Auswertung von Zähldaten mit Nullenüberschuss in SAS - dargestellt am Beispiel eines Fütterungsversuches mit Milchkühen

Katrin Thamm
Institut für Agrar- und Ernährungswissenschaften, Martin-Luther-Universität Halle-Wittenberg
Karl-Freiherr-von-Fritsch-Str.4
06120 Halle
Katrin.Thamm@landw.uni-halle.de

Karen Höltl
Institut für Agrar- und Ernährungswissenschaften, Martin-Luther-Universität Halle-Wittenberg
Karl-Freiherr-von-Fritsch-Str.4
06120 Halle
Karen.Hoeltl@landw.uni-halle.de

Norbert Mielenz
Institut für Agrar- und Ernährungswissenschaften, Martin-Luther-Universität Halle-Wittenberg
Karl-Freiherr-von-Fritsch-Str.4
06120 Halle
Norbert.Mielenz@landw.uni-halle.de

Joachim Spilke
Institut für Agrar- und Ernährungswissenschaften, Martin-Luther-Universität Halle-Wittenberg
Karl-Freiherr-von-Fritsch-Str.4
06120 Halle
Joachim.Spilke@landw.uni-halle.de

Michael Bulang
Institut für Agrar- und Ernährungswissenschaften, Martin-Luther-Universität Halle-Wittenberg
06099 Halle
Michael.Bulang@landw.uni-halle.de

Zusammenfassung

Für die statistische Modellierung von Zählmerkmalen wird üblicherweise das Poisson-Regressionsmodell verwendet. Übersteigt die Anzahl der beobachteten Nullen für die Zähldaten die erwartete Anzahl, abgeleitet aus einer positiven Ursprungsverteilung wie beispielsweise der Poissonverteilung, so liegt ein sogenannter „Nullenüberschuss“ vor. Für die Behandlung dieses Problems sind kombinierte Modelle erforderlich, bei denen für das Ereignis Null und die positiven Ereignisse verschiedene Verteilungen zugrunde gelegt werden. Bekannt sind vor allem Zero-Inflated- und Hurdle-Modelle unter Verwendung von Poisson- und negativer Binomialverteilung. Zur Anpassung der genannten Modelle stehen in SAS beispielsweise die Prozeduren GENMOD, NLMIXED und mit Einschränkung auch die Prozedur GLIMMIX zur Verfügung. Möglichkeiten zur Anwendung dieser Prozeduren werden an dem Zählmerkmal Anzahl Fressplatzbesuche einer Kuh pro Stunde demonstriert. Als Einflussgrößen sind unter anderem die fixen Effekte der Faktoren Ration, Stunde und Laktationsnummer, der zufällige Effekt einer Kuh und als stetige Kovariable der Laktationstag zu berücksichtigen. Der Vergleich von beobachteten und

vorhergesagten Wahrscheinlichkeiten zeigte für das Hurdle-Modell mit negativer Binomialverteilung die beste Übereinstimmung.

Schlüsselwörter: Regressionsmodelle, Poissonverteilung, negative Binomialverteilung, Hurdle-Modelle, Zero-Inflated-Modelle

1 Einleitung

Bei dem vorliegenden Fütterungsversuch an Milchkühen mit hoher Leistung sollten klassische Leistungsmerkmale wie die Milchmenge und Futteraufnahme analysiert werden (vgl. [2]). Die Erfassung der Futteraufnahme erfolgte über automatische Fress-Wiege-Tröge. Hierbei wurde zusätzlich das Fressverhalten in Bezug zur Tageszeit ausgewertet, in dem man die Anzahl Besuche pro Stunde am Futterautomat als Merkmal analysierte. Für die Auswertung von Zählmerkmalen stehen u. a. die Poisson(P)- oder Negative Binomialverteilung (NB) zur Verfügung. Liegt jedoch die Anzahl der beobachten Nullen über der erwarteten Anzahl, abgeleitet aus diesen beiden Verteilungen, so spricht man von einem Nullenüberschuss. Für die Modellierung von Zählmerkmalen mit Nullenüberschuss können Zero-Inflated(ZI)- und Hurdle(H)-Modelle, basierend auf der Poisson- und der negativen Binomialverteilung, verwendet werden. Hurdle-Modelle (vgl. [6]) sind kombinierte Modelle, wobei die Ereignisse kleiner gleich einer Hürde sowie Ereignisse oberhalb einer Hürde unterschieden werden und deren Verteilungen miteinander kombiniert werden. Liegt ein Nullenüberschuss vor, so wird als Hürde der Wert Null gewählt. Zero-Inflated-Modelle (vgl. [4]) sind ebenfalls kombinierte Modelle, wobei Nullen nicht nur durch einen Auswahlprozess entstehen, sondern auch durch den Ursprungsprozess selbst. Ziel der Arbeit war es, ein geeignetes Auswertungsmodell für das Merkmal Anzahl Besuche pro Stunde am Futterautomat zu finden und anzupassen.

