiden Parameter sind –2,0936 und 0,2094. Dabei sind 2 der 10

möglichen Nebenbedingungen „aktiv“, die Lösung liegt also auf dem Rand des Parameterraums. Bei der Prozedur GENMOD treten bei diesem Beispiel Konvergenzprobleme auf.

2 Lösungsansätze

Obige Fragestellung ist in der epidemiologischen Literatur intensiv diskutiert worden [8,9] und wird hier kurz zusammengefasst.

2.1 Startwerte

Wenn die Lösung nahe am Rand liegt, helfen manchmal geeignete Startwerte. So wählt man etwa oft Modelle, bei denen eine Komponente von x_i für alle Probanden gleich 1 ist, der entsprechende Parameter wird als „Intercept“ bezeichnet. Skov et al. [10] schlagen dann vor, den Startwert für den „Intercept“ auf - 4 zu setzten. In GENMOD setzt man dazu in dem MODEL-Statement hinter einem Schrägstrich die Option „INTERCEPT = -4“. Wenn die Lösung genau auf dem Rand liegt, hilft dieses einfache Verfahren nicht unbedingt.

2.2 Poisson-Regression

Die Poisson-Regression gehört ebenfalls zu den Generalisierten Linearen Modellen. Die Zielvariable wird als Poisson-verteilt angenommen und als Link-Funktion wird der log-Link verwendet. Damit wird dann eigentlich eine „Erkrankungsrate“ geschätzt. Diese Raten sind praktisch oft identisch mit Prävalenzen, aber nicht immer. Sie können auch größer als eins werden. Dafür gibt es aber keine Beschränkung des Parameterraumes. Die Poisson-Regression ist äquivalent zu einem Vorschlag, bei dem das „proportional hazards model“ von Cox genutzt wird [11]. Eine neuere Arbeit ergänzt diesen Ansatz durch eine „robuste“ Varianzschätzung [12,13].

2.3 COPY-Methode von Deddens et al.

Deddens et al. [6,7] schlagen folgendes Vorgehen vor.

- Kopiere Originaldatensatz oft, etwa 999-mal.
- Füge eine weitere Kopie hinzu, bei der die Werte der Zielvariablen vertauscht sind.
- Analysiere zusammengesetzten Datensatz mit log-Binomial-Modell.
- Korrigiere Standardfehler

Die Beobachtungen mit den vertauschten Zielgrößen verlagern die Lösung des Maximierungsproblem weg vom Rand des Parameterraumes, die 999 Kopien der Original-Beobachtungen sorgen dafür, dass die Lösung approximativ korrekt ist. Das Verfahren kostet aber Rechenzeit und Speicherplatz und ist bei großen Datensätzen nicht anwendbar. Außerdem ist es problematisch, p-Werte und andere Statistiken entsprechend zu korrigieren.

Wir schlagen daher folgende Modifikation vor, bei der nur zwei Kopien des Original-Datensatzes gebraucht werden:

- Jede Beobachtung wird mit einem Gewicht w versehen, etwa $w = 0.999$.
- Eine Kopie jeder Beobachtung erhält das Gewicht 1-w, und die Werte der Zielvariablen werden vertauscht.
- Analysiere den zusammengesetzten Datensatzes mit gewichtetem log-Binomial Modell.

Dieses Verfahren ist auch bei größeren Datensätzen praktikabel. Die Standardfehler und andere Statistiken müssen nicht weiter korrigiert werden.

