Open Access

Original study

Analysis of correlated count data using generalised linear mixed models exemplified by field data on aggressive behaviour of boars

Norbert Mielenz[1], Joachim Spilke[1] and Eberhard von Borell[2]

[1]Working group Biometry and Agricultural Informatics, [2]Department of Animal Husbandry and Livestock Ecology, Institute for Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Germany

Abstract

Population-averaged and subject-specific models are available to evaluate count data when repeated observations per subject are present. The latter are also known in the literature as generalised linear mixed models (GLMM). In GLMM repeated measures are taken into account explicitly through random animal effects in the linear predictor. In this paper the relevant GLMMs are presented based on conditional Poisson or negative binomial distribution of the response variable for given random animal effects. Equations for the repeatability of count data are derived assuming normal distribution and logarithmic gamma distribution for the random animal effects. Using count data on aggressive behaviour events of pigs (barrows, sows and boars) in mixed-sex housing, we demonstrate the use of the Poisson »log-gamma intercept«, the Poisson »normal intercept« and the »normal intercept« model with negative binomial distribution. Since not all count data can definitely be seen as Poisson or negative-binomially distributed, questions of model selection and model checking are examined. Emanating from the example, we also interpret the least squares means, estimated on the link as well as the response scale. Options provided by the SAS procedure NLMIXED for estimating model parameters and for estimating marginal expected values are presented.

Keywords: correlated count data, negative-binomial and Poisson distribution, aggressive events, boar

Abbreviations: GLMM: generalised linear mixed models; GnRH: gonadotropin-releasing hormone; LSM: least square means; ML: maximum likelihood; MNB: multivariate negative binomial; NAE: number of initiated aggressive events; NB: negative binomial; NBinN: negative binomial normal; OLS: ordinary least squares; PA: population-averaged; PoiLG: Poisson log-gamma; PoiN: Poisson normal; SE: standard errors; VC: variance-covariance

Introduction

Count data arise when the trait outcome is determined through a count process, whereby a count variable denotes the number of discrete occurrences of an event of interest. The case of repeated measurements occurs when the count trait on the same subject is recorded at several time points, usually consecutively. One widely-cited example is the number of epileptic seizures experienced by patients within 14 days, recorded over 8 weeks (Thall & Vail 1990). Count data do not fulfil the requirements of normally distributed data. Thus, for example, the variances of response variables, whose realisations are the observations, are deterministic functions of the expected values. In the case of continuous, normally distributed outcomes, the mean and the variance are entirely separate parameters. For normally distributed data, there is no relationship between the mean and the variance, while for Poisson distributed data, the variance is equal to the mean. The variance of the outcomes is thus assumed to be of the form dictated by the distribution. In the linear model assuming normality and heterogeneity, the residual variances are considered as additional model parameters to be estimated.

Two fundamentally different model approaches are available for the statistical evaluation of repeated observations per subject. In the subject-specific model (Breslow & Clayton 1993, McCulloch & Searle 2008), the assumption is that the observations for the given random effects of the subjects satisfy a Poisson or a negative binomial distribution. With the help of a link function, the relationship between the linear predictor and the conditional expected value is provided on the response scale. Since count traits can only take non-negative integer values and thus the expected value is always positive, the logarithm function is used as a link function and thus the exponential function is used as the inverse link function. Considering qualitative and quantitative explanatory factors in the linear predictor follows, analogous to the theory of the linear model, with the help of effects and regression coefficients, which can be both fixed as well as random. Repeated measures are taken into account via random subject effects, which are explicit components of the linear predictor. A further differentiation of the subject-specific model is given by making distributional assumptions for the random effects in the linear predictor. In the literature, subject-specific models are also termed generalised linear mixed models (GLMMs). On the response scale the GLMMs provide conditional expected values for the count trait, for example for an animal whose effect corresponds to the average of the herd or population. If statements about an animal randomly chosen from the population are desired, then a conversion of the conditional expected values to marginal expected values must be made. This conversion is carried out by considering the conditional expected values to be random variables and by forming the expected value in the distribution of the random effects. In the case of non-linear link functions, the conversion is only possible in special cases (Ritz & Spiegelman 2004).

Apart from the GLMM, one may also use so-called population-averaged (PA) models to evaluate the count data with repeated measures per subject (Liang & Zeger 1986; Hardin & Hilbe 2003). In PA models a direct connection is made between linear predictors and marginal expected values by using a link function. Consequently, the PA models are also denoted as marginal models. To account for correlation due to repeated observations per subject, a so-called working correlation matrix is used. For this correlation matrix a certain structure must be predefined to select the covariance structure of the model. Variance-covariance (VC) matrices for the observation vectors per subject can be formed from the correlation matrix, but only the diagonal elements of this VC matrix are derived from the distribution assumption for the observation. In this paper only subject-specific models, that are models with random effects in the linear predictor, are used to evaluate the count data. On particular it is shown how, for count data, the covariance between repeated observations per subject can be calculated on the response scale. Formulae for repeatability are derived, depending on the distribution assumption, not only with respect to the observations but also to the random effects.

To estimate the model parameters in GLMM, the maximum likelihood (ML) methods based on numerical integration according to Gaussian-Hermite quadrature, the integral method based on the Laplace approximation and the pseudo-likelihood methods based, for instance, on the linearization of the data exist (see McCulloch & Searle 2008). In this paper the ML method is used to estimate parameters including the Gaussian-Hermite quadrature. If the numerical integration technique for setting up the log-likelihood function works well, this method will be preferable to the pseudo-likelihood method. In a study based on stochastic simulation, this hypothesis was confirmed (Thamm 2012).

In chapter „Models for correlated count data" we present a brief review of models for count data with repeated measures. Chapter "Calculation of repeatability" includes formulas of correlations between repeated measures of counts. In chapter "Estimation of the model parameters" we discuss aspects of parameter estimation in GLMM by using the procedures GLIMMIX and NLMIXED of SAS. An example illustrates the application of the models with field data on the aggressive behaviour of fattening pigs. The field data are used to illustrate the options of the residual analysis in GLMM. Not every count trait can be assumed to have a Poisson distribution. Thus, questions on model selection and checking are extensively illustrated for the analysis of correlated count data.