2 Material und Methoden

Das Merkmal Anzahl Besuche pro Stunde wurde innerhalb eines Fütterungsversuches mit Milchkühen der Rasse "Deutsche Holstein" in der Landesanstalt für Landwirtschaft, Forsten und Gartenbau in Iden erfasst (vgl. [2]). Es wurden drei unterschiedliche Mischrationen Mais, Gras und Luzerne getestet. Die drei Behandlungen unterschieden sich dabei u. a. im Verhältnis der Grundfutterkomponenten Mais- bzw. Grassilage. Die Trockensubstanz der Mischration "Mais" bestand zu 41,5% aus Maissilage und 12% aus Grassilage, der Mischration "Gras" zu 18,4% aus Maissilage und 29,4% aus Grassilage und der Mischration "Luzerne" zu 29,2% aus Luzernesilage und 18,3% aus Grassilage. Die Anzahl Kühe je Behandlung umfasste 28 (Mais) bzw. jeweils 29 Tiere (Gras und Luzerne). Die Dauer des Fütterungsversuches betrug 169 Tage. Pro Tag und Tier wurden 24 Beobachtungen entsprechend der Stunden des Tages für das Untersuchungsmerkmal berücksichtigt.

Die beobachteten Frequenzen für das Untersuchungsmerkmal sind in Abbildung 1 dargestellt. Die Beobachtung Null, d. h. kein Besuch am Fütterungsautomat innerhalb einer Stunde, tritt mit 56,5% am häufigsten auf.

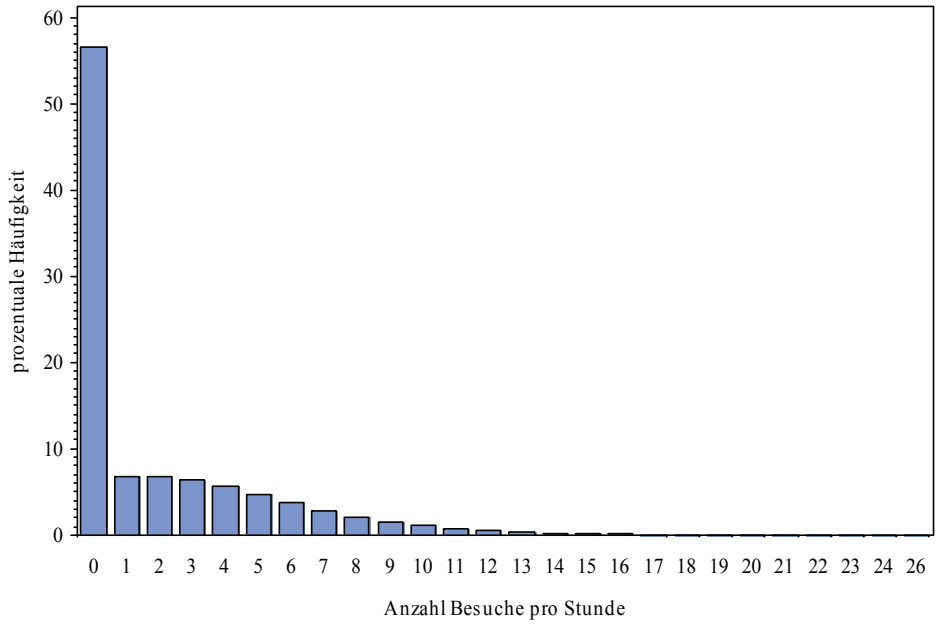

Abbildung 1: Empirische Verteilung des Merkmals Anzahl Besuche pro Stunde am Futterautomat

Das Fressverhalten der Milchkühe sollte im Verlauf der Tageszeit analysiert werden, da man einen bestimmten Tag-Nacht-Rhythmus vermutete. Es zeigte sich jedoch, dass durch den Einfluss der Melkzeiten (in den Stunden 4, 12, 20) Reinigung (in Stunde 7) und Sperrzeit der Tröge (in Stunde 23) kein Rhythmus zu beobachten war (siehe Abbildung 2). Die Einflussgröße Stunde konnte dadurch nur als qualitativer Faktor mit 24 Niveaustufen in das Auswertungsmodell einbezogen werden. Als weitere Einflussgrößen wurden die Ration und die Laktationsnummer als fixe Faktoren, die Kuh als zufälliger Faktor und der Laktationstag als quantitative Variable berücksichtigt.

