Colloquium Biometricum 41
2011, 115–126

AN APPLICATION OF THE GENERALIZED LINEAR MODELS FOR AN EXAMINATION OF THE PHENOTYPIC QUALITY OF ROE DEER

Joanna Ukalska[1], Krzysztof Ukalski[1], Jakub Borkowski [2]

[1] Biometry Division, Department of Econometrics and Statistics
Warsaw University of Life Sciences–SGGW
Nowoursynowska 159, 02–776 Warsaw, Poland
e–mail: joanna_ukalska@sggw.pl
[2] Department of Forest Ecology, Forest Research Institute
Braci Leśnej 3, Sękocin Stary, 05–090 Raszyn, Poland

Summary

The influence of forest environment (forest regeneration after a 1992 forest fire covered with young stands (low quality deer habitat) and unburned forest of diversified stand age classes (high quality deer habitat)) and climatic factors (the mean temperature and the total number of days with snow cover in January and February) on roe deer antler asymmetry in two age classes of roe deer males was studied. Data were collected by local hunters from 366 shot males during 1998–2007. We applied 4 generalized linear models: Poisson model, Poisson adjusted for overdispersion, negative binomial and negative binomial with log canonical link function. Goodness–of–fit statistics were checked as well as residuals plots. There was a significant difference in roe deer antler asymmetry incidence between age classes for both considered habitats while weather conditions didn't influence roe deer antler asymmetry.

Key words and phrases: count data, generalized linear model**,** negative binomial distribution, overdispersion, Poisson distribution, roe deer

Classification AMS 2010: 62J12

1. Introduction

Roe deer male quality can depend on an antler symmetry. By males with symmetric antlers we mean those with an even number of points on both antlers, otherwise they are asymmetric. Degree of antler symmetry can be an indicator of an environmental stress (e.g. malnutrition). Thus it is interesting to check the influence of forest environment quality on roe deer antler symmetry. Roe deer are the only cervid which antlers grow in winter, when the food conditions are the toughest among all the seasons. Furthermore it is also interesting if the snow cover and temperature influence antler growth (in this case its level of asymmetry).

The number of males are a typical count data therefore linear models are not suitable for them. Since 1972 when Nelder and Weddenburn (1972) used the term of generalized linear models (GzLM) and adapted linear model methodology for use with non–normal data – these models are typically applied for non–normal data. The main features of GzLM are the link function and the variance function (McCullagh and Nelder, 1989; Littel et al., 2002). The link function η is a mathematical model of the expected value μ of a random response variable *Y*. With normally distributed data we fit a linear model directly to the mean, but in GzLM we fit the linear model indirectly using a function of the mean – the link function. The variance function $V(\mu)$ describes the relationships between the expected value and the variance of the distribution of the response variable.

The log likelihood formula for exponential family distributions is considered as the function of θ, ϕ and y being given (McCullagh & Nelder):

$$\ell(\theta,\phi,y)=\frac{y\theta-b(\theta)}{\phi}+c(y,\theta) \tag{1.1}$$

where θ is the natural parameter, ϕ is a scale parameter and $b(\cdot)$, $c(\cdot)$ are specific functions. The second derivative of (1.1) is called the variance function $V(\mu)$. Thus var(y)= $\phi\,V(\mu)$.

Typically the Poisson distribution is used to model information on counts in situations where there is no natural “denominator” and thus no upper limit of size of an observed count. The probability of the count y is $P(Y=y)=\frac{\lambda^y e^{-\lambda}}{y!}$, where λ is the mean count. For the Poisson distribution the mean and the variance are equal $E(y)=\mathrm{var}(y)=\lambda$. The log likelihood of the count y is $\ell(y)=y\log(\lambda)-\lambda-\log(y!)$. According to (1.1) $\theta=\log(\lambda)$, ϕ=1 and $V(\mu)=\lambda$

(Littel et al., 2002). The link function in the Poisson distribution is η = log(λ) thus $\eta = \theta$. Link function having the form $\eta = \theta$ is called the canonical link function (Littel et al., 2002).