3 Lösung mit der SAS-Prozedur NLP

Alle bislang vorgestellten Ansätze umgehen das Problem der Optimierung unter Nebenbedingungen. Dabei gibt es dafür seit langem numerische Methoden, die hier eingesetzt werden können. Eine Reihe dieser Ansätze ist in der SAS-Prozedur NLP aus dem Bereich Operations Research implementiert. Wir benutzen hier ein Newton-Raphson-Verfahren. Der Programmcode sieht für den Datensatz von Deddens et al. folgendermaßen aus

```
PROC NLP TECHNIQUE=NEWRAP VARDEF=N COV=2 DATA=beispiel;
     PARMS beta0 = -4 -2, beta1=0.1;
     xbeta  = beta0 + x*beta1;
     prob   = EXP(xbeta);
     loglik = y*xbeta + (1-y)*LOG(1-prob);
     MAX loglik;
     LINCON beta0 +    beta1 < 0,
            beta0 + 10*beta1 < 0;
RUN;
```

Beim Aufruf der Prozedur wird das numerische Verfahren näher spezifiziert und der Eingabedatensatz beispiel übergeben. Die zweite Zeile listet die Parameter auf, die geschätzt werden sollen, und gibt Startwerte vor. In den nächsten drei Zeilen wird mit

Programmschritten, wie sie im DATA-Step üblich sind, die zu maximierende log-Likelihood beschrieben. Die Variable loglik enthält den Beitrag jeder Beobachtung zu log-Likelihood. Die Variable y, die dabei verwendet wird, sei die Zielvariable aus dem Eingabedatensatz beispiel. Die nächste Zeile gibt an, dass die so berechnete log-Likelihood maximiert werden soll. Das LINCON-Statement beschreibt die linearen Nebenbedingungen, die hier zum Tragen kommen.

Bei komplexeren Datensätzen ist es zu mühselig, alle Nebenbedingungen herzuleiten und aufzuschreiben. Diese Arbeit kann man vermeiden, wenn man die Option INEST von PROC NLP nutzt, mit der man die Nebenbedingungen in Form eines SAS-Datensatzes spezifizieren kann. Dazu kopiert man zunächst alle erklärenden Variablen des Analysedatensatzes in einen neuen Datensatz. Die Variablen müssen dann umbenannt werden, ihre Namen müssen den Namen entsprechen, die im PARMS-Statement aufgelistet werden. Darüber hinaus werden noch die Variablen _type_ und _rhs_ benötigt. _type_ ist eine Zeichenvariable und muss den Wert '<' erhalten, um die Art der Ungleichung anzuzeigen; _rhs_ spezifiziert die „right hand side“ der Ungleichungen und wird auf 0 gesetzt. Der Datensatz wird dann im Aufruf der Prozedur mit der Option INEST=<*datasetname*> übergeben. Damit werden genauso viele Nebenbedingungen spezifiziert wie Beobachtungen in dem Analysedatensatz sind, was aber scheinbar nicht weiter problematisch ist. Viele Nebenbedingungen sind redundant, einige sogar identisch. Letztere werden offenbar von SAS erkannt, denn es wird die Warnung ausgegeben: “A total of ... identical linear constraints are deleted”.

Wir haben dieses Verfahren in einer kleinen Simulation in Anlehnung an das Beispiel von Deddens et al. [6,7] ausprobiert mit folgenden Vorgaben

- $n = 100$ Beobachtungen pro simuliertem Datensatz.
- Eine erklärende Variable mit Werten $x = 1, 2, 3, ..., 10$, jeweils 10-mal.
- Modell $p = \exp(\beta_0 + x\beta_1)$, so dass $p = 1$ bei $x = 10$ und $p < 1$ bei allen anderen Ausprägungen von x.
- 1000 Simulationsläufe.

Tabelle 2: Ergebnisse der Simulation

Para-meter	wahrer Wert	Parameterschätzer		STDERR
		Mittelwert	STD	Mittelwert
β_0	-2,0	-2,0043	0,3081	0,3027
β_1	0,2	0,2000	0,0312	0,0310

STD = Standardabweichung der Schätzwerte
STDERR = Schätzwerte für die Standardfehler der Schätzer, wie von der Prozedur NLP berechnet
In 916 von 1000 Simulationsläufen wurde eine Nebenbedingung aktiv

Die Simulation macht deutlich, dass die Parameter β_0 und β_1 offenbar erwartungstreu geschätzt werden. Überdies ist die Standardabweichung dieser Schätzwerte nahe bei dem Mittelwert der Standardfehler, die von der Prozedur NLP ausgegeben werden. Das lässt vermuten, dass auch die Standardfehler erwartungstreu geschätzt werden.