Models for correlated count data

The Poisson »normal intercept« model

Let y_{ij} be the observation of subject i ($i=1,\ldots,a$) at time j ($j=1,\ldots,n_i$). Observations y_{ij} are regarded as realisations of random variables Y_{ij} . For the given random effect u_i from subject i, the Y_{ij} are assumed to be Poisson-distributed random variables. Thus, it follows that:

$$Y_{ij}|u_i \sim Poi(\lambda_{ij}) \ \ with \ \ log(\lambda_{ij}) = x_{ij}'\beta + u_i \tag{1}$$

In (1), β is the vector of the fixed model parameters and x_{ij} is the design-vector for the j-th observation of subject i.

From (1) the following conditional expected value is obtained:

$$\lambda_{ij} = E(Y_{ij}| u_i) = z_i \cdot exp(x_{ij}'\beta) \quad with \quad z_i = e^{u_i} \tag{2}$$

If we assume $u_i \sim N(0,\sigma^2)$ then model (2) will be named as Poisson »normal intercept« model (PoiN). In PoiN, the marginal expectation, the variance and the covariance can be expressed as (Molenberghs *et al.* 2007):

$$\mu_{ij} = E(Y_{ij}) = E_u\{E(Y_{ij}|u_i)\} = exp(x_{ij}'\beta + \tfrac{1}{2}\cdot\sigma^2)$$

$$Var(Y_{ij}) = \mu_{ij} + (e^{\sigma^2}-1)\cdot\mu_{ij}^2 \tag{3}$$

$$cov(Y_{ij},Y_{ij*}) = (e^{\sigma^2}-1)\cdot\mu_{ij}\cdot\mu_{ij*} \quad with \quad j\neq j^*$$

Due to the fact that $(e^{\sigma^2}-1) >0$, the random variables Y_{ij} show an over-dispersion on the original scale. Only for the conditional random variables $Y_{ij}|u_i$ do the expected value and variance agree.

The Poisson »log-gamma intercept« model

If the following approach is used in (2)

$$z_i \sim Gamma(\delta, 1/\delta) = \frac{\delta^\delta}{\Gamma(\delta)}\cdot z_i^{\delta-1}\cdot e^{-\delta z_i} \tag{4}$$

$$with \quad E(z_i) = 1 \quad and \quad Var(z_i) = 1/\delta = \Phi$$

or if we assume that u_i in (1) satisfies a log-gamma distribution, then by (2) and (4) the Poisson »log-gamma« intercept model (PoiLG) is given. According to Solis-Trapala & Farewell (2005) it follows that:

$$\mu_{ij} = E(Y_{ij}) = exp(x_{ij}'\beta)$$

$$Var(Y_{ij}) = \mu_{ij} + \Phi\cdot\mu_{ij}^2 \tag{5}$$

$$cov(Y_{ij},Y_{ij*}) = \Phi\cdot\mu_{ij}\cdot\mu_{ij*} \quad with \quad j\neq j^*$$

If we consider $Y_i = (Y_{i1}, ..., Y_{in_i})'$ and $\mu_{ij} = (\mu_{i1}, ..., \mu_{in_i})'$ then it can be shown that Y_i satisfies a multivariate negative binomial (MNB) distribution whose density function can be given as follows (see Fabio *et al.* 2012):

$$Y_i \sim MNB(\mu_i,\Phi) = f(y_i|\mu_i,\Phi) = \frac{\left(\prod_{j=1}^{n_i}(\mu_{ij})^{y_{ij}}\right)\Gamma(\Phi^{-1}+y_{i\cdot})\,(\Phi^{-1}+\mu_{i\cdot})^{-(\Phi^{-1}+y_{i\cdot})}}{\left(\prod_{j=1}^{n_i}y_{ij}!\right)\Gamma(\Phi^{-1})\,\Phi^{\Phi^{-1}}} \tag{6}$$

In (6) the following is true: $\Phi>0$; $y_{i\cdot} = \sum_{j=1}^{n_i}y_{ij}$ and $\mu_{i\cdot} = \sum_{j=1}^{n_i}\mu_{ij}$

The log-likelihood function of model (5) and (6) can be found in the appendix.

The negative binomial »normal intercept« model (NBinN)

Apart from the Poisson distribution, count data can be modelled using the negative binomial (NB) distribution. The NB distribution depends on an extra parameter compared to the Poisson distribution, which allows over-dispersion to be taken into account. To distinguish it from the MNB distribution, this parameter will be denoted by a (s. equation 7). The density function is given from (6) by setting $n_i=1$ and $\Phi=a$. The repeated measures per subject are modelled under the assumption of normally-distributed animal effects as follows:

$$Y_{ij} | u_i \sim NB(\lambda_{ij}, a) \quad with \quad log(\lambda_{ij}) = x'_{ij}\beta + u_i \quad and \quad u_i \sim N(0, \sigma^2) \tag{7}$$

In (7) the marginal variances and covariances can be calculated as (see Carrasco 2010; Molenberghs *et al.* 2007):

$$\mu_{ij} = exp(x'_{ij}\beta + \tfrac{1}{2}\cdot\sigma^2)$$

$$Var(Y_{ij}) = \mu_{ij} + [e^{\sigma^2}(1+a) - 1]\cdot\mu_{ij}^2 \tag{8}$$

$$cov(Y_{ij}, Y_{ij*}) = (e^{\sigma^2}-1)\cdot\mu_{ij}\cdot\mu_{ij*} \quad with \quad j\neq j* $$

Calculation of repeatability

The correlations between two measures on the same animal can be derived with the help of the marginal variances and covariances given by (3), (5) and (8). For the PoiN, PoiLG and NBinN models, the following expression is obtained:

$$corr(Y_{ij}, Y_{ij*}) = \frac{\lambda_1\cdot\mu_{ij}\cdot\mu_{ij*}}{\sqrt{\mu_{ij}+\lambda_2\cdot\mu_{ij}^2}\cdot\sqrt{\mu_{ij*}+\lambda_2\cdot\mu_{ij*}^2}} \tag{9}$$

In (9) λ_1, λ_2 and μ_{ij} are given by:

$$\mu_{ij} = \begin{cases} exp(x'_{ij}\beta) & for & PoiLG \\ exp(x'_{ij}\beta + 0.5\sigma^2) & for & PoiN\ and\ NBinN \end{cases} \tag{10}$$

$$\lambda_1 = \begin{cases} \Phi & for & PoiLG \\ (e^{\sigma^2}-1) & for & PoiN \\ (e^{\sigma^2}-1) & for & NBinN \end{cases} \quad and \quad \lambda_2 = \begin{cases} \Phi & for & PoiLG \\ (e^{\sigma^2}-1) & for & PoiN \\ [e^{\sigma^2}(1+a)-1] & for & NBinN \end{cases} \tag{11}$$

If, for instance, animals are randomly distributed to treatments with different mean values, the correlations will be dependent on the considered treatment. If in (9), for instance, an average estimated value is used for μ_{ij}, count traits can be compared to one another in their repeatability. If μ_{ij} is replaced by μ, the derived equation (from Foulley & Im 1993) will follow for heritability in the broad sense of a Poisson distributed trait, as seen for the special case of (9). The prerequisite for this is that variability between animals is only genetically caused. If random permanent environmental effects occur, equation (9) will only provide an upper limit

for the heritability. A further possibility to eliminate the dependence of the fixed model effects of qualitative explanatory factors in equation (9) can be found in Carrasco & Joyer (2005). These authors derive the marginal expected value given through equation (10) under the assumption that all model effects are random and satisfy a normal distribution. The variances of qualitative explanatory factors are estimated using sums of the squared fixed effects. This approach allows the definition of the intra-class correlation coefficients in the linear model to be transferred to the more comprehensive classes of the GLMM (Carrasco 2010).