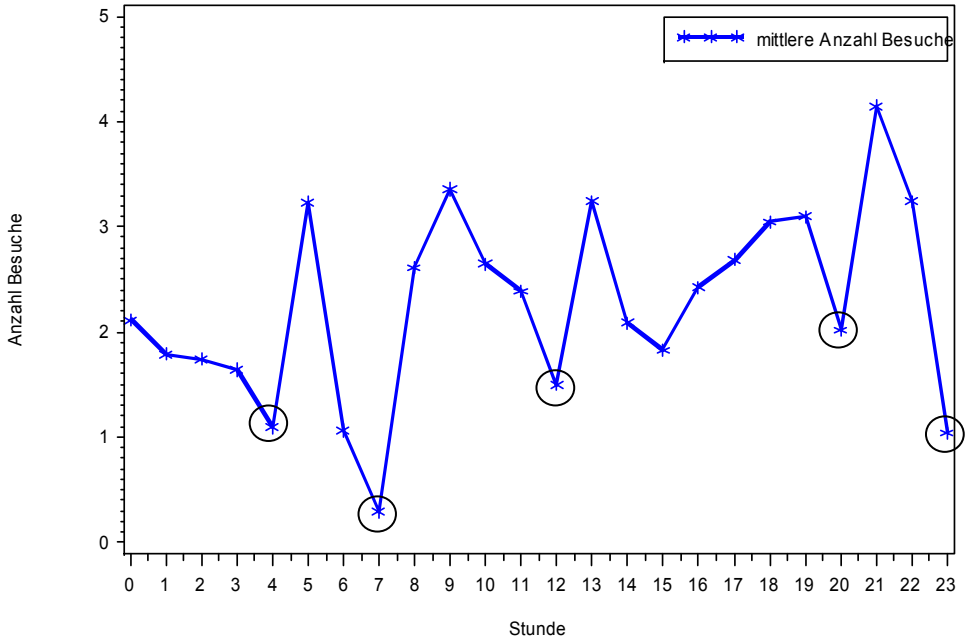

Abbildung 2: Anzahl Besuche pro Stunde in Abhängigkeit der Tagesstunden

Um ein geeignetes Modell für das Untersuchungsmerkmal zu finden, wurden sechs verschiedene Verteilungsannahmen wie die Poissonverteilung, die Negative Binomialverteilung und die Zero-Inflated- und Hurdle-Modelle, basierend auf der Poisson- und der negativen Binomialverteilung, verglichen. Für die unterschiedlichen Modelle wurde der Ansatz für den linearen Prädiktor identisch gewählt. Dabei ist y_{ij} die Anzahl der Besuche am Futterautomat zu Ration i und Stunde j. Es wird z. B. im Hurdle-Modell die Wahrscheinlichkeit, dass Y_{ij} den Wert y_{ij} annimmt, wie folgt modelliert:

$$P(Y_{ij} = y_{ij}) = \begin{cases} p_{0,ij} & \text{für } y_{ij} = 0 \\ (1 - p_{0,ij}) \cdot \dfrac{f(y_{ij} | \lambda_{ij})}{(1 - f(0 | \lambda_{ij}))} & \text{für } y_{ij} > 0 \end{cases}$$

$$\text{mit } \operatorname{logit}(p_{0,ij}) = \eta_{ij}^{(1)} \text{ und } \log(\lambda_{ij}) = \eta_{ij}^{(2)}$$

$$\eta_{ij}^{(k)} = \mu^{(k)} + \alpha_{ij}^{(k)} + \sum_{m=1}^{4} b_{mi}^{(k)} \cdot x_m(t_{ij}) \quad \text{mit } k = 1,2$$

