Archiv Tierzucht 54 (2011) 6, 661-675, ISSN 0003-9438

Generalized linear models with random effects for the description of data with excess zeros

Norbert Mielenz[1], Katrin Thamm[1], Michael Bulang[2] and Joachim Spilke[1]

[1]Biometrics and Informatics in Agriculture Group, [2]Research Centre for Animal Sciences Merbitz, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Abstract

In this paper count data with excess zeros and repeated observations per subject are evaluated. If the number of values observed for the zero event in the trial substantially exceeds the expected number (derived from the Poisson or from the negative binomial distribution), then there is an excess of zeros. Hurdle and zero-inflated models with random effects are available in order to evaluate this type of data. In this paper both model approaches are presented and are used for the evaluation of the number of visits to the feeder per cow per hour. Finally, for the analysis of the target trait a hurdle model with random effects based on a negative binomial distribution was used. This analysis was derived from a detailed comparison of models and was needed because of a simpler computer implementation. For improved interpretation of the results, the levels of the explanatory factors (for example, the classes of lactation) were not averaged in the link scale, but rather in the response scale. The deciding explanatory variables for the pattern of visiting activities in the 24-hour cycle are the milking and cleaning times at hours 4, 7, 12 and 20. The highly significant differences in the visiting frequencies of cows of the first lactation and those of higher lactations were explained by competition for access to the feeder and thus to the feed.

Keywords: excess zeros, hurdle and zero-inflated models

Zusammenfassung

Generalisierte lineare Modelle mit zufälligen Effekten zur Beschreibung von Daten mit Nullenüberschuss

In vorliegender Arbeit wird die Auswertung eines Zählmerkmals bei Vorliegen von Nullenüberschuss und Messwiederholung demonstriert. Übersteigt die Anzahl der im Versuch beobachteten Werte für das Ereignis Null die erwartete Anzahl abgeleitet aus der Poisson- oder der negativen Binomialverteilung erheblich, so liegt Nullenüberschuss vor. Zur Auswertung derartiger Daten stehen Hurdle und Zero-Inflated Modelle mit zufälligen Effekten zur Verfügung. In dieser Arbeit werden beide Modellansätze vorgestellt und zur Auswertung des Merkmals Anzahl Besuche einer Kuh am Fütterungsautomat pro Stunde eingesetzt. Abgeleitet aus einer ausführlichen Modellüberprüfung und bedingt durch die einfachere rechentechnische Umsetzung wurde zur Analyse des Zielmerkmals letztendlich ein Hurdle-Modell mit zufälligen Effekten bei Annahme von negativer Binomialverteilung verwendet. Zur besseren Interpretation der Ergebnisse erfolgte die Mittelwertbildung über die Stufen

der Einflussfaktoren (beispielsweise der Laktationsklassen) nicht in der Link- sondern in der Response-Skala. Die entscheidenden Einflussgrößen für das Muster der Besuchsaktivitäten im 24-Stundenzyklus sind die Melk- und Reinigungszeiten zu den Stunden 4, 7 12 und 20. Die hoch signifikanten Unterschiede in den Besuchsfrequenzen von Kühen der ersten und Kühen höherer Laktationen wurden mit der Konkurrenz um den Zugang zum Trog und somit um das Futter erklärt.

Schlüsselwörter: Nullenüberschuss, Hurdle und Zero-Inflated Modelle

Introduction

The evaluation of count data typically occurs using the Poisson or the negative binomial distribution. However, if there is an excess of zeros, then the standard distributions mentioned can no longer be used. In livestock breeding, for instance, excess zeros can be observed for the number of clinical cases of mastitis per cow during the lactation (Rodrigues-Motta *et al.* 2007). Zero events are observed for a group of animals that are already resistant to mastitis. For other cows zero is observed according to a random variable with a Poisson distribution. An example that can be similarly explained comes from the medical field in the form of the number of visits to the doctor per person in a certain period (Min & Agresti 2005). For some people the observation of zero occurs through random processes, whilst other individuals might decline going to the doctor, for example, because of a phobia. To generalize, one can imagine a population consisting of two groups. In the first group the objects for the count data can only reflect the zero, whilst in the other group the observations per object suffice for a discrete distribution for count traits. For the modelling of this circumstance, zero-inflated Poisson (ZIP) models and zero-inflated negative binomial (ZINB) models have been developed in the literature (Lambert 1992). In the following these are abbreviated as ZI models. In contrast to this the hurdle model (Mullahy 1986) is a model consisting of two parts for count data. The first part consists of a binary model, where the response outcome is »zero« or »greater than zero«. The second part uses a truncated model, which modifies a distribution for count traits so that only positive events can occur. In this paper these different models are used to evaluate the number of visits to the feeder per hour. The observation trait was recorded within a feeding experiment over a period of 141 days, at all hours of the day for 22 cows. On average, over all of the hourly observations, the zero event (no visit per hour) occurred in about 50 % of the cases. In contrast, the positive number of visits varied between 1 and 12, whereby the corresponding relative frequency stayed under 7 %. Due to the repeated sampling of traits, the trial data have a longitudinal character so that hurdle or ZI models with random effects (Hall 2000, Min & Agresti 2005) must be used. For example, Yau & Lee (2001) analyse the number of work injuries of persons using a hurdle-Poisson model with random effects. Solution methods and uses of ZINB models with random effects can be found in Yau *et al.* (2003) and Xiang *et al.* (2007). Lee *et al.* (2006) present ZIP models with a hierarchical structure for the random effects.

