Modellierung rechts-zensierter Zähldaten ein SAS-Makro unter Nutzung der Prozedur NLMIXED

Hannes-Friedrich Ulbrich
Bayer Pharma AG, GDDS
Research & Industrial Statistics
13342 Berlin
hannes-friedrich.ulbrich@bayer.com

Zusammenfassung

Zähldaten, d. h. nichtnegative ganzzahlige Werte, genügen wegen Überdispersion häufig nicht einer Poissonverteilung. Unter den Alternativen wird die negative Binomialverteilung häufig bevorzugt. Für Poisson- und Negativ-Binomialregression stellt SAS verschiedene Prozeduren zur Verfügung.
In speziellen Situationen hingegen sind einzelne Werte nicht exakt ermittelbar, sondern liegen nur zensiert vor. Das bedeutet z. B. für rechts-zensierte Daten, dass der wahre Wert größer ist als der beobachtete. Für diese Fälle fehlt in SAS bisher an entsprechenden Analysemöglichkeiten. Anhand eines Beispiels aus der Onkologie-Forschung wird der Bedarf dargestellt. Ein Makro auf Basis der Prozedur NLMIXED bietet eine Lösung für rechtszensierte Zähldaten.

Schlüsselwörter: zensierte Zähldaten, Negativ-Binomialmodell, Makro, PROC NLMIXED

1 Einleitung

Tumore können metastasieren, d. h. ein Tumor verursacht im selben Organismus weitere Tumore, genannt Metastasen. Neben der Behandlung des Primärtumors ist die Behandlung von Metastasen von großer Bedeutung. Die Entwicklung wirksamer Medikamente greift dabei auch auf solche Tiermodelle zurück, bei denen das Auftreten von Metastasen gut quantifizierbar ist.
Für ein Tiermodell, bei dem die Anzahl der Metastasen in der Lunge bestimmt wird, ist eine adäquate statistische Auswertestrategie zu entwickeln und programmtechnisch umzusetzen. Dabei ist davon auszugehen, dass sich die Metastasenanzahl in der Lunge nur bis zu einer bestimmten Anzahl exakt bestimmen lässt – bei größeren Anzahlen an Metastasen kommt es dazu, dass mehrere Metastasen zusammenwachsen und nicht mehr sicher einzeln gezählt werden können.
Mittels dieses Tiermodells soll untersucht werden, ob sich verschiedene Behandlungen in ihrer Wirkung auf die Metastasenanzahlen unterscheiden.
Als Beispiel diene ein Versuch mit 7 miteinander zu vergleichenden Therapieschemen[1]: Zwei Substanzen (A und B) in je 2 verschiedenen Dosierungen und ihre jeweiligen Ve-

[1] basierend auf einem realen Versuch, Datenpunkte und Gruppenzuordnungen leicht modifiziert.

hikel[2] stehen neben der Nichtbehandlung. Jede dieser Behandlungen wird zwischen 10 und 12 Tieren verabreicht. Vor Beginn des Versuches ist festgehalten worden, dass eine Metastasenanzahl bis 40 exakt zu bestimmen ist, größere Metastasenanzahlen werden mit 41 kodiert.
Mit dem Ziel, die wirksamen unter den Substanzdosen zu erkennen, interessiert den Onkologen dabei weder der Globaltest noch alle 21 möglichen paarweisen Gruppenvergleiche. Die wichtigen Gruppenvergleiche sind:

- die Nichtbehandlung gegen das Vehikel zur Substanz A,
- die Nichtbehandlung gegen das Vehikel zur Substanz B,
- das Vehikel zur Substanz A gegen die niedrige Dosis A,
- das Vehikel zur Substanz A gegen die höhere Dosis A,
- das Vehikel zur Substanz B gegen die niedrige Dosis B und
- das Vehikel zur Substanz B gegen die höhere Dosis B.

Für diese eingeschränkte Menge an Hypothesen ist ein möglichst trennscharfes Testen gewünscht.