Estimation of the model parameters

Estimating the parameters in the Poisson normal (PoiN), in the Poisson log-gamma (PoiLG) and in the negative binomial normal (NBinN) model occurs through the use of the ML method, i.e. through setting up and maximising the log-likelihood function. For this, numerical integration based on the Gaussian-Hermite quadrature formula is necessary. The parameter estimation using the ML method for the models PoiN and NBinN can be carried out in SAS (SAS Institute Inc., Cary, NC, USA) with the procedures GLIMMIX and NLMIXED. Characteristics of GLIMMIX are: three methods of parameter estimation (Gaussian quadrature, Laplace integral approximation and a pseudo-likelihood method based on linearization of the data), a class statement for qualitative covariates, high complexity of the linear predictor and the covariance structure, but only testing of linear contrasts and of differences of least square means (LSM) on the link scale. Advantages and restrictions of NLMIXED are: parameter estimation with Gaussian quadrature only, user-defined likelihood functions are supported (PoiLG and MNB can be implemented), calculation of standard errors for nonlinear parameter functions by using the delta method (testing differences of marginal means can be implemented easily, see appendix). Problems with NLMIXED are: no class statement is available, the complexity of covariance structures is restricted. If only testing of LSM (from model PoiN or NBinN) on the link scale is desired, GLIMMIX should be the procedure of choice.

For the random effects in the linear predictor, only a normal distribution can be supposed in NLMIXED by default. If model PoiLG is to fit with NLMIXED, the following relationship can be used (Nelson *et al.* 2006): Let $a_i \sim N(0,1)$ then $p_i = \Phi(a_i)$ has a uniform distribution, whereby $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Let $F^{-1}(\cdot)$ denote the inverse Gamma distribution. Then, $g_i = F^{-1}(p_i)$ is gamma distributed. The commands in NLMIXED for gamma random effects are given in the appendix. Using the adaptive Gaussian quadrature can lead to convergence problems in the case of the PoiLG model. Therefore, Nelson *et al.* (2006) recommend the non-adaptive Gaussian quadrature (NOAD option) with an increase in the number of quadrature points. Better and more stable convergence is achieved by setting up the log-likelihood of the MNB distribution (s. appendix) and maximising this. This approach can be used within NLMIXED if, for each subject (in this example an animal), a dataset with all of the repeated observations per animal is generated. The dataset must also include all qualitative and quantitative factors that occur in the linear predictor. Since NLMIXED does not have a class-statement, numbering all qualitative explanatory variables is recommended. Full numbers, beginning with one, were assigned to the successive levels of the explanatory factors. This enables an assignment of field elements to the levels of explanatory factors in NLMIXED. Implementing the multivariate negative binomial distribution in NLMIXED can be found in the appendix for an example from the aggressive behaviour of boars.

An illustrative example

Field data on the aggressive behaviour of fattening pigs

The conventional castration of male piglets without anaesthesia has been recently questioned due to animal welfare concerns. As an alternative to surgical castration, active immunisation against the endogenous gonadotropin-releasing hormone (GnRH) has been established to avoid boar taint in male pigs due to sexual maturation and to reduce levels of associated aggression and sexual behaviour (Schmidt *et al.* 2010). Following two vaccinations with an interval of at least 4 weeks, the function of the testicles is temporarily suppressed and thus the production of the unwanted sex-specific boar taint is avoided. In the following, normally castrated boars will be denoted as barrows and those in which the sexual-function is suppressed by immunisation will continue to be referred to as boars. We study the effect that different sex combinations in fattening groups have on the frequency of agonistic encounters between boars. The following 4 combinations of single-sex and mixed-sex housing with 30 animals in each group are described:

Treatment 1: 30 boars
Treatment 2: 15 boars, 15 sows
Treatment 3: 15 boars, 15 barrows
Treatment 4: 10 boars, 10 sows, 10 barrows.

Treatments 1 and 2 were examined in four pens each, treatment 3 was examined in three pens and treatment 4 took place in one pen. All 12 pens were located in one room. The distribution of the treatments within the pens took place randomly. The vaccination of the uncastrated male piglets took place on the 1st and 69th day after the piglets had been stabled in the fattening room. The suppression of the testicle function begins a few days after the second immunisation with GnRH. Apart from recording the quality and classical performance traits such as body weight and daily weight gain (not reported in this paper), traits were also recorded that described the agonistic behaviour through video observations. Six recorded days (time points) were evaluated during an interval of 14 days. To evaluate agonistic behaviour, the number of initiated aggressive events (NAE) per boar and the number of mounting attempts (not considered for analysis in this paper) per boar in two selected daylight hours were determined. The NAE per boar within two hours of the day provides a count trait with a maximum of six repeated measures per subject. The arithmetic mean per treatment and time are listed in Table 1. A usable video recording was not available for all boars in the 11th and 15th h. Recordings for treatment 4 at time point 4 were completely missing.