The aim of the evaluation is (a) the selection of a suitable model for the description of the excess zeros in the number of visits per hour in the 24-hour cycle, (b) the modelling of this trait, dependent on the lactation stage, (c) the quantification of differences between the cows

of the first lactation (primiparous cows) and higher lactations (multiparous cows) and (d) the exposure of differences between day and night hours. In addition, the hourly observation of the number of visits per cow allows the investigation of the following questions. Which pattern is shown by the visiting activity as a function of the hours of a day? To what extent do milking and cleaning times influence the visiting frequency of the cows? Do differences exist in the visiting frequency of primiparous and multiparous cows, which have consequences for the housing of primiparous and multiparous cows and for the organisation of ratio between animal and feeder?

Materials and methods

Data and animals

In a feeding trial not only the milk yield and the feed intake per day were recorded, but also traits of the feeding behaviour. The recording of feed intake took place using computer-supported automatic feeders, which were equipped with electronic animal recognition devices using transponders. This facilitated an individual recording of feeding behaviour per cow, characterised by the feed intake per visit, the number and duration of visits at the feeder. Derived from the questions formulated in the introduction, we limited ourselves to the evaluation of the number of visits per cow per hour in this study. In order to calculate this trait, all visits of a cow registered within an hour at the feeder were added together. The alfalfa-mixed ration was offered to the animals in all 12 feeders. The evaluation of performance traits, such as the milk yield and the content of the milk in connection with the traits of the feeding behaviour through extrapolation of all investigation traits to daily observations can be found in Bulang *et al.* (2006) and Thamm *et al.* (2011). 22 cows were included in the trial; these were examined over a period of 141 days. All cows were given a mixed ration with a high proportion of alfalfa. The 7 cows of the first lactation and the 15 cows with two or more lactations were each grouped together in one class. In the following, primiparous and multiparous cows are differentiated according to the class affiliation. Table 1 shows the absolute frequency of the number of visits per hour for the cows of the first (CL 1) and the second class of lactation (CL 2)

Table 1
Absolute frequency for the number of visits per hour (trait Y) dependent on the class of lactation (CL)

	Number of observations with Y=k							
Class	k=0	1	2	3	4	5	6 to 12	n
CL 1	11 843	1 336	1 405	1 434	1 270	1 074	938 to 190	22 272
CL 2	24 584	2 905	3 052	2 728	2 514	1 952	1 537 to 161	42 456

From Table 1 it is clear that cows from the CL 1 or CL 2 with a relative frequency of 53.2 % or 57.9 % do not frequent the feeder within an hour. The relative frequency for the occurrence of 12 visits lies under one percent in CL 1 and in CL 2 is already under 0.5 %.

However, no measurements were available in order to derive a circadian rhythm for the visiting activity of the cows dependent on external stimuli such as light and temperature. The analysis of the averages per hour showed a very strong dependence of the visiting activity

on the three milking times in the hours 4, 12 and 20, as well as on the cleaning time of the feeders, mainly in hour 7. Within the hours 4, 7, 12 and 20 the limited access to the feeder, dependent on the sequence of milking and the location in the stable can vary from cow to cow, so that visits can also occur in these hours. The milking times and the cleaning times will be combined in the term »service time« from now on. A formal division of the 24-hour cycle into service, day and night hours was made using the service time in hour 7 and the service time in hour 20. The average number of visits in the night (h>20 and <7) and the day hours (h>7 and <20) with the exception of the service times (h=4, 7, 12 und 20) are shown in table 2.

Table 2
Average number of visits per hour (±SD) in the service, day and night hours and over all hours for three selected periods dependent on the class of lactation (CL)

	hours, time of day			Period, days in milk		
class	service-time	night-time	day-time	0, 60	60, 100	100, 200
CL 1	1.27±2.42	2.47±3.69	2.75±3.64	2.33±3.31	2.54±3.65	2.38±3.56
CL 2	1.06±1.99	1.63±2.79	2.10±2.84	1.51±2.32	1.93±2.92	1.78±2.78

According to table 2 the number of visits is influenced not only by the time of day and the CL, but also by the day of lactation, i.e. the time after calving. It can be assumed that cows show an increased frequency of visits at the time of the highest milk yield, experience showing that this is at about the 40th day of lactation. The standard deviations shown in table 2 are a manifestation of the large variability of the visiting frequency between the hours, the cows and the day of lactation. The days in milk (DIM) varies in the sample due to the different calving dates of the cows. For primiparous cows observations are available between DIM 15 and 224 and for multiparous cows between DIM 8 and 229. The following information signifies the high performance level of the herd. In the trial period, the primiparous cows achieved an average daily energy corrected milk (ECM) of 30.4 kg/d with an average daily dry matter intake (DMI) of 20.3 kg/d. With a DMI of 24.5 kg/d, the multiparous cows achieved an ECM of 41.4 kg/d. For the primiparous cows the daily ECM values varied between 20 and 40 kg/d and for the multiparous cows between 25 and 61 kg/d. The observed daily DMI values lay between 10.2 and 30.9 kg/d for the primiparous cows and between 10.1 and 40.4 kg/d for the multiparous cows.

Models for the evaluation of count data with excess zeros

Hurdle and ZI models, each based on Poisson or negative binomial distribution come into question for the evaluation of the trait *Y*, explained in the »data and animals« section. Within a day, the analysis of the average value per hour did not show any repetition with certain period lengths that can be described by the overlaying of a few sine and cosine functions (see figure 9 and 10 in the results section). Thus, the hourly influence was taken into consideration by fixed effects in the evaluation model. Let $y_{ijk}(t)$ be an observation of cow k (k=1,…,n_i) at the hour j (j=1,…,24) from CL i (i=1,2) on day of lactation t and let $Y_{ijk}(t)$ denote a random variable associated with the observations $y_{ijk}(t)$ of the trial.