When the variance is larger or smaller than expected in a given model, it indicates overdispersion ($\phi > 1$) or underdispersion ($\phi < 1$) (Cameron & Trivedi, 1998; McCullagh & Nelder, 1989). The scale parameter ϕ is also called the dispersion parameter (McCullagh & Nelder, 1989). Overdispersion causes that standard errors are underestimated and test statistics are overestimated. For biological count data overdispersion occurs quite often (Littel et al., 2002) and the distribution where the variance is bigger than the mean could be more appropriate than the Poisson distribution. One such distribution could be a negative binomial distribution (Dean & Lawless, 1989).

The aim of this study was firstly, to check which of environmental factors and weather conditions influence roe deer's antler growth and secondly, to find out the most adequate generalized linear model describing those data.

2. Materials and methods

2.1. Data

Observations were collected by local hunters from shot roe deer bucks during 1998–2007. The total number of males was 366 where 106 bucks were with asymmetric antlers. The date and place (burned/unburned forest) of each shot buck were noticed. The number of points on each antler from a given male was counted and its age was estimated on the basis of tooth wear. We took into consideration 2 age classes – yearlings (i.e. 1.5 –years–olds) and three–year–old bucks. Roe deer antlers were compared between two contrasting areas: forest regeneration after a 1992 forest fire covered with young stands (low quality deer habitat) and unburned forest of diversified stand age classes (high quality deer habitat). In addition to environmental factors, the influence of climatic variables such as the mean temperature and the total number of days with snow cover in January and February was examined.

2.2. Statistical methods

We were interested in examining the asymmetry incidence i.e. the number of bucks with asymmetric antlers *y* divided by the total number of observed bucks in the ij^{th} age–habitat combination in k^{th} year N_{ijk}. Therefore the response variable *Y* was count variable which is usually analyzed using generalized linear models (GzLM). At the beginning we used a Poisson regression model which form was:

$$\eta_{ijk} = \log(\lambda_{ijk}) = \log(N_{ijk}) + m + a_i + h_j + ah_{ij} + \beta_1 t + \beta_2 s\,, \qquad (2.2.1)$$

where $\eta_{ijk} = \log(\lambda_{ijk})$ is the canonical log link function and $\log(\lambda_{ijk})$ is the mean count for the ij^{th} age–habitat combination in k^{th} year (k=1,…,10), m the intercept, a_i the i^{th} age effect (i=1, 2), h_j the j^{th} habitat effect (j=1, 2), ah_{ij} the interaction effect for ij^{th} age–habitat combination, t and s are respectively the mean temperature and the mean number of days of snow cover variables in January and February in considered years. β_1 and β_2 are regression slopes, which have convenient interpretation as the natural log of the antler asymmetry rate ratio for comparing a one unit increase of t or s respectively. The logarithm of the total number of bucks in each class $\log(N_{ijk})$ is an offset term allowing different number of males in considered classes. In matrix notation the model (2.2.1) has the form (Littel et al., 2002):

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}\,, \qquad (2.2.2)$$

where $\boldsymbol{\eta}$ is the N×1 vector of the link function, $\mathbf{X}$ is is the N×p design matrix and $\boldsymbol{\beta}$ is the p×1 vector of the model parameters.

For model (2.2.1) we checked the evidence of overdispersion using goodness–of–fit statistics. The overdispersion parameter ϕ is unknown and therefore must be estimated. A method suggested by McCullagh and Nelder (1989) is using the *deviance*, which is the measure of discrepancy between observed and fitted values. The *deviance* is defined as (Littel et al., 2002):

$$2\left[\ell(\theta(\mathbf{y});\mathbf{y}) - \ell\left(\theta\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right);\mathbf{y}\right)\right], \qquad (2.2.3)$$

where $\ell(\theta;\mathbf{y})$ is the log likelihood with $\theta(\mathbf{y})$ value determined from the data and $\theta\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)$ attained from the estimate of $\boldsymbol{\beta}$ under the fitted model, $\mathbf{y}$ is N×1 vector of observations. The *deviance* has an approximate χ^2 distribution with N–p degrees of freedom (N=total number of observation, p= number of the model parameters; rank of the design matrix $\mathbf{X}$). The *deviance* divided by N–p is the estimator of the unknown overdispersion parameter ϕ (McCullagh & Nelder, 1989) and it is used to detect over – or under – dispersion in Poisson models:

$$\hat{\phi} = \frac{deviance}{N - p}. \tag{2.2.4}$$

There are at least two ways to account for over – or under – dispersion in GzLM. One way is to adjust the covariance matrix of the Poisson model with the overdispersion parameter ϕ. Then the covariance matrix is pre–multiplied by ϕ and the scaled deviance and the log–likelihood ratio tests are divided by ϕ (Stokes et al., 2000). This approach was suggested by McCullagh and Nelder (1989).