Table 1
Average values for the number of initiated aggressive events per boar dependent on treatment and time

Treatment	Time					
	1	2	3	4	5	6
1	0.783	0.342	0.267	0.218	0.235	0.050
2	0.948	0.893	0.464	0.263	0.351	0.000
3	0.633	0.356	0.600	0.533	0.295	0.140
4	2.200	0.400	0.800	×	0.200	0.100

Statistical models

For the analysis of NAE a GLMM was used. Let y_{ijkl} be an observation of treatment i ($i=1,\ldots,4$) at time point k ($k=1,\ldots,6$) on boar j from pen l ($l=1,\ldots,n_i$). The four treatments were randomly allocated to 12 pens in one room. The number of pens varied from 1 to 4. The pens were therefore considered as randomization units. Thus, in a first step of model selection, pens were included as random effects in the model. In a GLMM framework the following linear predicator was used

$$\eta_{ijkl} = \mu + a_{ik} + b_{il} + u_{ijl} \quad with \quad (b_{il}, u_{ijl})' \sim N(0,D) \quad and \quad D=Diag(\sigma_b^2, \sigma_u^2) \tag{12}$$

where μ is the general mean (or intercept), a_{ik} is the fixed effect of k-th time point within i-th treatment, b_{il} is the random effect of l-th pen within i-th treatment and u_{ijl} is the random effect of j-th boar from l-th pen within i-th treatment. Under the assumption of Poisson as well as negative binomial distribution, the use of pseudo-likelihood methods (GLIMMIX with option method=rspl) lead to convergence problems during the parameter estimation. When using the Laplace method in GLIMMIX, the pen variance is estimated by zero. Therefore, in the next step of model selection the pen effects of (12) were taken as fixed. In this case, the statistical analysis using (12) showed that there are no significant differences between the fixed pen effects (*p*-value of F-test equal to 0.39 for Poisson and to 0.26 for NB distribution). The pens are nested within treatments. Therefore, the time-depending treatment effects can be expressed as a linear combination of time-by-pen effects. Thus, the LSM for comparing the treatments within time cannot be analysed. One possible solution of modelling time-specific pen effects is to use random regression models. The time of recording has to be considered as continuous covariate t. For example, if a linear time trend is assumed, then linear predictor can have the following simple form:

$$\eta_{ijl}(t) = (\beta_{0i} + b_{0il}) + (\beta_{1i} + b_{1il}) \cdot t + u_{ijl} \tag{13}$$

Here, β_{0i} and β_{1i} are treatment-specific fixed regression coefficients and b_{0il} and b_{1il} are normally distributed pen effects of intercept and slope. Perhaps, the simplest assumption in (13) is

$$(b_{0il}, b_{1il}, u_{ijl})' \sim N(0, D) \quad with \quad D=diag(\sigma_0^2, \sigma_1^2, \sigma_u^2) \tag{14}$$

When the Laplace method of GLIMMIX is used, the pen variance in (14) is estimated to be zero. The consideration of linear regression functions with treatment-specific slopes was discarded to describe the increase in average values between time points 2 and 3 (see Table 1). The model selection steps can be summarized as follows: In principle, a model with random pen or time-by-pen effects should be used. Because of the non-significance of the fixed main pen effects, no such effects were fitted. We also tried a simple random regression on time by pen. This model also yielded a zero variance for pens. As a consequence, pen effects were dropped from the analysis. The pen effects in (12) can be interpreted as permanent environmental effects. For NAE no such environmental effects could be detected. The analysis of NAE can be carried out using PoiGL, PoiN and NBinN from section Models for correlated count data. If negative binomial distribution is assumed, GLMM will be given as follows:

$$Y_{ijk}|u_{ij} \sim NB(\lambda_{ijk}, a) \quad with \quad \lambda_{ijk} = E(Y_{ijk}|u_{ij})$$
$$and \quad log(\lambda_{ijk}) = \mu + a_{ik} + u_{ij} \quad with \quad u_{ij} \sim N(0,\sigma_u^2) \tag{15}$$

In model (15) the marginal expected values (or marginal means) were obtained as:

$$\mu_{ik} = E_u \{E(Y_{ijk} \mid u_{ij})\} = exp\,(\mu + a_{ik} + 0.5 \cdot \sigma_u^2) \quad and \quad \mu_i = \frac{1}{6} \cdot \sum_{k=1}^{6} \mu_{ik} \tag{16}$$

The number of aggressive events is given by (16), which we can expect (at time k in a pen of treatment i) of a randomly chosen boar from the population. The correlations between two observations at time k and k^* on the same boar from treatment i can be calculated according to equation (9) as follows:

$$r_{k,k^*} = \frac{\lambda_1 \cdot \mu_{ik} \cdot \mu_{ik^*}}{\sqrt{\mu_{ik} + \lambda_2 \cdot \mu_{ik}^2} \cdot \sqrt{\mu_{ik^*} + \lambda_2 \cdot \mu_{ik^*}^2}} \tag{17}$$

Checking the distribution assumptions

Not every count trait can be described by a Poisson or negative binomial distribution. This is particularly true for the case that zero outcomes compared to the other outcomes occur with a relatively high frequency in the random sample. Thus, in the first step of the model checking, a comparison of the observed relative frequencies is carried out with the probabilities predicted from the model. In addition, the distribution parameter λ_{ijk} of model (15) was estimated for each record of the sample. Following this, estimating the probability $P(Y_{ijk}=m)$ with $m=0,1,\ldots,M$ was carried out for each record using the density function of the Poisson or NB distribution. The largest number of aggressive events initiated by one boar within two hours was eight. Thus, M was set equal to 8. Predicted probabilities of count outcomes between 0 and 8 were calculated by averaging across all records. In Table 2 the predicted probabilities and the relative frequencies observed in the random sample are summarized (in percentages). The described approach requires strictly independent observations. Only the sum of independent Poisson distributed random variables is again Poisson distributed. The assumption of independence can be satisfied if the described analysis is carried out separately for each time point. According to Table 2, around 77.1 % of the boars did not initiate an aggressive event during the two observed hours.

Table 2
Observed absolute and relative frequencies in %, as well as predicted probabilities in % using PoiN and NBinN for the number of initiated aggressive events per boar

| | Number of initiated aggressive events per boar | | | | | | | | |
	0	1	2	3	4	5	6	7	8
Absolute	914	170	67	23	3	2	4	1	1
Relative	77.1	14.3	5.65	1.94	0.25	0.17	0.34	0.08	0.08
PoiN	74.2	19.6	4.50	1.11	0.32	0.11	0.05	0.02	0.01
NBinN	77.2	15.8	4.42	1.49	0.58	0.25	0.12	0.06	0.03

One or two aggressive events were carried out by about 14.3 % or 5.7 % of the boars. More than two aggressive events were only carried out by 2.9 % of the boars. The observed relative frequencies were particularly well predicted by NBinN, i.e. by model (15). The results listed in Table 2 show that the use of Hurdle (Mullahy 1986, Mielenz *et al.* 2011) or

zero-inflated models (Lambert 1992) is not necessarily required to describe the many zeros in the existing data.