1. The hurdle model with random effects

In the hurdle model based on the negative binomial distribution the probability that Y_{ijk} on day of lactation t will take the value y_{ijk} (y_{ijk}=0,1,2,…) is modelled as follows:

$$P(Y_{ijk}=y_{ijk}\mid t)=\begin{Bmatrix} p_{0,ijk} & for\ y_{ijk}=0 \\ (1-p_{0,ijk})\, f_{TNB}(y_{ijk}\mid \lambda_{ijk},\alpha) & for\ y_{ijk}>0 \end{Bmatrix} \quad (1)$$

In (1) $f_{TNB}(\cdot)$ denotes the density function of the truncated negative binomial distribution. If $f_{NB}(\cdot)$ is the density of the negative binomial distribution, then the following is valid:

$$f_{TNB}(y\mid\lambda,\alpha)=\frac{f_{NB}(y\mid\lambda,\alpha)}{(1-f_{NB}(0\mid\lambda,\alpha))}$$
$$with\ f_{NB}(y\mid\lambda,\alpha)=\frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\cdot y!}\cdot(1+\alpha\cdot\lambda)^{-\frac{1}{\alpha}}\left|\frac{\alpha\cdot\lambda}{1+\alpha\cdot\lambda}\right|^{y} \quad (2)$$

The distribution parameters are dependent on the explanatory variables as follows:

$$logit(p_{0,ijk})=\eta_{0,ij}+u_{ik}\quad with\quad \eta_{0,ij}=\beta_{0,ij}+\beta_{1,i}\cdot x_1(t)+\cdots+\beta_{4,i}\cdot x_4(t)$$
$$log(\lambda_{ijk})=\eta_{1,ij}+v_{ik}\quad with\quad \eta_{1,ij}=\alpha_{0,ij}+\alpha_{1,i}\cdot y_1(t)+\cdots+\alpha_{4,i}\cdot y_4(t) \quad (3)$$

The probabilities defined in formula (1) are conditional probabilities given the random effects u_{ik} and v_{ik} of a cow from class i. For the covariates $x_1(t)$ to $x_4(t)$ in (3) an approach with sine and cosine functions proved to be advantageous in comparison with an approach with polynomials of the fourth degree. Legendre polynomials of the first to the fourth degree were used for improvement of convergence for the covariates $y_1(t)$ to $y_4(t)$.

For example the following is valid:

$$x_i(t)=sin(2\pi\cdot t/T_i)\qquad i=1,2$$
$$x_{2+i}(t)=cos(2\pi\cdot t/T_i)\quad \text{with } T_1=300;\ T_2=100 \quad (4)$$

Through the special selection of T_1 and T_2 the lactation lengths and the occurrence of two local minima with an interval of about 100 days were considered (compare figure 1 and 2). The regression coefficients of the covariates in (3) were seen as being specific for each CL. Normal distribution was assumed and postulated for the vectors *u* and *v* of the random animal effects.

$$Var(u)=I_q\cdot\sigma_u^2\qquad and\qquad Var(v)=I_q\cdot\sigma_v^2\qquad q=number\ of\ animals \quad (5)$$

The maximization of the log-likelihood function of model (1) in two steps for $p_{0,ijk}$ and (λ_{ijk}, α) is achieved through the assumption cov(u_{ik},v_{ik})=0 (compare Min & Agresti 2005). The conditional expected value of CL i and hour j on day of lactation t can be calculated by the following formula:

$$\mu_{ijk}(t)=E(Y_{ijk}\mid u_{ik},v_{ik},t)=\frac{(1-p_{0,ijk})}{(1-f_{NB}(0\mid\lambda_{ijk},\alpha))}\cdot\lambda_{ijk} \quad (6)$$

Let $\text{cov}(u_{ik}, v_{ik})=0$ be, then the marginal expected values can be calculated by using the following approximations (compare Ritz & Spiegelman 2004):

$$\mu_{ij}(t) \approx \frac{(1-p_{0,ij})}{(1-f_{NB}(0|\lambda_{ij},a))} \cdot \lambda_{ij}$$

$$\text{with } p_{0,ij}(t) \approx \frac{\exp(a \cdot \eta_{0,ij})}{1+\exp(a \cdot \eta_{0,ij})} \qquad a = \frac{1}{\sqrt{(1+c^2 \cdot \sigma_u^2)}} \qquad c = \frac{16 \cdot \sqrt{3}}{15 \cdot \pi} \tag{7}$$

$$\text{and } \lambda_{ij} = \exp(\eta_{1,ij} + 0.5 \cdot \sigma_v^2)$$

2. The zero-inflated model with random effects

The ZI model for the negative binomial distribution has the form:

$$P(Y_{ijk} = y_{ijk} \mid t) = \begin{cases} p_{0,ijk} + (1-p_{0,ijk}) \cdot f_{NB}(0 \mid \lambda_{ijk}, a) & \text{for } y_{ijk} = 0 \\ (1-p_{0,ijk}) \cdot f_{NB}(y_{ijk} \mid \lambda_{ijk}, a) & \text{for } y_{ijk} > 0 \end{cases} \tag{8}$$

If one allows correlations between the random effects of an animal, then two-dimensional improper integrals are to be calculated numerically for the establishment of the log-likelihood function (Min & Agresti 2005).