Die Verteilungsparameter $p_{0,ij}$ und λ_{ij} sind über die Logit- bzw. Log-Linkfunktion mit dem linearen Prädiktor $\eta_{ij}^{(1)}$ bzw. $\eta_{ij}^{(2)}$ verknüpft. In diesem Modellansatz bezeichnet μ das allgemeine Mittel, α_{ij} den festen Effekt für die Kombination von Ration i und Stunde j, b_{mi} die fixen Regressionskoeffizienten innerhalb Ration i, $x_m(t)$ ein Legendre Polynom m-ten Grades und t_{ij} den auf das Intervall [-1,1] standardisierten Laktationstag. Das geeignete Modell wurde über den Vergleich der Wahrscheinlichkeiten P(Y=m), geschätzt aus der Stichprobe, und den mittleren Wahrscheinlichkeiten:

$$\overline{P}(Y = m) = \frac{1}{N} \cdot \sum_{k=1}^{N} P(Y_k = m),$$

vorhergesagt mit dem unterstellten Schätzmodell, ermittelt (vgl. [3]). Dazu erfolgte für jeden Datensatz mit Hilfe des Auswertungsmodells die Berechnung der zugehörigen Wahrscheinlichkeiten $\hat{P}_{km} = \hat{P}(Y_k = m)$ mit m=0,1,...,26 unter Verwendung der geschätzten Modellparameter, spezifisch für jeden Datensatz k (mit k=1 bis N). Durch Mittelwertbildung über $\hat{P}_{km}$ ergeben sich dann Vorhersagen für die relative Häufigkeit der im Versuch beobachteten Merkmalsausprägung.
Zusätzlich wurde über die Informationskriterien von Akaike und Schwarz (vgl. [1] und [10]) der endgültige Modellansatz gewählt.

3 Numerische Umsetzung in SAS

In der vorliegenden Arbeit wurden die Prozeduren GLIMMIX, GENMOD und NLMIXED verwendet. In Tabelle 1 sind die Modelle zur Berücksichtigung des Nullenüberschusses aufgeführt und die Möglichkeit der Umsetzung innerhalb der ausgewählten SAS-Prozeduren dargestellt.

Tabelle 1: Implementierung von Modellen mit Nullenüberschuss in ausgewählten SAS-Prozeduren

Prozedur	ZIP	ZINB	HP	HNB
GENMOD	ja	nein	nein	nein
GLIMMIX	nein	nein	ja[1]	ja[1]
NLMIXED	ja[1]	ja[1]	ja[1]	ja[1]

[1] Berücksichtigung von zufälligen Effekten im linearen Prädiktor möglich

Die numerische Umsetzung von ZI- und Hurdle-Modellen innerhalb von NLMIXED ist bei [5] beschrieben. Die Parameterschätzung für die ZIP-Modelle lässt sich zusätzlich mit der Prozedur GENMOD umsetzen. Hier ist allerdings zu beachten, dass keine zufälligen Effekte und somit keine wiederholten Messungen pro Objekt berücksichtigt werden können.
Für die Hurdle-Modelle wurde die numerische Umsetzung in GLIMMIX mit einem Zwei-Schritt-Verfahren gelöst. Das unten aufgeführte Beispiel bezieht sich auf das HNB-Modell. Die Log-Likelihoodfunktion der HNB-Verteilung besteht aus zwei Summanden, wobei der erste Summand nur von $p_{0,ij}$ abhängt (vgl. [5]). Folglich kann die Parameterschätzung mit GLIMMIX in zwei Schritten erfolgen. Im ersten Auswertungsschritt werden die Modellparameter im linearen Prädiktor $\eta_{ij}^{(1)}$ von $p_{0,ij}$ geschätzt. Dies gelingt durch Einführung der binären Variable "visit" mit visit=0, wenn kein Besuch stattgefunden hat und visit=1, falls mindestens ein Besuch stattgefunden hat. Mit Hilfe der id-Anweisung werden die geschätzten Werte für $\eta_{ij}^{(1)}$ in eine SAS-interne Datei abgespeichert und stehen somit zur Weiterverarbeitung für den zweiten Schritt zur Verfügung. Im zweiten Auswertungsschritt muss die Log-Likelihoodfunktion der HNB-Ver-

teilung vom Nutzer programmiert werden. Die NB-Verteilung hängt neben λ_{ij} zusätzlich vom Verteilungsparameter α ab. Da α nur Werte zwischen Null und Eins annehmen darf, wählt man α in Abhängigkeit des SAS-internen Skalierungsfaktor _phi_ zweckmäßig wie folgt (vgl. [9]):