4 Diskussion

Wir möchten auf dem Hintergrund der angesprochenen Diskussion in der Epidemiologie noch einmal betonen, dass die logistische Regression ein statistisch korrektes Verfahren zur Auswertung von Querschnittsstudien darstellt. Sie ist mit Standard-Software relativ einfach durchzuführen und wegen der Äquivarianzeigenschaft des OR mathematisch sogar eleganter als ein log-Binomial-Modell.
Dagegen hat das log-Binomial-Modell den Vorteil, als Ergebnisse die leichter zu interpretierenden relativen Risiken zu liefern. Die Anwendung dieses Modells in der Epidemiologie wird aber immer noch durch einen Mangel an geeigneter Software erschwert. Wir haben hier einen Ausweg aus diesem Dilemma gezeigt, der zumindest für SAS-Anwender mit relativ geringem Aufwand gangbar ist. Unser Ansatz hat den Vorteil, das eigentliche Problem, nämlich eine Maximierung der log-Likelihood unter Nebenbedingungen, direkt anzugehen, während andere Vorschläge darauf beruhen, dieses Problem zu umgehen.

Eine Frage, die unsere Arbeit noch nicht befriedigend beantwortet, ist die nach den asymptotischen Eigenschaften der so gewonnenen Schätzer. Schließlich gilt die übliche asymptotische Theorie der Maximum-Likelihood-Schätzer nur, wenn die Lösung des Maximierungsproblems nicht auf dem Rand des Parameterraumes liegt [14]. Immerhin deutet aber unsere Simulation an, dass es zumindest in dem dort untersuchten Spezialfall keine Verzerrung in den Schätzern der Parameter oder der zugehörigen

Standardfehler gibt. Abschließend bleibt noch zu vermerken, dass unserer Erfahrung nach die beschriebenen numerischen Probleme eher die Ausnahme als die Regel sind, da die Lösung des Maximierungsproblems meist deutlich innerhalb des Parameterraumes liegt und dann auch PROC GENMOD ohne Probleme konvergiert.

Literatur

[1] Zocchetti C, Consonni D, Bertazzi PA. Relationship between Prevalence Rate Ratios and Odds Ratios in Cross-Sectional Studies. Int J Epidemiol 1997;26(1):220–223.

[2] Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol 1982;116(3):547–553.

[3] Miettinen O. Estimability and estimation in case-referent studies. Am J Epidemiol 1976;103(2):226–235.

[4] Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med 1998;55:272–277.

[5] Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol 1986;123(1):174–184.

[6] Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge. In: Proceedings of the Twenty-Eighth Annual SAS® Users Group International Conference. Cary, NC, USA: SAS Institute, Inc; 2003. (Paper 270-28).

[7] Deddens JA, Petersen MR. RE: Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol 2004;159(2):213–214.

[8] McNutt LA, Wu C, Xue X, Hafner JP. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol 2003;157(10):940–943.

[9] Behrens T, Taeger D, Wellmann J, Keil U. Different Methods to Calculate Effect Estimates in Cross-sectional Studies. A Comparison between Prevalence Odds Ratio and Prevalence Ratio. Methods of Information in Medicine 2004;43(5):505–509.

[10] Skov T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol 1998;27:91–95.

[11] Zocchetti C, Consonni D, Bertazzi PA. Estimation of Prevalence Rate Ratios from Cross-Sectional Data. Int J Epidemiol 1995;24(5):1064–1065.

[12] Barros A, Hirakata V. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol 2003;3(21).

[13] Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. Am J Epidemiol 2004;159(7):702–706.

[14] Lehmann EL, Casella G. Theory of point estimation. 2nd ed. Springer Texts in Statistics. New York, Berlin, Heidelberg: Springer-Verlag; 1998.