Computational implementation

The parameter estimation in the hurdle and ZI models was carried out using the statistics programme SAS 9.2 (SAS Institute Inc., Cary, NC, USA) using the procedures GENMOD, GLIMMIX and NLMIXED. ZIP models without random effects are directly implemented in GENMOD as standard through the option distribution=ZIP. In GLIMMIX, the implementation of hurdle models occurs through entering a user-defined log-likelihood function for the truncated distribution of the count trait with values larger than zero. The parameter estimation in the hurdle models can be carried out in two steps. In the first step the binary trait (»visit« or »no-visit«) is analysed and the parameters estimated in the linear predictor of the logit scale. In the second step the maximization of the likelihood of the truncated Poisson or negative binomial distribution occurs with given model parameters for modelling $p_{0,ijk}$ (formula 1). In our example 56 fixed model parameters on each step are estimated. The number 56 arises from the sum of the number of CL multiplied by the number of hours, and the number of CL multiplied by the number of covariates. The implementation of hurdle and ZI models based on the Poisson or negative binomial distribution within NLMIXED is well known (Liu & Cela 2008). In general, NLMIXED can also be used to formulate ZIP and ZINB models with random effects. Unfortunately, not only long calculating times but also large problems with convergence already occurred with the ZIP and ZINB models with only fixed effects. In the dataset presented here, the simultaneous estimation of 112 fixed model parameters and at least 3 parameters of the variability present a prerequisite for the fitting of a ZINB model (Table 4). The maximum likelihood method implemented in NLMIXED requires the numerical solution of improper integrals in each iteration step if random effects are presented (Min & Agresti 2005). NLMIXED was not able to fulfil this requirement for the dataset available in an acceptable calculation

time (under one week). Apart from the estimation of model parameters with the help of the ML methods using Gaussian quadrature formulas, numerous approximations exist, which circumvent the numerical solution of multi-dimensional integrals. For instance, in connection with the ZI models, the deployment and maximization of the »penalised quasi-likelihood« is recommended according to the theory of the generalized linear mixed models (Yau *et al.* 2003, Lee *et al.* 2006). The »penalized quasi-likelihood«, or also the term »best linear unbiased prediction (BLUP) type likelihood«, is made up of two terms (McGilchrist 1994). The first corresponds to the log-likelihood function of the ZIP model under the assumption that the random effects can be seen as being given and fixed. The second term is derived from the density function of the random model effects and takes the interpretation of a penalty function. The estimation of the model parameters occurs through the use of an »expectation-maximisation« (EM) algorithm using partial derivations of the second order. The use of the EM algorithm enables the partition of the log-likelihood into two parts, which can be maximized independently of one another. The methods shown were implemented, for example, by Lee *et al.* (2006) and Xiang *et al.* (2007) in S-Plus (TIBCO Software Inc., Palo Alto, CA, USA) R (R Foundation, Vienna, Austria) macros. As an alternative to the commercial software packages S-Plus and SAS, the freely available software was included in the investigation. If one limits oneself to the fixed model parameters, the R functions zeroinfl(·) and hurdle(·) within the library pscl offer an alternative to the commercial solutions. According to the knowledge of the authors, hurdle and ZI models with random effects, for instance using the »penalised quasi-likelihood« methods have not yet been implemented in R.

Model comparison

For given animal effects the random variable $Y_{ijk}(t)$ suffices for the distribution assumption fixed through the hurdle or ZI model. If a negative binomial distribution is supposed then a triple of the estimated distribution parameters $(\hat{p}_{0,k}(t), \hat{\lambda}_k(t), \hat{\alpha})$ and a vector from the predictions $\hat{u}_k, \hat{v}_k$ can be assigned to each record k. Using this parameter set the probabilities $P(Y_k=y|u_k,v_k)$ with $y = 0,1,2,\ldots$ can be estimated for each record, for example with the formulas (1) or (8). The following estimations for the relative frequency are given through summation over all N records.

$$\hat{P}(Y=y) = \frac{1}{N} \cdot \sum_{k=1}^{N} \hat{P}(Y_k = y \mid \hat{u}_k, \hat{v}_k) \qquad \text{with } y = 0, 1, 2, \ldots \tag{9}$$

Through comparison with the frequencies estimated by formula (1) and the relative frequency observed in the sample, one obtains first information about the goodness of fit of the assumed evaluation model. The observed relative frequencies, estimated using formula (9) are given in Table 3.

According to Table 3, the observed relative frequencies are shown well, not only through the hurdle model, but also through ZI models based on the negative binomial distribution. In contrast, no sufficient fitting could be achieved for either of the models with the use of the Poisson distribution. The relative frequencies estimated by the HNB and ZINB models are identical. A differentiation between these models can be achieved with the help of the information criterion AIC (Akaike 1973) and BIC (Schwarz 1978). In Table 4 the penalty term of the BIC values was calculated with the help of the number of cows (subjects). According to

Table 4 the ZINB model with only fixed model parameters leads to lower AIC and BIC values than the comparable hurdle model (HNB_fix). Another decisive lowering of the AIC and BIC values is achieved by the transfer to model HNB_rand, that is to a hurdle model with random cow effects based on the negative binomial distribution.

Table 3
Observed and estimated relative frequency for the occurrences 0 to 8 for hurdle and ZI models without random effects

Source	P(Y=k) k=0	1	2	3	4	5	6	7	8
Observ.	56.3	6.55	6.89	6.43	5.85	4.67	3.82	2.77	2.08
HP_fix	56.3	2.91	5.52	7.42	7.85	6.90	5.24	3.51	2.12
HNB_fix	56.3	6.23	7.11	6.77	5.83	4.71	3.64	2.73	1.99
ZIP_fix	56.3	2.91	5.52	7.42	7.84	6.90	5.24	3.52	2.12
ZINB_fix	56.3	6.23	7.10	6.77	5.83	4.71	3.64	2.73	1.99

Table 4
Number of fixed model parameters (p), number of parameters of variability (q), number of cows (s) with −2 multiplied log-likelihood function, AIC and BIC values (relative to HNB_rand) for hurdle (H) and ZI models for the description of excess zeros

Model	Procedure	No. of parameters p	q	s	Statistics -2logL	AIC	BIC
HP_fix	Glimmix	112	0	22	222 447.6	12 068.7	12 065.4
HNB_fix	Glimmix	112	1	22	213 146.2	2 769.3	2 767.1
ZIP_fix	Genmod	112	0	22	222 436.3	12 057.4	12 054.1
ZINB_fix	Nlmixed	112	1	22	213 137.2	2 760.3	2 758.1
HP_rand	Glimmix	112	2	22	217 073.3	6 698.4	6 697.3
HNB_rand	Glimmix	112	3	22	210 372.9	0	0

Unfortunately, it was not possible to fit a ZINB model with random effects to this dataset with NLMIXED. The calculation time lasted several weeks, without achievement of successful convergence.