Marginal means and repeatability

The LSM of μ_i (see equation 16), also known as marginal means of the i-th treatment depending on the three investigated models, are contained in Table 3. NLMIXED is used to calculate the LSM and its standard error. In this procedure calculating averages can be allocated to the response scale. NLMIXED uses the approximate delta method to calculate the standard errors (SE). The greatest SE of all marginal means were found with NBinN.

Table 3
LSM (±standard error) of the treatments by averaging on the response scale compared to the observed means ($\bar{y}_{obs}$)

	Model			
Treatment	PoiLG	PoiN	NBinN	$\bar{y}_{obs}$
1	0.319±0.032	0.323±0.033	0.319±0.035	0.316
2	0.494±0.065	0.493±0.067	0.491±0.070	0.487
3	0.423±0.064	0.417±0.064	0.419±0.067	0.426
4	0.740±0.196	0.720±0.195	0.737±0.225	0.740

The estimated differences between the treatments and the results of the Wald t-tests are given in Table 4. Observations were not available for treatment 4 at time point 4. Thus, this treatment was dropped from comparison of means across the six time points. In PoiN and NBinN, the number of degrees of freedom is equal to the number of independent subjects minus one. This selection agrees with the standard within the NLMIXED procedure. The PoiLG and MNB models (see section The Poisson »log-gamma intercept« model) are equivalent. Since MNB has better convergence characteristics, the LSMs of the treatments were calculated using this model. In MNB the degrees of freedom are equal to the number of animals minus the number of model parameters. According to Table 3, the lowest NAE can be expected of boars from treatment 1 (housing type 1). From Table 4 it follows that the difference between treatment 1 and treatment 2 is significantly different from zero in all three models with an error probability of 5 %.

Table 4
Differences of the LSM (±standard error) and *p*-values derived from the Wald *t*-test for comparison of treatments across time for given degrees of freedom (df)

	MNB (df=209)		PoiN (df=232)		NBinN (df=232)	
treatment	difference	*p*-value	difference	*p*-value	difference	*p*-value
1-2	−0.175±0.073	0.018	−0.171±0.074	0.022	−0.172±0.078	0.029
1-3	−0.104±0.071	0.147	−0.094±0.072	0.191	−0.101±0.075	0.183
2-3	0.071±0.091	0.436	0.076±0.092	0.409	0.071±0.097	0.462

For a boar from treatment 2, a significantly higher number of initiated aggressive events can be expected compared to the boars from treatment 1. Predictions can be calculated for

the random boar effects in model (15). Since 233 boars were included in the trial and they are presumed to be unrelated, the distribution of the boar effects is approximately normal. The estimations of correlations between sequential measurements using PoiN and NBinN are found in Table 5. The estimated correlations between time points 1 and 2 vary between the treatments for PoiN from 0.193 to 0.317 and for NBinB from 0.046 to 0.066. In PoiN the correlations calculated by averaging over the treatments have a 95 % confidence interval (presuming a normal distribution) with positive lower limits. The correlations between successive estimated time points with PoiLG and PoiN are in the same range. Therefore, Table 5 includes results for PoiN and NBinN.

Table 5
Correlations estimated with PoiN and NBinN between sequential time points for the number of initiated aggressive events per boar and treatment (trt)

trt.	PoiN					NBinN				
	r_{12}	r_{23}	r_{34}	r_{45}	r_{56}	r_{12}	r_{23}	r_{34}	r_{45}	r_{56}
1	0.205	0.131	0.108	0.102	0.051	0.048	0.035	0.030	0.028	0.015
2	0.317	0.247	0.149	0.130	0.000	0.066	0.054	0.037	0.034	0.000
3	0.193	0.184	0.205	0.153	0.089	0.046	0.045	0.050	0.040	0.025
4	0.289	0.212	×	×	0.064	0.061	0.051	×	×	0.019
$\bar{r}$	0.251	0.193	0.154	0.128	0.051	0.055	0.046	0.039	0.034	0.015
$se(\bar{r})$	0.057	0.049	0.040	0.035	0.019	0.038	0.032	0.027	0.023	0.011

Akaike information criterion and residual analysis

According to Table 5 the correlations between repeated measures per boar with PoiN and NBinN have widely different estimates. The question then becomes how to select a suitable evaluation model. Apart from the results from Table 2, comparing the AIC values (Akaike 1974) of the three models (see Table 6) provides further information. In treatment 4 no observations existed for measuring time 4. Thus, just 23 independent model parameters exist for the combination of treatments and time points. Since one parameter of variability is explained in both PoiN and PoiLG and two parameters are explained in NBinN, 24 or 25 independent model parameters ensue. According to Table 6, PoiLG and PoiN have similar statistics. A much better fit is provided by NBinN.

Table 6
Number of model parameters (p) with −2 multiplied log-likelihood function, criteria of the model fit and estimates for selected model parameters

Model	Statistics				Model parameters		
	p	-2logL	AIC	BIC	$\hat{\Phi}$	$\hat{a}$	$\hat{\sigma}_u^2$
PoiLG	24	1757.0	1805.0	1887.8	0.4337	×	×
PoiN	24	1756.0	1804.0	1886.8	×	×	0.4027
NBinN	25	1708.0	1758.0	1844.3	×	1.0910	0.1519

A further possibility for checking the model is provided through the graphical representation of absolute, scaled and standardised residuals, not only dependent on the linear predictor,

but also on the fitted means (McCullagh & Nelder 1989). For example, Pearson residuals, which represent scaled absolute residuals, are well-known and easy to calculate. The conditional Pearson residuals for model (15) have the following form:

$$r^{(P)}_{ijk} = \frac{(y_{ijk} - \hat{\mu}_{ijk})}{\sqrt{\hat{V}ar(y_{ijk} \mid u_{ij})}} \quad \text{with } \hat{\mu}_{ijk} = exp(\hat{\mu} + \hat{a}_{ik} + \hat{u}_{ij}) \tag{18}$$

In (18), the variance function in the denominator depends on the distribution assumption. In the case of negative binomial distribution, the following is true:

$$\hat{V}ar(y_{ijk} \mid u_{ij}) = \hat{\mu}_{ijk} + \hat{\Phi} \cdot \hat{\mu}^{2}_{ijk} \tag{19}$$