2 Zähldatenmodelle und SAS

Zähldaten sind nichtnegative ganzzahlige Werte und als solche potenziell unbeschränkt. Rechtszensierung liegt vor, wenn für einen Wert nur bekannt ist, dass er größer als die beobachtete oder beobachtbare Anzahl ist. Für die Auswertung steht somit jeweils die Anzahl (z. B. ermittelter Metastasen) zur Verfügung, ergänzt um die zugehörige Angabe, ob diese Anzahl exakt erhoben (unzensiert) oder potenziell kleiner als der wahre Wert ist. Letzteres bedeutet, dass die Anzahl zensiert ist und nur eine untere Schranke des wahren Wertes angibt.
Die Anpassung einer Poissonverteilung an empirische Zähldaten erweist sich häufig als unbefriedigend. Die Poissonverteilung setzt mit $V(\mu) = \mu$ die Gleichheit von Erwartungswert μ und Varianz V voraus, Abweichungen hin zu höherer Variabilität werden als Überdispersion bezeichnet. Diese häufig bei Anwendung der Poissonverteilung zu beobachtende Überdispersion ist der Grund, nach tragfähigen Alternativen zu suchen.
Die negative Binomialverteilung stellt dabei die wohl beliebteste Modellierungsalternative dar [3], so beliebt, dass Warnungen vor allzu kritikloser Anwendung angezeigt sind [1].
Wird von negativer Binomialverteilung gesprochen, handelt es sich in der überwiegenden Anzahl von Fällen um die NB-2-Verteilung [3], in SAS-Dokumenten erscheint diese als 'negative binomial' (SAS/STAT 9.22 User's Guide). In den SAS-Prozeduren wird der Parameter als NEGBIN oder NB (GENMOD), NEGBINOMIAL, NEGBIN oder NB (GLIMMIX), NEGBIN oder NB (MCMC) bzw. NEGBIN (NLMIXED) bezeichnet.

[2] die reine Trägerlösung ohne Substanzbeigabe

Für die NB-2-Verteilung besteht – wie für die Poissonverteilung – ein funktionaler Zusammenhang zwischen Erwartungswert μ und Varianz V. Statt $V(\mu) = \mu$ wird jedoch von $V(\mu) = \mu + k\mu^2$ ausgegangen, d. h., die Varianz steigt mit dem Quadrat der Erwartung μ. k ist dabei ein zu schätzender Parameter (*ancillary* oder *heterogeneity parameter* [3]); dabei entspricht $k = 0$ einer Poissonverteilung und $k = 1$ einer geometrischen Verteilung [6, S. 2433].
Neben der NB-2-Verteilung werden weitere Verteilungen als negative Binomialverteilungen angesehen. Modifiziert man die Varianz zu $V(\mu) = \mu + k\mu^1 = \mu + k\mu$, so ergibt sich für die NB-1-Verteilung eine lineare Beziehung zwischen Erwartung und Varianz. Allgemeiner werden auch NB-p-Verteilungen mit einer Varianz von $V(\mu) = \mu + k\mu^p$ betrachtet, der Exponent p wird zum Parameter. NB-p-Verteilungen werden im Folgenden nicht weiter betrachtet.
Sowohl NB-1- als auch NB-2-Verteilung lassen sich im Rahmen verallgemeinerter linearer Modelle nutzen. Die Negativ-Binomialregression modelliert dabei die bedingte Erwartung für y_i als $E(y_i|\boldsymbol{x}_i) = \mu_i = e^{\eta_i} = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$ für die jeweils gegebene Werte-Konstellation der Kovariaten und Faktoren (in Form eines Vektors) $\boldsymbol{x}_i$. Dabei ist $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ der lineare Prädiktor. NB-1- und NB-2-Modell unterschieden sich bezüglich der Varianz.
Zu schätzen sind in beiden Fällen der Parametervektor $\boldsymbol{\beta}$ und der gemeinsame *heterogeneity parameter* k.
Bisher ist die Prozedur COUNTREG im Modul SAS/ETS die einzige Implementierung von NB-1- und NB-2-Modell nebeneinander (dort als NEGBIN1- bzw. NEGBIN2-Modell bezeichnet), alle anderen betreffenden Prozeduren (GENMOD, GLIMMIX, MCMC, NLMIXED) bieten als negative Binomialverteilung nur die NB-2-Verteilung an. Seit Version 9.22 stehen ggf. auch Erweiterungen für *zero inflated*-Situationen zur Verfügung. Die Modellierung von zensierten Zähldaten hingegen wird von SAS bisher nicht direkt unterstützt.[3]
Während bei Anwendung von PROC GENMOD oder PROC GLIMMIX auf die angebotenen Verteilungen zurückgegriffen werden muss, lässt sich mit der Prozedur NLMIXED eine beliebige Verteilung über die Spezifizierung ihrer Loglikelihood-Funktion anpassen.