Results

Comparison of predictions and trend analyses

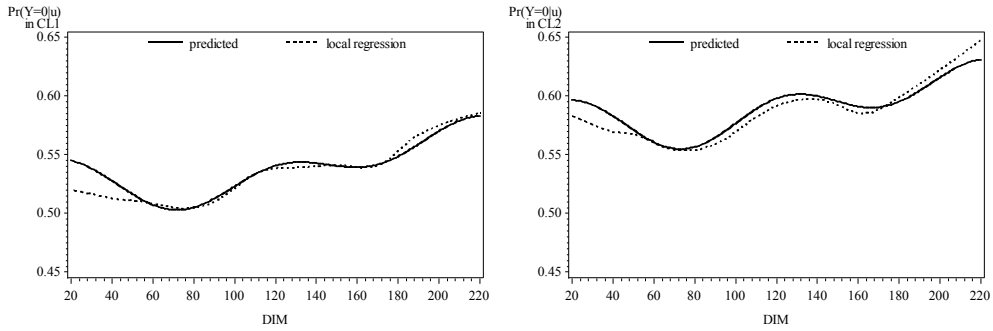
An extra examination of the model exists in the comparison of the predictions for the number of visits on the day of lactation, derived from model HNB_rand with the corresponding number of visits, determined from a trend analysis using local regression. Predictions for the probability that on day of lactation t in CL i no visit occurs, can be calculated as follows:

$$\hat{p}_{0i}(t) = \frac{1}{(b \cdot n_i)} \cdot \sum_{k=1}^{n_i} \sum_{j=1}^{b} \hat{p}_{0,ijk}(t) \qquad with\ \hat{p}_{0,ijk}(t) = h(\hat{\eta}_{0,ij} + \hat{u}_{ik}) \tag{10}$$

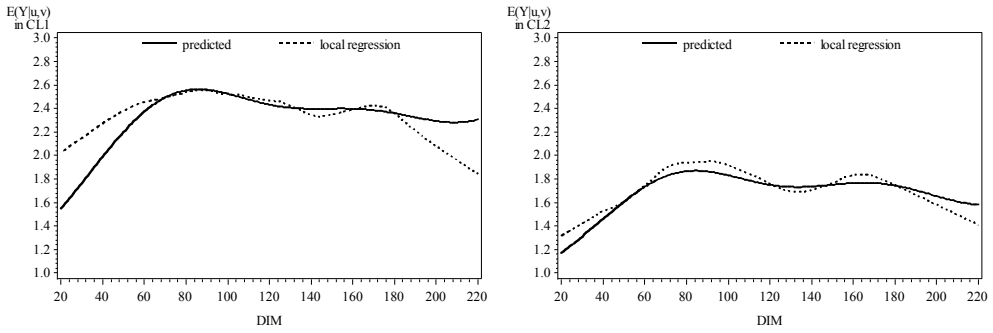
In (11) h(·) stands for the inverse link function in the logit model. Analogously, predictions can be made using formula (6) for the average number of visits per hour for day of lactation t within the two CL.

$$\hat{\mu}_i(t) = \frac{1}{(b \cdot n_i)} \cdot \sum_{k=1}^{n_i} \sum_{j=1}^{b} \hat{\mu}_{ijk}(t) \qquad with\ \hat{\mu}_{ijk}(t) = E(Y_{ijk} \mid \hat{u}_{ik}, \hat{v}_{ik}, t) \tag{11}$$

In order to calculate the functions $\hat{p}_{0i}(t)$ and $\hat{\mu}_i(t)$ the predictions for the random cow effects were used in the linear predictor. The predictions given in formula (10) or (11) can be seen as estimations for the conditional probabilities or conditional expected values. To check the model, the curves based on model estimations were compared with the corresponding smoothed trend curves that had been calculated with the help of local regression. For fitting of the trend curves with the help of the SAS procedure LOESS, generally a smoothing parameter of 0.3 was chosen. In figures 1 and 2 the estimated probabilities (according to formula (10)) and the calculated trend curves are compared to one another. For both CL there is very good agreement for DIM, between 50 and 200. The global minimum for $\hat{p}_{0i}(t)$ is achieved at about the 70th day of lactation, independent of the CL.



Figures 1+2
With model HNB_rand within class of lactation 1 and 2 (CL1 and CL2) estimated probability of zero event (no visit per h) in comparison to smoothed trend curves, dependent on the days in milk (DIM)



Figures 3+4
With model HNB_rand within class of lactation 1 and 2 (CL1 and CL2) estimated average number of visits per hour in comparison to smoothed trend curves, dependent on the days in milk (DIM)

The comparison of the estimated expected value with the corresponding trend curves is shown in figures 3 and 4. There is good agreement between expected values and trend curves between DIM 80 to 180. The largest number of visits is found in both CL at about

90th day of lactation. The reason for the poorer agreement at the beginning and end of the lactation might be the relatively low number of observations in these periods.

The results of figures 1 to 4 are closely connected to the typical course of a lactation curve. After calving, the daily milk yield increases and achieves its maximum between day 40 and 80 and falls linearly until the end of lactation. The prediction curves in figures 3 and 4 follow this course. However, the estimated conditional expected values achieve their maximum about 2 to 3 weeks later, in comparison to the lactation curves.