The so-called residual plots are obtained by removing the observation number, the linear predictor, a possible continuous explanatory variable, or the estimated conditional expected values on the abscissa and the residuals on the ordinate.
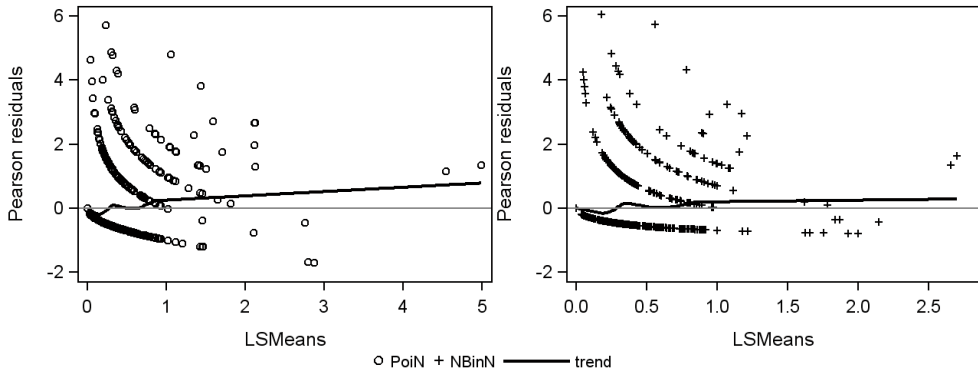

Figure 1
Pearson Residuals for model PoiN and NBinN with smoothed trend plotted against the LSM of μ_{ijk} (on the response scale)

If smoothing the residuals (by using local regression) yields a trend, the choice of variance function, link function or the linear predictor will be wrong (Fahrmeier & Tutz 1994). Figure 1 shows that the dependence of the variance on fitted means for large values in PoiN is small and is adequately described in NBinN. Another possibility of checking the model is to compare the empirical variance function, based on ordinary least squares (OLS) residuals, with the theoretical variance functions of the models PoiN, PoiLG, and NBinN. In the Poisson model, for example, the variance is a function of the means and of the variance of the random boar effects. The variance functions of the three models can be found in equations (3), (5) and (8) of section Models for correlated count data. In the first step of the comparison, the observations are evaluated with a linear model by considering the fixed effects for the combination of treatment, time and pen. The empirical variance function is obtained by smoothing the squared OLS residuals against fitted means using local regression.
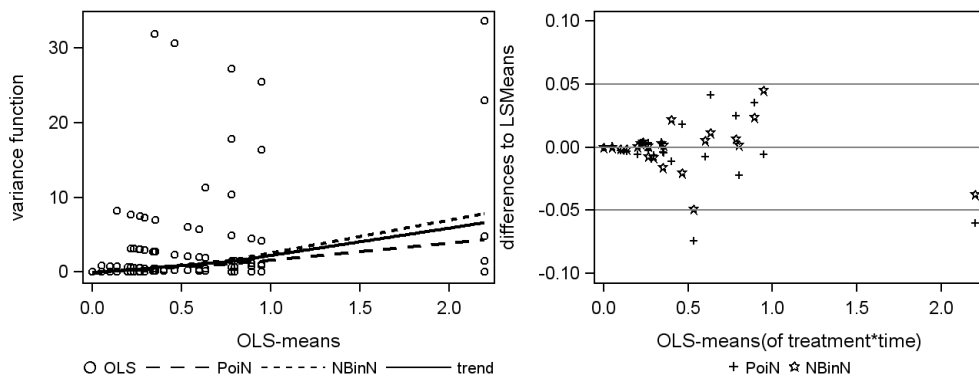
Figure 2
Empirical variance function (trend), squared OLS residuals, variance functions of the models and differences between LSM and OLS-means for the models per treatment and time point

To verify the empirical variance function (solid line), the variance functions generated by PoiN (dotted line) and NBinN (dashed lines), as well as the squared OLS residuals (circle signs) are all plotted against the OLS means (Figure 2). In the example, the number of observations ranged from 10 to 120 per treatment and time point. Thus, the differences between the LSM, which were estimated by equation (16), and the respective OLS-means were compared graphically (Figure 2). For the PoiN or NBinN models, the absolute deviations lay below 0.074 or 0.049, respectively. The empirical variance function and the NBinN variance function showed very good visual agreement for means below 1.2. The correlations listed in Table 5 were also checked using OLS residuals. For this purpose the six repeated observations per boar were interpreted as different traits. The correlations between the OLS residuals of this multiple trait model and the repeatability estimated using NBinN agree quite well.

Interpretation of LSM on the response scale

Following the results of section Akaike information criterion and residual analysis, our analysis continues only with the consideration of model NBinN. Special care is required in interpreting the results for models with random effects. In Table 7, the results of comparing the LSM between the treatments are shown for time points 1, 2 and 3. According to the definition of the LSM from equation (16), the comparisons refer to the differences on the response scale averaged over all boars in the population. A comparison is made for the difference of the number of aggressive events from two treatments, which can be expected by averaging across the distribution of the random boar effects. This interpretation is achieved by transforming the conditional means (taking into account the variability between the boars) into marginal means. Consequently, the standard errors of the two marginal means differences are thus dependent on the common variance-covariance matrix of the estimated fixed model parameters and of the estimated boar variance.

Table 7
Differences of the LSM (±standard error) and *p*-values for comparing the treatments at time point (TP) 1, 2 and 3, estimated using NBinN

Treatment	TP=1		TP=2		TP=3	
	difference	*P*-value	difference	*P*-value	difference	*P*-value
1-2	−0.204±0.264	0.411	−0.572±0.281	0.043	0.346±0.231	0.136
1-3	0.145±0.262	0.582	−0.013±0.127	0.920	0.185±0.229	0.421
1-4	−1.372±0.938	0.145	−0.077±0.270	0.776	−0.012±0.442	0.978
2-3	0.348±0.288	0.227	0.559±0.295	0.059	0.389±0.258	0.134
2-4	−1.169±0.946	0.218	0.495±0.378	0.192	0.191±0.458	0.677
3-4	−1.517±0.946	0.110	−0.064±0.284	0.822	−0.157±0.457	0.731