Second way to manage overdispersion is to assume a more flexible distribution e.g. the negative binomial distribution with the mean λ and the variance function $\lambda + k\lambda^2$, where k is the aggregation parameter (Littel et al., 2002). The limiting distribution for negative binomial is Poisson when k=0. The negative binomial distribution adds a quadratic term to the variance representing overdispersion. For k>0 the variance is larger than the mean and the data are more aggregated (clustered) than would be expected in the Poisson distribution (Littel et al., 2002). The canonical link function for the negative binomial distribution is $\eta = \log\frac{\lambda}{\lambda + \frac{1}{k}}$. When k is unknown it has to be estimated. A simpler approach to the negative binomial distribution is using the log(λ) as the canonical link.

We decided to perform the analysis of the experimental data described by the model (2.2.1) in four ways. Firstly we used the “classical” Poisson model (model 1) assuming that the vector of observations **y** is a realization of the random variable Y having Poisson distribution. Secondly, assuming the same distribution of Y as in the model 1, we used adjusted for over – or under – dispersion Poisson model (model 2). Next we assumed that response variable Y had the negative binomial distribution and we used the negative binomial model with the $\log(\lambda_{ijk})$ as link function (model 3) and the negative binomial model with the link function of the form $\eta_{ijk} = \log\frac{\lambda_{ijk}}{\lambda_{ijk} + \frac{1}{k}}$ for estimated k (model 4).

To compare negative binomial models (models 3 and 4) with the Poisson ones (model 1 and 2) we carried out the likelihood ratio test for significance of overdispesion (Cameron & Trivedi, 1998), that is, the test of the hypothesis $H_0 : k = 0$ against $H_1 : k > 0$.

To check assessing of the fit of these models we analyzed two kinds of residuals plots suggested by McCullagh and Nelder (1989). One of them is the plot of standardized deviance residuals

$$\text{std}\, r_{ijk}^{D} = \text{sign}\left(y_{ijk} - \hat{\lambda}_{ijk}\right)\sqrt{deviance_{ijk}} \,/ \sqrt{1 - h_{ijk}}\ ,$$

where h_{ijk} is a function of the Hessian matrix, against the predicted counts $\hat{\lambda}_{ijk}$. The unequal scatter on this plot indicates violation of the homogeneity of variance (Littel et al., 2002). The next is the plot checking the link function. It is the plot of linear predictors $y_{ijk}^{*} = \hat{\eta}_{ijk} + \dfrac{y_{ijk} - \hat{\lambda}_{ijk}}{\hat{\lambda}_{ijk}}$ against estimated link function $\hat{\eta}_{ijk}$.

We used PROC GENMOD in SAS System for all computations (SAS Institute, 2008). Residuals values needed for model checking plots were obtained using OBSTAT statement of this procedure.

3. Results

The goodness–of–fit statistics for models 1–4 are presented in Table 1. The deviance has approximately chi–square distribution with the number of degrees of freedom presented in column df. For model 1 $\hat{\phi} = 39.628/33 = 1.201$ what indicates slight overdispersion. The scaled deviance for model 2 is the deviance for model 1 divided by the estimated overdispersion parameter $\hat{\phi}$. For models 3 and 4 both criteria are the same. The third row of Table 1 consists of chi–square statistics, and referred p–values for test fitting of presented models. For none of models 1–4 we do not reject the hypothesis that data come from assuming distributions.