$$\alpha = 1 - 1/\exp(\_phi\_).$$

Die nachfolgenden SAS-Anweisungen demonstrieren die Vorgehensweise an einem ausgewählten Beispiel.

```
/*1. Schritt */
proc glimmix data=besuch1 method=quad ;
CLASS trt stunde;
MODEL visit(event='0')= trt*stunde x1(trt) x2(trt) x3(trt) x4(trt)/
dist=binary link=logit noint ;
output out=df0_out pred=eta0 residual=r;
id anzahl visit trt stunde x1 x2 x3 x4;
RUN;

/*2. Schritt */
proc glimmix data=df0_out method=quad;
CLASS trt stunde;
MODEL anzahl= trt*stunde x1(trt) x2(trt) x3(trt) x4(trt)/ link=log
noint;
alp=1-1/exp(_phi_);
if(_mu_=.) or (_linp_=.) then _logl_=.;
else do;
 p0=exp(eta0)/(1+exp(eta0));
 p1=1.0-p0;
 if(p0>1E-12) then log_p0=log(p0); else log_p0=-1E20;
 if(p1>1E-12) then log_p1=log(p1); else log_p1=-1E20;
 if(anzahl=0) then _logl_=log_p0;
 else do;
  f1=lgamma(anzahl+1/alp)-lgamma(1/alp)-lgamma(anzahl+1);
  mu1=alp*_mu_;
  if(mu1>1E-12) then log_m1=log(mu1); else log_m1=-1E20;
  mu2=1+alp*_mu_;
  if(mu2>1E-12) then log_m2=log(mu2); else log_m2=-1E20;
  f2=anzahl*log_m1;
  f3=(anzahl+1/alp)*log_m2;
  log_f1=f1+f2-f3;
  f0=1-mu2**(-1/alp);
  if(f0>1E-12) then log_f0=log(f0); else log_f0=-1E20;
  _logl_=log_p1+log_f1-log_f0;
 end;
end;
output out=df_out pred=p residual=r;
RUN;
```

4 Ergebnisse

Die Auswahl einer geeigneten Verteilungsannahme erfolgte zunächst über einen grafischen Vergleich der beobachteten und der vorhergesagten Wahrscheinlichkeiten. Im folgenden Abschnitt werden dazu nur ausgewählte Modelle vorgestellt. In Abbildung 3 ist zu erkennen, dass das ZI-Modell mit negativer Binomialverteilung im Vergleich zu den beobachteten Wahrscheinlichkeiten im Besonderen das Ereignis Null als auch die Ereignisse Zwei bis Sechs nicht adäquat wiederspiegelt.

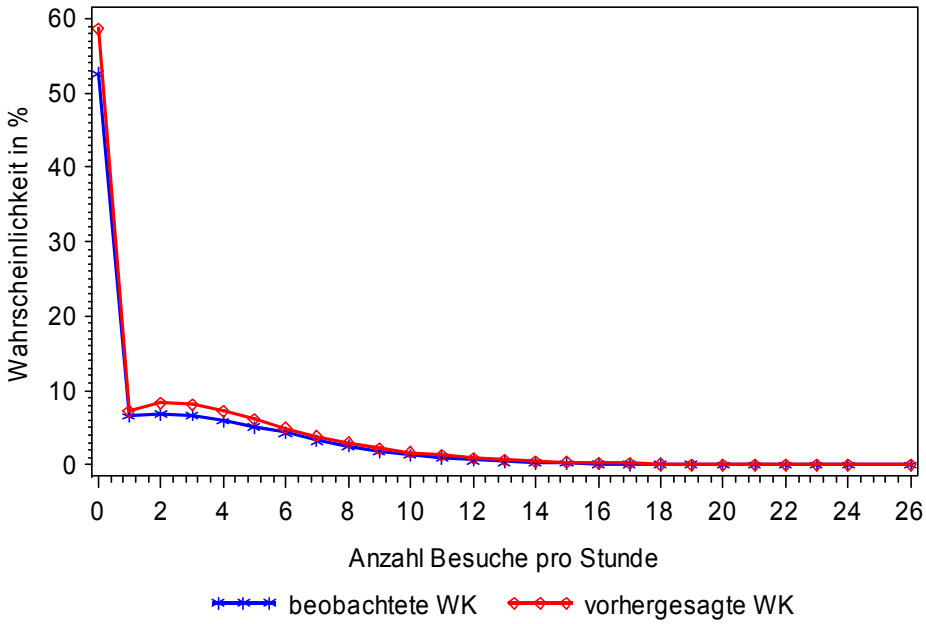

Abbildung 3: ZI-Modell mit negativer Binomialverteilung

Dagegen zeigte das Hurdle-Modell mit negativer Binomialverteilung eine sehr gute Übereinstimmung der beobachteten und vorhergesagten Werte (siehe Abbildung 4).