A further possibility of transforming from the link to the response scale is provided by setting the animal effect to zero. In this case, the standard error of the conditional means can only be approximated by using the variance-covariance matrix of the fixed-effects parameter estimates. In the case of the boar data, we obtain estimations for the number of aggressive events that is expected for a boar that corresponds to the population's average. By setting the boar effects to zero, the comparison delivered a *p*-value of 0.043 or 0.059, respectively, for the differences in treatments 1 and 2, or 2 and 3 at time point 2. This result agrees with the claims of Table 7. The estimation of means has only been carried out on the response scale up to this point. If average values and their comparisons on the link scale are created, significant differences can be found at time point 2, between treatments 1 and 2 as well as between treatment 2 and 3 with a significance level of 5%. Testing of hypotheses on the link scale is the standard in the SAS procedure GLIMMIX. In this paper, GLIMMIX was used to control the convergence behaviour of the optimisation algorithm implemented in NLMIXED by comparing the log-likelihood functions and the estimated model parameters. The difference of marginal means cannot be tested efficiently with GLIMMIX. The advantage of testing on the link scale is that the approximate delta method must not have to be used for calculating the standard error. A difficulty is the interpretation of differences on the response scale if significant differences are found on the link scale. If we generate LSM for treatment i (i=1,…,4) on the link scale, in (15) α_{ik} must be averaged for all time points and u_{ij} must be averaged for all animals from pens of treatment i. The means of treatment i are expressed as $\eta_i = \mu + \bar{\alpha}_{i.} + \bar{u}_{i.}$ on the link scale. In the equation, a dot in place of a subscript indicates that all values of that subscript have been summed over. If a bar is placed over a dot notation expression, it means division by the number of levels across which the sum was computed. To simplify the notation we limit the comparison of treatments 1 and 2 in the following. In order to be able to test the hypothesis $\Delta\eta = (\eta_1 - \eta_2) = 0$ it is usually assumed that the related difference of the average animal effects of treatments 1 and 2 is equal to zero. Thus, on the response scale the following is formally true:

$$\mu_1 = exp(\eta_1) = e^{\Delta\eta} \cdot exp(\eta_2) \quad with \quad \Delta\eta = (\bar{\alpha}_{1.} - \bar{\alpha}_{2.}) \tag{20}$$

Thus, a multiplicative connection exists between the means μ_1 and $\mu_2 = exp(\eta_2)$. If $\Delta\eta<0$, then $0<exp(\Delta\eta)<1$. The mean of treatment 2 is $exp(\Delta\eta)$-times smaller than the mean of treatment 1. For $\Delta\eta>0$, it follows that μ_2 is $exp(\Delta\eta)$-times larger than μ_1. A difference on the link scale can be associated with a multiplicative factor on the response scale. If the difference is significantly

different from zero, then the related factor can be formally seen to be significantly different from one. The estimated values of μ_1 (or μ_2) are generated in the respective SAS procedures under the assumption of $\bar{u}_{1.}=0$ (bzw. $\bar{u}_{2.}=0$). The estimations thus refer to animals whose effects equate to the mean of the population. No statements can be made about the differences from the marginal means given by (16) with the approach described here.

General discussion

The GLMM theory is valuable for evaluating count traits that are repeatedly recorded on the same subject. In this model, the repeated measures are taken into account by including random animal effects in the linear predictor. The observations for given animal effects are assumed to be realisations of Poisson or negative-binomially distributed random variables. In the so-called population-averaged models, the repeated measurements are depicted by a working correlation matrix. The assumption of normally distributed animal effects is then dropped. As a result, no log-likelihood function is available for selecting the model or estimating the model parameters. The convergence of the generalised estimating equations (GEE) in the solution becomes problematic if complex correlation structures are considered, or the correlation matrices are, for instance, seen to be dependent on the treatments. In this paper the repeated measurements (as is usual in animal breeding) are explicitly taken into account in the linear predictor. Thus, different correlation matrices can be specified for each treatment at every time point for the repeated observations per animal.

Count traits described by random variables, which satisfy a Poisson or NB-distribution, characteristically have variances that are functions of the expected values. Consequently, the correlations between repeated observations per animal are dependent on the expected values and the variance between the animals. Not every count trait must satisfy a Poisson or NB distribution. Thus, as part of the model selection, verification must be made as to whether the observed relative frequencies of the count outcomes are in accordance with the probabilities derived from the assumed distribution.

Aside from determining assumed distributions with respect to the observations and the random animal effects, considering qualitative and quantitative explanatory factors in the linear predictor is a central task of the model selection. The likelihood ratio test and the information criteria are available for this. With the help of local regression it must be decided which quantitative explanatory variable can be taken into account with a suitable, at least quasi-linear, regression function in the predictor. At this stage of the model selection the competitive models can be put into sequence with the help of the AIC values. A prerequisite of this recommendation is that if presenting random effects in the linear predictor, the model parameters must also be estimated during the setup and maximisation of the log-likelihood function. A residual analysis should be carried out on the selected model, as in the case of the linear models. The residual analysis for checking models (mainly known in connection with the use of linear mixed models) must be modified since in GLMM no normal distribution can be assumed for the scaled or standardised residuals. In GLMM, the residuals must have only the expected value of zero, independent of the observation number, quantitative covariables, or the estimated means for the level combination of qualitative explanatory factors. This characteristic can be checked in residual plots with the help of local regression, i.e. through smoothed trend curves.

Using a count trait exemplified by behavioural data derived from pigs with six consecutive observations per animal, it is shown which model approach within the GLMM with repeated measurements is best-suited for analysis. Assuming a Poisson or negative binomial distribution, a different (6×6) correlation matrix can be formally calculated for each of the four treatments investigated. In the GLMM, the conditional expected values are linked to the linear predictor by a so-called inverse link function. Converting the conditional to marginal expected values is demonstrated in the example. Marginal expected values take into account the existing variability between the animals and permit statements to be made, for example, about the influence of different treatments by averaging over all animals in a population. The basis for calculating the marginal expected values is the calculation of the averages over the levels of qualitative explanatory factors on the response scale. If means and a comparison of means are calculated on the link scale, the statements will refer to those animals whose effects correspond to the population's average. Using data on the aggressive behaviour of pigs, the paper illustrates and discusses which interpretations have significant differences on the links scale when it is transformed back to the response scale.

Emanating from the example it was shown how the empirical variance function, estimated with OLS residues and the variance functions of the GLMM, can be used for model verification. The presented graphical residual analysis can be extended to other types of residuals, such as the deviance or the Anscombe residuals. Additional diagnostic approaches can be found in McCullagh & Nelder (1989) and in Collett (1991).