Table 1. Criteria for assessing goodness–of–fit for models 1–4

		Model 1 and 2		Model 3		Model 4	
Criterion	df	Value	Value/df	Value	Value/df	Value	Value/df
Deviance	33[a]	39.628$^{\#}$	1.201$^{\#}$	43.141	1.307	44.117	1.337
Scaled deviance	33	33$^{\#\#}$	1$^{\#\#}$	43.141	1.307	44.117	1.337
Pearson χ^2 (*p–value*)	33	35.779 (0.339)	1.084	39.455 (0.204)	1.196	41.112 (0.157)	1.246
Log likelihood	Model 1		Model 2	Model 3		Model 4	
$\ell\left(\hat{\theta}\right)$ (p–value)	37.930		31.586 (<0.001)	38.047 (0.314)		38.075 (0.296)	

$^{\#}$ for model 1; $^{\#\#}$ for model 2

[a] – df=33 – any 3 year–old buck was shot at burned forest in 1999 thus *N*=39 (instead of 40), *p*=6

The results of the likelihood ratio test for significance of overdispesion i.e. the hypothesis H_0: k=0 are presented in the last row of Table 1. We do not reject H_0 for both negative binomial models and conclude that the Poisson model fits well. However, we used this test also to compare both Poisson models i.e. model 2 with adjustments for overdispersion with model 1. There is a significant difference between these models, so model 2 seems to be appropriate.

We checked residuals plots next. Figure 1 presents standardized deviance residuals $\mathrm{std}\, r_{ijk}^{D}$ versus the predicted mean $\hat{\lambda}_{ijk}$ for models 3 (Fig. 1b) and 4 (Fig. 1c) and for model 2 $\mathrm{std}\, r_{ijk}^{D}$ against predicted mean adjusted to a constant information scale $2\sqrt{\hat{\lambda}_{ijk}}$ (Fig. 1a) as was suggested by McCullagh and Nelder (1989). There is no overt visual evidence on unequal scatter or systematic pattern on plots in Fig. 1 and the absolute values of standardized deviance residuals aren't greater than 2.5, so none of compared models can be rejected on the basis of these plots.

Figure 2 presents linear predictors y^* plotted against the estimated link function $\hat{\eta}_{ijk}$. These plots are unique to GzLMs (Littel et al., 2002). They should be linear, departure form linearity suggests a poor choice of the link function (Littel et al., 2002). There is the visible scatter on each of Fig.2 a–c plots and no overt departures from linearity and hence no obvious evidence of a poor choice of the link functions for considered models.

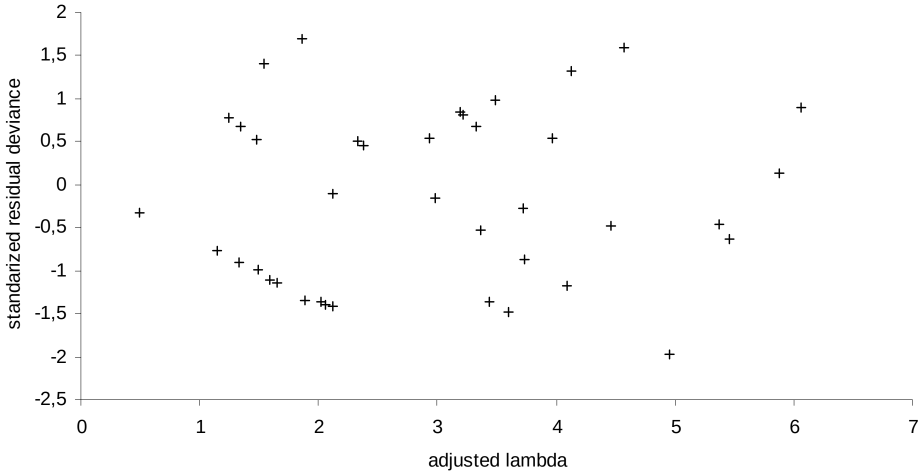

Fig. 1a. The plot of standardized deviance residuals against adjusted predicted mean (adjusted lambda) for model 2