Calculation of average values for selected daily intervals

The following listed results are based on the hurdle model with random independent effects (see formulas (1) to (5)). The results within the link scale, that is for $p_{0,ijk}$ within the logit scale and for λ_{ijk} within the log scale are less clear. Thus, a re-calculation is carried out into the response scale. The delta method (Greene 2008) was used in order to calculate the standard error in the response scale. According to the formulas (1) to (5), the distribution assumptions made are valid within all combination levels from CL i and hour j to the given day of lactation t. Consequently the conditional or marginal expected values (see formula (6) and (7)) within each combination level must be calculated. The creation of the average value must then occur in the response scale. This means that within each possible combination level of the testing factors CL and day hour the linear predictors are transformed back. After this transformation the accumulation then occurs in the original scale. Through the creation of the average value, for instance over all 24 hours, over the day hours or over the night hours, marginal expected values can be calculated as follows.

$$\bar{\mu}_i(t) = \frac{1}{b} \cdot \sum_{j=1}^{b} \mu_{ij}(t) \qquad \bar{p}_{0i}(t) = \frac{1}{b} \cdot \sum_{j=1}^{b} p_{0,ij}(t) \tag{12}$$

The change to marginal expected values guarantees that all made statements are valid for randomly selected cows of both CL. In Table 5 the estimated marginal expected values for the night and day hours are summarised for an average day in milk of $\bar{t}$=112.

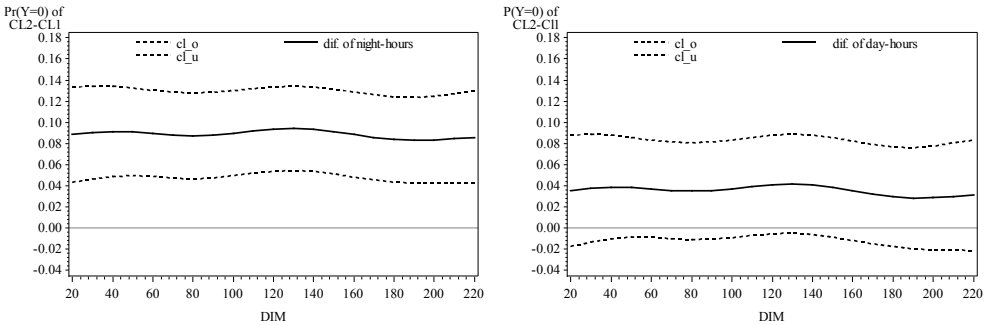
Table 5
Marginal probabilities ($\hat{p}_0$) for zero event (no visit per h) and marginal expected values ($\hat{\mu}$) for the number of visits, dependent on the classes of lactation (CL) and estimated for the day and night hours for given average day of lactation

	$\hat{p}_0$±se		$\hat{\mu}$±se	
class	Night-time	Day-time	Night-time	Day-time
CL1	0.552±0.0146	0.467±0.0169	2.365±0.192	2.646±0.213
CL2	0.641±0.0109	0.503±0.0123	1.602±0.093	2.029±0.110
Dif. CL2-CL1	0.089±0.0182	0.0363±0.0209	−0.762±0.213	−0.617±0.239
p-value	<0.0001	0.0971	0.0018	0.0174

In comparison to the cows of CL 2, the cows of CL 1 have a lower probability that no visit took place. The difference of 0.552 to 0.641 in the night hours is highly significant. Within the day hours the estimated difference of 0.0363 is not statistically significant at the 5 % level, also for infinite degrees of freedom (*p*-value=0.0824) and for 21 degrees of freedom

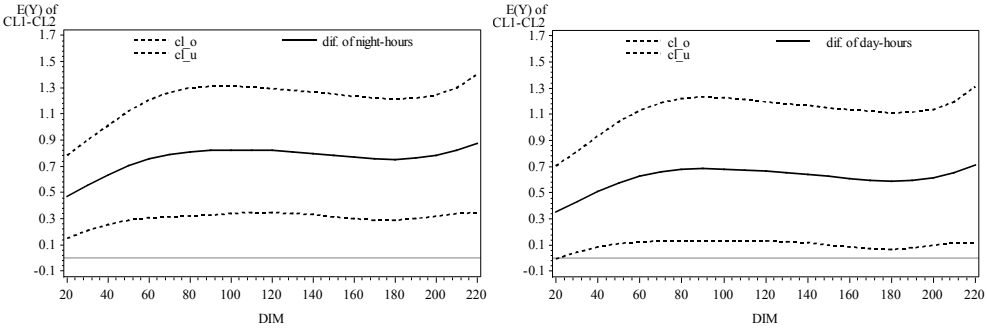
(*p*-value=0.0971). For the degrees of freedom (DF) the number of subjects minus the number of random factors in the linear predictor was used.

The results in Table 5 can be calculated for every day of lactation in the observation period. The figures 5 and 6 show the estimated difference $d(\bar{p}_0)=\bar{p}_{02}(t)-\bar{p}_{01}(t)$ with averaging over the day and night hours. Apart from the estimated difference, figures 5 and 6 contain the corresponding confidence intervals to P=0.95 with assumption of t-distribution for the estimated differences with 21 DF. Figures 5 and 6 confirm and enhance the results of Table 5. The differences between the CL are highly significant in the night hours for all DIM between 20 and 220. In contrast, differences in the day hours for the probability that no visit takes place per hour, cannot be proven to be significant to α=0.05.



Figures 5 und 6
Estimated difference between classes (CL) 2 and 1 for the marginal probability Pr(Y=0) of the zero event (no visit per h), dependent on the days in milk (DIM) and determined over the day and night hours.