3 Das Makro %rcnsNB

Das Makro %rcnsNB (rechtszensierte Negativbinomial-Modellierung) hat das Ziel, Zähldaten, von denen möglicherweise einige einer Rechtszensierung unterliegen, modellieren zu können. Es bedient sich dabei der in der Prozedur NLMIXED gegebenen Möglichkeit der Spezifizierung einer Loglikelihood-Funktion.

[3] In [2] wird eine SAS-basierte Nutzerlösung vorgestellt, das zugehörige Makro jedoch nicht offengelegt.

3.1 Parameter und Aufruf

Mit

```
%MACRO rcnsNB (dsn=%sysfunc(compbl(&syslast.)),
               prch=TECH=quanew,
               prci= ,
               outp=work._rcnsNB_,
               oout=ODS LISTING close%str(;)
                    ODS OUTPUT ParameterEstimates=work.tp_ParEst,
               cnbm=2,
               cnts= ,
               cnsd= ,
               prms= ,
               eta= ,
               chks=10,
               notes=N);
```

lässt sich das Makro aufrufen. Dabei sind die Parameter gemäß Tabelle 1 zu spezifizieren:

Tabelle 1: Parameter des Makroaufrufs

Parameter	Erläuterung
`dsn=`	Name des data sets mit Zähldaten, Kovariaten und Identifikatoren (ggf. mit data set options, default: &syslast.)
`prch=`	Zusatz-Optionen im Aufruf von proc NLMIXED (default: TECH=quanew)
`prci=`	Zusatz-Anweisungen innerhalb der Prozedur NLMIXED (mehrere sind mit maskierten Semikolons voneinander trennbar)
`outp=`	Name des data sets mit PROC NLMIXED Prädiktionen (default: work._rcnsNB_)
`oout=`	ODS Anweisungen für PROC NLMIXED (mehrere sind mit maskierten Semikolons voneinander trennbar, default sichert Parameterschätzungen in data set work.tp_ParEst)
`cnbm=`	Modell-Identifikator: 1 (NB-1-) oder 2 (NB-2-Verteilung)
`cnts=`	Variablenname der Zähldaten
`cnsd=`	Variablenname des Zensierungsindikators: 0 (unzensiert), 1 (zensiert)
`prms=`	Parameternamen und -startwerte für den linearen Prädiktor
`eta=`	linearer Prädiktor
`chks=`	(positiv-ganzzahlige) obere Schranke der Prädiktionen in `outp=` (Prädiktionen erfolgen für 0, 1, … &chks., default=10)
`notes=`	Note-Meldungen im Logfile: Y (mit) oder N (ohne)

`'eta=b0+b1*xx'` forderte hier z. B. ein Intercept mit dem Koeffizienten `b0` sowie für `b1` eine lineare Beziehung mit der Kovariaten `xx`. Dazu passend könnte `'prms=b0=0.3 b1=1.5'` die Startwertfestlegung sein.

Ein Aufruf mit `'notes=N'` führt dazu, dass während der Abarbeitung des Makros keine 'NOTE: '-Zeilen in die Logdatei eingehen, 'ERROR: '- und 'WARNING: '-Zeilen erscheinen jedoch genauso wie explizit programmierte Informationen.

3.2 Programmierstil

Mit dem Aufruf des Makros wird die Start- und unmittelbar vor Beendigung die End-Zeit in die Logdatei eingetragen.
Das Makro trennt klar zwischen Innen- und Außenwelt.
Das mit `dsn=` angeforderte data set wird innerhalb des Makros lediglich gelesen und steht auch nach Ende des Makroaufrufs unverändert zur Verfügung. Die Übergabe eines mittels data set options eingeschränkten data sets ist dabei möglich, makrointern wird mit einer temporären Kopie gearbeitet.
Makrointerne data sets unterliegen einer Namenskonvention. Das gemeinsame Präfix lautet 'work.rcnsNB_', alle data sets dieses Typs werden im Makro vor Beendigung desselben gelöscht. Ein versehentliches Nachnutzen von Dateien eines vorangehenden Makroaufrufs ist somit ausgeschlossen.
Durch die Deklaration aller makrointernen Makrovariablen als % Local wird eine Interaktion mit (zufällig) namensgleichen Makrovariablen außerhalb des Makros unterbunden.