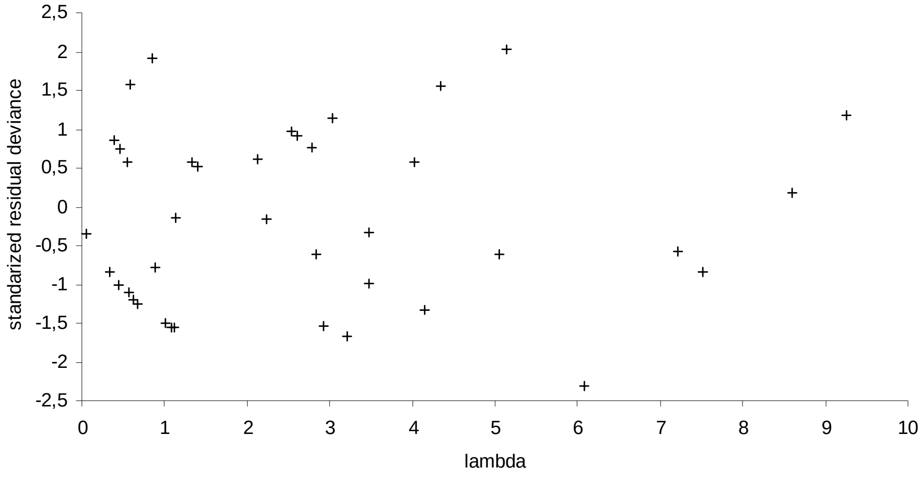

Fig. 1b. The plot of standardized deviance residuals against adjusted mean $\hat{\lambda}_{ijk}$ for model 3

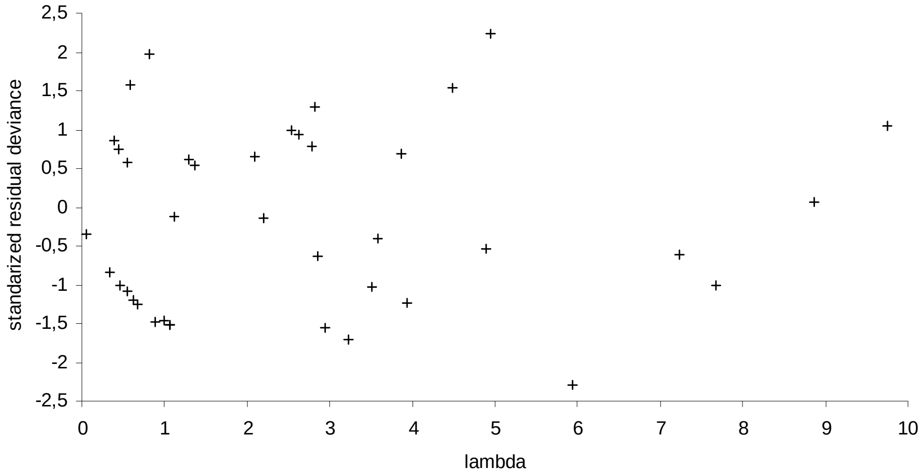

Fig. 1c. The plot of standardized deviance residuals against predicted mean $\hat{\lambda}_{ijk}$ for model 4

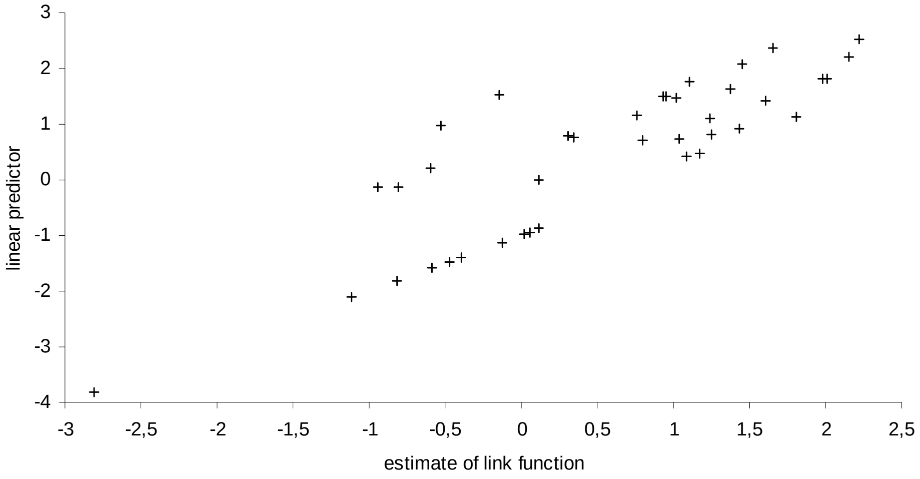

Fig. 2a. The plot of linear predictors y^* against estimated link function $\hat{\eta}_{ijk}$ for model 2