In figures 7 and 8 the differences $d(\bar{\mu})=\bar{\mu}_1(t)-\bar{\mu}_2(t)$ formed for the day and night hours, dependent on day of lactation t with corresponding confidence bands (P=0.95, DF=21) are graphically displayed. The differences between the CL for the average number of visits in the night hours are significant for all DIM between 20 and 220. With the exception of a few days at the beginning of the lactation, this statement is also true for the differences between the CL for the average number of visits in the day hours.



Figures 7 and 8
Estimated difference between classes of lactation (CL) 1 and 2 for the marginal expected value E(Y) of the trait number of visits, dependent on the days in milk (DIM) and determined over the day and night hours.

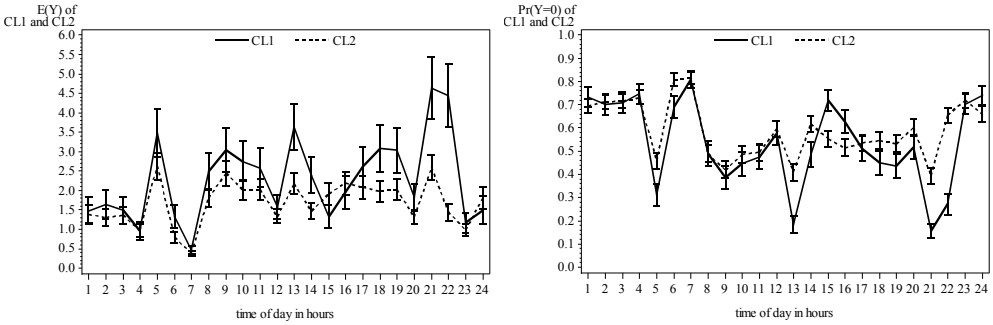
For primiparous cows, the average number of visits per night-hour increases from 0.47 at the beginning of the lactation to 0.87 at the end of the lactation compared to multiparous cows (see figures 7 and 8). For the average number of visits per day-hour, the increase lies between 0.35 at the beginning, and 0.71 at the end of the lactation. In order to cover the energy and nutritional requirement, primiparous cows probably have to adjust their visiting activities throughout the lactation.

Calculation of averages for selected hours

In the following, the marginal probabilities and expected values are calculated using formula (7) for selected hours with an averaging through all DIM between 20 and 220. In all figures the confidence intervals (P=0.95, DF=21) are given for the estimated parameters. The connection of the estimated values per hour for each CL gives trend curves as an extra component of the figures. Figure 9 shows the estimated expected value for the number of visits for the primiparous and multiparous cows for each hour.

A clear decrease in the visits is seen at the four service times (hour 4, 12, 7 and 20) for both primiparous and multiparous cows. The trend curves illustrate that, apart from the hours 15 and 16, the number of visits clearly increased for the primiparous in comparison to the multiparous cows. At hours 5, 13 and 21, which directly follow the milking times, a much higher number of visits was shown for the primiparous cows than for multiparous cows. With the exception of the service time at hour 7, the multiparous cows show a much reduced variation in the number of visits over the day and night hours in comparison to primiparous cows. The visiting activity of the primiparous cows is not only raised, but also shows a much larger upwards displacement. The reasons are the limited access to the feeders and probably the low rank of the primiparous cows compared to the multiparous cows. The high visiting activity in hours 21 and 22 is very noticeable. After the third milking at hour 20, the average number of visits increases to values over 4. Within the hours 23 and 4, the primiparous and multiparous cows achieve a similar level with an average of 1 or 2 visits per hour.

The probability that no visit will take place is shown in figure 10. As expected, a similar activity pattern arises linked to figure 9. After the milking and cleaning times in hours 4, 7, 12 and 20 there is an extreme decrease of the probability for the absence of a visit in the



Figures 9+10
Estimated marginal expected values and probabilities for the primiparous (CL1) and multiparous cows (CL2), dependent on the time of day and determined over days in milk between 20 to 220.

following hours: 5, 8, 13 and 21. Thus, the probability decreases for example from 0.75 during hour 4 to 0.31 during hour 5. Figures 9 to 10 lead to the following conclusions. The deciding independent variable and thus the »time giver« or »synchronizer« for the visiting activity are the four hours with limited access to the feeders. The visiting time of the primiparous and multiparous cows show extreme differences, not only in the absolute level, but also in the variation. For primiparous cows, the largest increase of visiting frequency is shown after the three milking times in hours 5, 13 and 21. The confidence intervals in figure 9 show that the increases are found to be significant in comparison to the multiparous cows.

Discussion

In order to evaluate the traits with excess zeros on presentation of the repeated samples per object, there is a competition between ZI and hurdle models (augmented by random effects) based on Poisson or negative binomial distribution. The ZI models explain the excess zero through the division of individuals into two groups. The individuals of the first group can (for example, due to their genetic disposition) only take the zero event. For individuals of the second group, the zero event occurs according to a random variable, for which a distribution for a count trait suffices. With respect to the example investigated, cows must exist that did not visit the feeder at a certain hour over the trial period of 141 days. The advantage for the interpretation of the excess zeros becomes a negative effect for the estimation of the model parameters. In the hurdle model, the estimation of the parameters is simplified considerably, if there is a prerequisite that the random effects per object are uncorrelated for the zero event and the event larger than zero. In this case, the estimation of the model parameters takes place in two steps. In the first step, a binary trait is analysed and in the second step the evaluation takes place, under the assumption of a distribution truncted-at-zero. In contrast, in ZI models all parameters have to be estimated simultaneously, even with independent random effects per animal. In particular, with the explanatory factors with many levels and a large number of covariates, the parameter estimation with the maximum likelihood method in the ZI model with random effects leads to unacceptable calculation times and increasing problems with convergence. The numerical problems also arise from the calculation of integrals with the help of the Gauss-Hermite quadrature formulas, which are necessary for the deployment of the log-likelihood function. Thus, the hurdle model with random effects is used for the evaluation of the number of visits per hour. The dataset presented contained 64 728 observations for the trait, with which (using the maximum likelihood method) 112 fixed parameters and, in the case of the negative binomial distribution, 3 parameters of the variability were estimated.