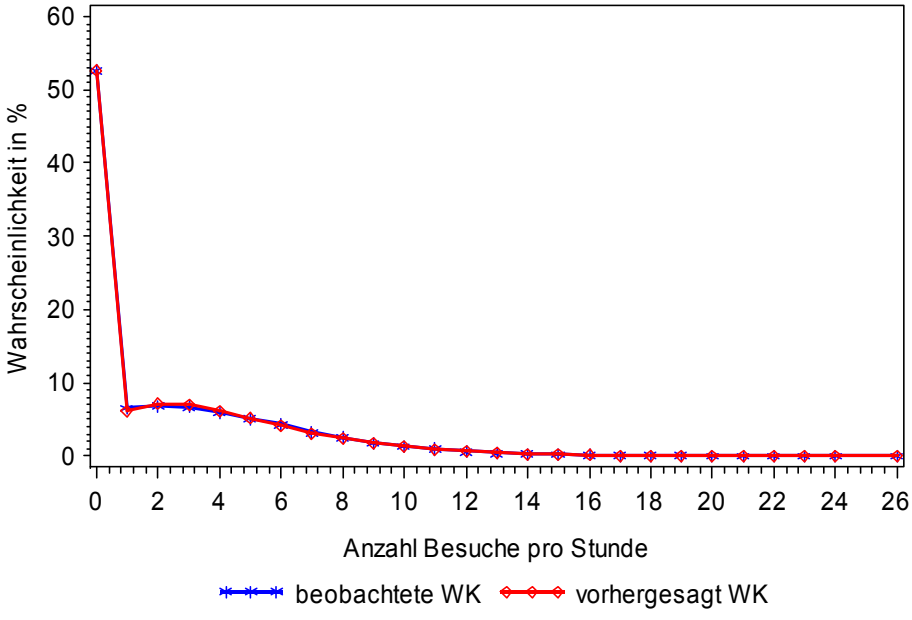

Abbildung 4: Hurdle-Modell mit negativer Binomialverteilung

Zusätzlich wurden die Modelle anhand von Informationskriterien wie dem AIC und dem BIC miteinander verglichen. Aus Tabelle 2 wird ersichtlich, dass sowohl AIC als auch BIC das HNB-Modell bevorzugen. Weiterhin zeigt Tabelle 2, dass der Übergang von Poisson- zur negativen Binomialverteilung zu einer besseren Modellanpassung führt.

Tabelle 2: Modellvergleich mittels Informationskriterien

Modell (Verteilung)	d	Statistiken		
		-2LogL	ΔAIC	ΔBIC
Poisson	56	247304	88067	87569
NB	57	169238	10003	9514
ZIP	112	167402	8277	8268
HP	112	167171	8046	8037
ZINB	113	159217	94	95
HNB	113	159123	0	0

AIC, BIC als Abweichungen vom kleinsten Wert (HNB-Modell)

d - Anzahl der Modellparameter

5 Diskussion

Die durchgeführte Modellwahl bevorzugte Hurdle-Modelle gegenüber den ZI-Modellen. Dagegen kommen in der Tierzucht zur Auswertung von Zählmerkmalen zunehmend ZI-Modelle zum Einsatz (vgl. [7] und [8]). Beispielsweise lässt sich für die Anzahl klinischer Mastitisfälle pro Kuh im Verlauf einer Laktation ein Nullenüberschuss beobachten (vgl. [8]). Das Ereignis Null entsteht, weil einerseits in der Population bereits eine Gruppe von Tieren Resistenz gegenüber Mastitis aufweist. Andererseits existieren Tiere, deren Merkmalsausprägungen abhängig von verschiedenen Umwelteinflüssen, einer Poissonverteilung genügen, so dass zusätzlich Nullen auftreten können. Die in der Tierzucht häufige Interpretation zur Erklärung des Nullenüberschusses lässt sich nicht auf das vorliegende Beispiel übertragen. Daraus ergibt sich ein zusätzliches Argument für die Anwendung der Hurdle-Modelle.