3.3 Die Prozedur NLMIXED und die Loglikelihood-Funktion

In [4] ist die Loglikelihood-Funktion[4] der rechtszensierten negativen Binomialverteilung für NB-1- und NB-2-Verteilungen angegeben, die Unterschiede zwischen beiden ergeben sich mit dem Indikator-Parameter a.
Bei der Spezifikation des Loglikelihood-Anteils einer jeden Beobachtung erfolgt im Makro eine klare Fallunterscheidung zwischen NB-1- und NB-2-Verteilung. Für jede der beiden möglichen Belegungen des `'cnbm= '`-Parameters existieren separate Passagen im Makro. Damit können Summanden wie `'mean**0'` ($x^0 \equiv 1$, für $x \neq 0$) mit `mean` als Variable vermieden werden.
In beiden Fällen geht der vom Nutzer als Makro-Parameter `'eta= '` formulierte lineare Prädiktor η als Exponent der Exponentialfunktion in die Schätzung des bedingten Erwartungswertes $\mu = e^{\eta}$ ein.

3.4 Ergebnisausgaben

Über den PROC NLMIXED-eigenen Mechanismus der Prädiktionsgenerierung (PREDICT OUTP= ;) werden gemäß gewählten Makroparameterwerten für `'outp= '` und `'chks= '` Prädiktionen für die Anzahlen von 0, 1 bis &chks. bereitgestellt. Makrointern wird dabei für jede Anzahl ein data set generiert, die zum data set &outp. zusammengefasst werden.

[4] dabei ist $\Gamma((y_i + \mu_i^{1-a})/\tau)$ in Gleichung (2) durch $\Gamma(y_i + (\mu_i^{1-a}/\tau))$ zu ersetzen

Alle anderen Ergebnis- und Ausgabeinteressen sind über 'oout= ' aus ODS Anweisungen zusammenzustellen. Mehrere aufeinanderfolgende Anweisungen sind mit je einem maskierten Semikolon (z. B. mit %str(;)) voneinander zu trennen. In SAS V 9.22 lässt sich durch 'oout=ODS LISTING close%str(;) ODS OUTPUT ParameterEstimates=work.tp_ParEst' jegliche Ausgabe in das listing verhindern. Die Parameterschätzungen könnten nach Abschluss des Makroaufrufs z. B. durch eine PROC PRINT Anweisung sichtbar werden.

3.5 Nutzung ohne CLASS Anweisung

Wegen der Nutzung von PROC NLMIXED sind lineare Prädiktoren ohne CLASS Anweisung zu formulieren. Merkmale mit (mehr als 2) Faktorstufen müssen in Indikatorvariablen aufgeteilt werden. Diese können vorab dem in 'dsn= ' spezifizierten data set mitgegeben sein[5]. Alternativ steht die Möglichkeit der expliziten Formulierung in 'eta= ' zur Verfügung, wie z. B. für die Modellierung eines 4-stufigen Merkmals 'group' (bei 'group=4' als Referenzstufe oder -kategorie) in 'eta=b0+b11*(group=1)+b12*(group=2)+b13*(group=3)'.

4 Das Beispiel

Für den in der Einleitung beschriebenen Versuch sind die erhobenen Metastasenanzahlen in Tabelle 2 zusammengestellt.

Tabelle 2: Ermittelte Metastasenanzahlen

Gruppe	Behandlung	# exakt	exakte Zählwerte	# zensiert (>=41)
1	unbehandelt	**7**	5 5 6 13 23 31 33	**3**
2	Vehikel A	**2**	20 29	**8**
3	Vehikel B	**8**	15 20 21 26 29 32 34 36	**4**
4	A, niedrig dosiert	**11**	7 14 14 15 16 22 25 25 25 27 30	**.**
5	A, höher dosiert	**8**	1 10 12 13 17 18 24 34	**4**
6	B, niedrig dosiert	**10**	2 3 5 14 15 17 19 21 22 35	**2**
7	B, höher dosiert	**10**	4 6 8 9 12 17 19 22 27 36	**2**

Der Versuch ist leicht unbalanciert. Die Anteile an zensierten Werten liegen dabei zwischen 0% in Gruppe 4 und 80% in Gruppe 2. Für Gruppe 2 lässt sich somit die mediane Anzahl an Metastasen nicht ermitteln.