For improved interpretation of the results, averaging over the levels of the explanatory factors in the response scale was carried out. Through the conversation to marginal expected values it is guaranteed that all conclusions for randomly selected cows are valid for the whole herd and thus are independent of the animals in the trial.

The comparison of our results with values from the literature is restricted. In other studies the trait number of feeder visits (NFV) was analyzed per day. In numerous papers (Tolkamp & Kyriazakis 1999, Azizi *et al.* 2010) a grouping of the visits into meals is carried out to characterize the feeding behaviour. Work up to now has not analyzed the feeding behavior

of cows dependent on the day of lactation. Friggens *et al.* (1998) divided the lactation into four periods with the use of average performance per period in order to reflect the lactation dynamic. Azizi *et al.* (2010) are carrying out a division between 7th and 105th day of lactation into three periods of equal length. Kaufmann *et al.* (2007) determined an increase of NFV from 28 visits per day in the 2nd week of lactation to 35 visits per day in the 15th week of lactation. In our study the trait NFV per hour is evaluated by using a Hurdle model based on negative binomial distribution. In order to reflect the lactation dynamic, polynomials of fourth degree and an approach with sine and cosine functions are used.

The evaluation of the number of visits per hour showed that the deciding »time giver« for the visiting activity of the cows is given by the three milking times and the cleaning times of the feeders. In comparison to the multiparous cows, the visiting frequency of the primiparous cows was strongly increased. The highly significant differences found can be explained as follows: the primiparous cows are forced to achieve the necessary feed intake through higher visiting frequency than the multiparous cows. Conditioned by the 2 to 1 ratio between animal and feeder in the trial, a competitive situation probably arises for access to the feeders. The multiparous cows, which are higher in rank, displace the lower-ranking primiparous cows. These are then forced to change feeders or to visit the feeder again. A separate housing of primiparous and multiparous cows, at least during the first 100 DIM, could lessen the competition between the two CL. In this trial there were no results, for instance, for the number of aggressive displacements or for the number of direct confrontations between primiparous and multiparous cows. Thus, the explanations and conclusions given must be confirmed by further investigations.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds.) Proc. 2nd Intern Symp on Information Theory. Akademiai Kiado, Budapest, Hungary, 267-281

Azizi O, Hasselmann L, Kaufmann O (2010) Variations in feeding behaviour of high-yielding dairy cows in relation to parity during early to peak lactation. Arch Tierz 53, 130-140

Bulang M, Kluth H, Engelhard T, Spilke J, Rodehutscord M (2006) Studies on the use of lucerne silage as a forage source for high-yielding dairy cows. J Anim Physiol Anim Nutr 90, 89-102 [in German]

Friggens NC, Nielsen BL, Kyriazakis I, Tolkamp BJ, Emmans GC (1998) Effects of feed composition and stage of lactation on the short-term feeding behavior of dairy cows. J. Dairy Sci 81, 3268-3277

Greene WH (2008) Econometric analysis. 6th ed. Pearson-Prentice Hall/Upper Saddle River, NJ, USA

Hall DB (2000) Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics 56, 1030-1039

Kaufmann O, Azizi O, Hasselmann L (2007) Feeding behaviour of high yielding dairy cows during early lactation. Züchtungsk 79, 219-230 [in German]

Lambert D (1992) Zero-inflated poisson regression with application to defects in manufacturing. Technometrics 34, 1-14

Lee AH, Wang K, Scott JA., Yau KKW, McLachlan GJ (2006) Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Stat Methods Med Res 15, 47-61

Liu WS, Cela J (2008) Count data models in SAS. SAS Global Forum 317, 1-12

McGilchrist CA (1994) Estimation in generalised mixed models. J R Stat Soc Series B Stat Methodol 56, 61-69

Min Y, Agresti A (2005) Random effect models for repeated measures of zero-inflated count data. Stat Modelling 5, 1-19

Mullahy J (1986) Specification and testing of some modified count data models. J Econ 33, 341-365

Ritz J, Spiegelman D (2004) Equivalence of conditional and marginal regression models for clustered and longitudinal data. Stat Methods Med Res 13, 309-323

Rodrigues-Motta M, Gianola D, Heringstad B, Rosa GJM, Chang YM (2007) A zero-inflated poisson model for genetic analysis of the number of mastitis cases in Norwegian Red cows. J Dairy Sci 90, 5306-5315

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6, 461-464

Thamm K, Bulang M, Mielenz N, Spilke J (2011) Analysis of milk yield, dry matter intake, number of feeder visits and feeding time in dairy cows fed grass-, alfalfa- or corn silage based mixed diets by using of random regression models. J Anim Nutr (presented)

Tolkamp BJ, Kyriazakis I (1999) To split behaviour into bouts, log-transform the intervals. Anim Behav 57, 807-817

Xiang L, Lee AH, Yau KKW, McLachlan GJ (2007) A score test for overdispersion in zero-inflated poisson mixed regression model. Stat Med 26, 1608-1622

Yau KKW, Lee AH (2001) Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Stat Med 20, 2907-2920

Yau KKW, Wang K, Lee AH (2003) Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biom J 5, 437-452

Received 14 April 2011, accepted 23 June 2011.

Corresponding author:

Norbert Mielenz
email: norbert.mielenz@landw.uni-halle.de

Martin-Luther-University Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, 06099 Halle (Saale), Germany