Die Auswahl einer geeigneten Verteilungsannahme sollte nicht nur anhand der Log-Likelihood und den Informationskriterien AIC und BIC erfolgen. Sinnvoll ist hierfür zusätzlich einen grafischen Vergleich der beobachten und vorhergesagten Wahrscheinlichkeiten durchzuführen.

Für die unterschiedlichen Verteilungsannahmen stehen verschiedene Prozeduren zur Verfügung. Der Nachteil der Prozedur NLMIXED gegenüber GLIMMIX besteht einer-

seits im hohen Programmieraufwand, da keine class-Anweisung existiert. Außerdem kommt es zu langen Rechenzeiten und Konvergenzproblemen, wenn fixe Einflussfaktoren mit hohen Stufenzahlen auftreten und zusätzlich zufällige Effekte berücksichtigt werden müssen. Der Vorteil besteht allerdings in der estimate-Anweisung, die im Gegensatz zur estimate-Anweisung der Prozedur GLIMMIX, auch auf nichtlineare Funktionen der Modellparameter angewendet werden kann. Während in GLIMMIX mit der estimate-Anweisung nur Hypothesen über die Effekte im linearen Prädiktor geprüft werden können, gestattet die estimate-Anweisung innerhalb von NLMIXED die Überprüfung von Hypothesen formuliert unter Verwendung der inversen Linkfunktion. Somit ist in NLMIXED die Berechnung von Standardfehlern beispielsweise für die geschätzten Erwartungswerte, basierend auf dem Hurdle-Modell innerhalb der Stufenkombinationen von Einflussfaktoren, nutzerfreundlich realisierbar.

Die hier vorgestellten Ergebnisse dienten der Auffindung eines geeigneten Verteilungstyps zur Beschreibung des untersuchten Zählmerkmals. Deshalb wurden die festen Effekte der Testtage und der zufällige Effekt der Kuh nicht im linearen Prädiktor berücksichtigt. Die Einbeziehung dieser Einflussgrößen und somit der Übergang zu einem generalisierten linearen gemischten Modell ist jedoch innerhalb der hier vorgestellten Zwei-Schritt-Prozedur bei akzeptablen Rechenzeiten und guten Konvergenzverhalten relativ einfach durchführbar. Allerdings muss im Zwei-Schritt-Verfahren vorausgesetzt werden, dass die zufälligen Tiereffekte im linearen Prädiktor der Verteilungsparameter $p_{0,ij}$ und λ_{ij} voneinander unabhängig sind. Soll auf die Voraussetzung der Unabhängigkeit verzichtet werden, so kann die Prozedur GLIMMIX zumindest genutzt werden, um geeignete Startwerte für die Regressionskoeffizienten in NLMIXED vorzugeben.

Literatur

[1] Akaike, H.: Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory BN Petrov and F Csaki ed Akademiai Kiado Budapest Hungary (1973): 267-281.

[2] Bulang, M.; Kluth, H.; Engelhard, T., Spilke, J., Rodehutscord, M.: Zum Einsatz von Luzernesilage bei Kühen mit hoher Milchleistung. Journal of Animal Physiology and Animal Nutrition 90 (2006): 89-102.

[3] Erdman, D.; Jackson, L.; Sinko, A.: Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure. SAS Global Forum, Paper 322 (2008): 1-11.

[4] Lambert, D: Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 34 (1992): 1-14.

[5] Liu, W; Cela, J: Count Data Models in SAS. SAS Global Forum, Paper 317 (2008): 1-12.

[6] Mullahy, J: Specification and Testing of Some Modified count Data Models. Journal of Ecometrics 33 (1986): 341-365.

[7] Naya,H.; Urioste,J.I.; Chang,Y.M.; Rodrigues-Motta,M.; Kremer,R.; Gianola,D: A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep. Genetics Selection Evolution 40 (2008): 379-394.

[8] Rodrigues-Motta, M.; Gianola, D.; Heringstad, B.; Rosa, G.J.M.; Chang, Y.M.: A zero-inflated poisson model for genetic analysis of the number of mastitis cases in Norwegian red cows. Journal of Dairy Science 90 (2007): 5306-5315.

[9] Schabenberger, O.: Example 38.14 Generalized Poisson Mixed Model for Overdispersed Count Data. SAS-Hilfe für PROC GLIMMIX in SAS 9.2 (2008).

[10] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6 (1978): 461-464.