[5] sei es durch Nutzung eines Datenschritts oder Anwendung von PROC GLMMOD

Gesucht werden die erwarteten mittleren Metastasenanzahlen einer jeden Gruppe sowie deren Unterschiede für die Gruppenvergleiche gemäß den präspezifizierten Kontrasten sowohl unter der Annahme einer NB-2- als auch einer NB-1-Verteilung.

Tabelle 3: Schätzungen der mittleren Metastasenanzahlen

Gruppe	*NB-2*		*NB-1*	
	Schätzung	**Rang**	**Schätzung**	**Rang**
1	**29.3**	**4**	**26.0**	**4**
2	**95.4**	**7**	**62.3**	**7**
3	**40.0**	**6**	**40.1**	**6**
4	**20.0**	**1**	**24.5**	**3**
5	**30.7**	**5**	**27.2**	**5**
6	**21.7**	**2**	**21.1**	**1**
7	**22.4**	**3**	**22.7**	**2**

In den Tabellen 3 und 4 sind die wichtigsten Ergebnisse zusammengefasst.
Zwischen NB-2- und NB-1-Verteilung tauschen die Gruppen mit den niedrigsten Metastasenanzahlen (4, 6 und 7) ihre anhand der Schätzung mittlerer Metastasenanzahlen bestimmte Positionen, während sich für die Gruppen mit höheren Metastasenanzahlen die Ordnung zueinander nicht ändert.

Tabelle 4: Gruppenvergleiche

Gruppe	*NB-2*			*NB-1*		
	Schätzung	**t-Statistik**	**p-Wert***	**Schätzung**	**t-Statistik**	**p-Wert***
1~2	**-1.18**	**-2.42**	**0.0180**	**-0.87**	**-3.23**	**0.0018**
1~3	**-0.31**	**-0.91**	**0.3669**	**-0.43**	**-1.65**	**0.1026**
2~4	**1.56**	**3.26**	**0.0016**	**0.94**	**3.56**	**0.0006**
2~5	**1.13**	**2.36**	**0.0205**	**0.83**	**3.23**	**0.0018**
3~6	**0.61**	**1.90**	**0.0607**	**0.64**	**2.48**	**0.0151**
3~7	**0.58**	**1.81**	**0.0739**	**0.57**	**2.24**	**0.0278**

* p-Werte sind nicht adjustiert

Modellentsprechend beziehen sich die Werte der Kontraste auf die logarithmische Skala. Zugehörige t-Statistiken und p-Werte verstehen sich als jeweils zur Nullhypothese eines Kontrastes vom Wert 0.

Auffälligste Unterschiede in der Versuchsauswertung zwischen NB-2- und NB-1-Modellierung ergeben sich für die erwartete Anzahl an Metastasen in Gruppe 2. Der Unterschied zwischen 62 bzw. 95 zu erwartenden Metastasen mag zur Frage nach dem "besseren Modell" führen.
Legt man eines der von der Prozedur NLMIXED angebotenen Informationskriterien AIC, AICC oder BIC zugrunde, fällt die Wahl auf das NB-1-Modell.

5 Ausblick

Die mit dem Makro %rcnsNB zur Verfügung stehende Auswertestrategie lässt sich mit geringem Aufwand für linkszensierte Zähldaten modifizieren.
Das Makro %rcnsNB kann beim Autor angefordert werden. Die Weitergabe des Makros %rcnsNB erfolgt ohne jegliche Garantien seitens des Autors oder der Bayer Pharma AG.

Literatur

[1] Berk R, MacDonald JM. Overdispersion and Poisson Regression. Journal of Quantitative Criminology 24: 269-284, 2008.

[2] Chou N-T, Steenhard D. A Flexible Count Data Regression Model Using SAS PROC NLMIXED. SAS Global Forum 2009, paper 250.

[3] Hilbe JM. Negative Binomial Regression. Cambridge: Cambridge University Press, 2007

[4] Jung BC, Jhun M, Song SH. Testing for overdispersion in a censored Poisson regression model. Statistics 40: 533–543, 2006.

[5] SAS Institute. SAS/ETS 9.22 User's Guide. Cary, NC: SAS Institute Inc., 2010

[6] SAS Institute. SAS/STAT 9.22 User's Guide. Cary, NC: SAS Institute Inc., 2010