The framework of generalised linear mixed models is available for analysing correlated count data. Repeated measures cannot be treated as independent observations because the degrees of freedom of statistical tests would be overestimated. If the correlation structure of repeated measurements on the same animal is ignored, an inappropriate covariance structure has to be used for testing the fixed effects. In the case of continuous, normally distributed traits, the mean and the variance are separate parameters. For counts following a Poisson or negative binomial model, there is a relationship between the variance and the mean. Thus, for model fitting and checking, the mean-variance link has to be taken into account. If count data were analysed using linear models under the assumptions of normality and variance homogeneity, the standard errors of the estimates would be incorrect and the Type I error of the statistical tests could dramatically exceed the nominal significance level α. The probability of incorrectly rejecting the null hypothesis can be larger than the nominal α value and statistical conclusions from the studies are invalid. Therefore the previous statements have to be tested by stochastic simulation.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19, 716-723

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88, 9-25

Carrasco JL (2010) A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data. Biometrics 66, 897-904

Carrasco JL, Jover L (2005) Concordance correlation coefficient applied to discrete data. Statist Med 24, 4021-4034

Collett D (1991) Modelling binary data. Chapman & Hall, London

Fabio LC, Paula GA, de Castro M (2012) A Poisson mixed model with nonnormal random effect distribution. Comput Stat Data An 56,1499-1510

Fahrmeir L, Tutz G (1994) Multivariate statistical modelling based on generalized linear models. Berlin, Heidelberg, New York, Springer

Foulley JL, Im S (1993) A marginal quasi-likelihood approach to the analysis of Poisson variables with generalized linear mixed models. Genet Sel Evol 25, 101-107

Hardin JW, Hilbe JM (2003) Generalized Estimating Equations. Boca Raton, FL, Chapman & Hall/CRC.

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

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22

McCullagh P, Nelder JA (1989) Generalized Linear Models (2nd edn). London: Chapman & Hall.

McCulloch CE, Searle SR (2008) Generalized, Linear and Mixed Models. New York, NY, John Wiley & Sons, Inc.

Mielenz N, Thamm K, Bulang M, Spilke J (2011) Generalized linear models with random effects for description of data with excess zeros. Arch Tierz 54, 661-675

Molenberghs G, Verbeke G, Demétrio CGB (2007) An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Anal 13, 513-531

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

Nelson KP, Lipsitz SR, Fitzmaurice GM, Ibrahim J, Parzen M, Strawderman R (2006) Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. J Comput Graph Stat 15, 39-57

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

Schmidt T, Calabrese JM, Grodzycki M, Paulick M, Pearce MC, Rau F, von Borell E (2010) Impact of single-sex and mixed-sex group housing of boars vaccinated against GnRF or physically castrated on body lesions, feeding behaviour and weight gain. Appl Anim Behav Sci 130, 42-52

Solis-Trapala IL, Farewell VT (2005) Regression analysis of overdispersed correlated count data with subject specific covariates. Statist Med 24, 2557-2575

Thall PF, Vail SC (1990) Some covariance models for longitudinal count data with overdispersion. Biometrics 46, 657-671

Thamm K (2012) Analysis of count data with repeated observations exemplified by trials from agricultural sciences. Diss., Martin-Luther-University Halle-Wittenberg [in German]

Appendix

Loglikelihood function of the multivariate negative binomial distribution (MNB):

$$logL(\beta, \Phi | y) = \sum_{i=1}^{a} log\ f(y_i | \mu_{i}, \Phi)$$

$$with\ log\ f(y_i | \mu_{i}, \Phi) = \sum_{j=1}^{n_i} y_{ij}\ log(\mu_{ij}) + y_{i.}\ log(\Phi) + log\Gamma\ (\Phi^{-1} + \mu_{i.})$$

$$-log\Gamma\ (\Phi^{-1}) - (\Phi^{-1} + y_{i.})\ log(1 + \Phi \cdot \mu_{i.}) - \sum_{j=1}^{n_i} log\ (y_{ij}!)$$

$$where\ y_{i.} = \sum_{i=1}^{n_i} y_{ij}\ and\ \mu_{i.} = \sum_{i=1}^{n_i} \mu_{ij}$$

Implementation of Model PoiLG in NLMIXED of SAS:

```
proc sort data=boar; by animal;
proc nlmixed data=boar method=gauss noad;
p_i=cdf('normal',a_i);
g_i1=gaminv(p_i, 1/theta_1);
g_i=theta_1*g_i1;
eta=<linear predictor>+log(g_i);
mu=exp(eta);
model trait~poisson(mu);
random a_i~normal(0,1) subject=animal;
run;
```

NLMIXED commands for the boar data using MNB

```
/* DF=number of animals – model parameters=233–(23+1)=209 */
/* one record per boar */
proc nlmixed data=sum1 df=209;
array b{4,6} b1–b24;
/* trt=4 and tp=4 does not exist */
parms b1–b21=0.1 b23=0.1 b24=0.1 phi=0.5;
bounds phi>0;
b22=.;
im=bdlg;
array y{6} y1–y6;
ysum=sum(of y1–y6);
mup=0.0;
yp=0.0;
loglik1=0.0;
const=0.0;
do jm=1 to 6;
 if(y{jm}^=.) then do;
 eta=b{im,jm};
 mup=mup+exp(eta);
 yp=yp+y{jm};
```

```
 loglik1=loglik1+y{jm}*eta;
 const=const+lgamma(1+y{jm});
 end;
end;
loglik2=yp*log(phi)+lgamma(1/phi+yp);
loglik3=−lgamma(1/phi)−(1/phi+yp)*log(1+mup*phi);
loglik=−const+loglik1+loglik2+loglik3;
model ysum~general(loglik);
run;
```

NLMIXED commands for the boar data using NBinN

```
proc sort data=boar; by animal;
proc nlmixed data=pigs1;
array b{4,6} b1−b24;
/* trt=4 and tp=4 does not exist */
parms b1−b21=0.1 b23=0.1 b24=0.1 phi=1.0 su2=0.2;
bounds phi>0, su2>0;
b22=.;
y=begRK_11_15;
im=bdlg;
km=mzp;
loglik1=0.0;
const=0.0;
/* linear predictor */
eta=b{im,km}+u;
mu=exp(eta);
loglik1=loglik1+y*eta;
const=const+lgamma(1+y);
loglik2=y*log(phi)+lgamma(1/phi+y);
loglik3=−lgamma(1/phi)−(1/phi+y)*log(1+mu*phi);
loglik=−const+loglik1+loglik2+loglik3;
model begRK_11_15~general(loglik);
random u~normal(0,su2) subject=tier_nr;
/* estimation of marginal means */
array bm_mu[4,6];
do i=1 to 4;
 do j=1 to 6;
  bm_mu[i,j]=exp(b[i,j]+0.5*su2);
 end;
end;
/* testing of differences */
estimate 'd12_1' (bm_mu[1,1]−bm_mu[2,1]);
estimate 'd13_1' (bm_mu[1,1]−bm_mu[3,1]);
estimate 'd23_1' (bm_mu[2,1]−bm_mu[3,1]);
run;
```