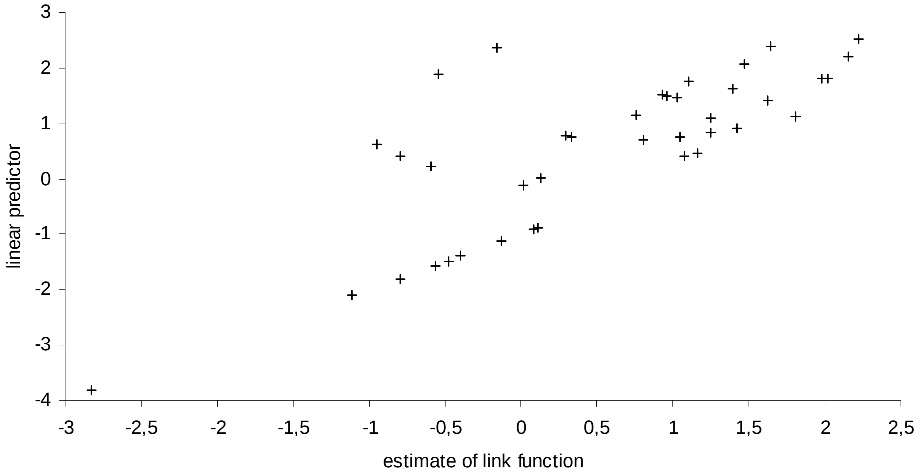

Fig. 2b. The plot of linear predictors y^* against estimated link function $\hat{\eta}_{ijk}$ for model 3

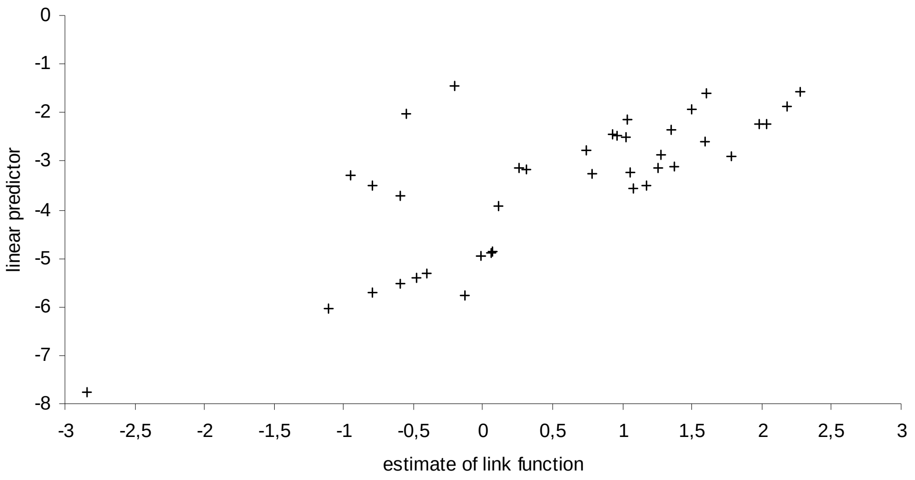

Fig. 2c. The plot of linear predictors y^* against estimated link function $\hat{\eta}_{ijk}$ for model 4

None from all models 2–4 was eliminated due to lack of the fit. Therefore we wanted to check if there were differences between results of testing the effects of the model (2.2.1). Table 2 presents the results of the type 3 analysis for all considered models. It contains likelihood ratio chi–square test statistics and associated p–values. It's worth pointing out that LR chi–square values for model 1 (the Poisson model without adjustments) are bigger than in model 2 (the Poisson model with adjustments for overdispersion). It increases because the standard errors of the estimated effects are biased and they are too small what makes test statistics overestimated.

Table 2. Likelihood ratio statistics (LR) for type 3 analysis for models 1–4

Source of variation	**df**	**Chi–square statistics (p–value)**			
		Model 1	**Model 2**	**Model 3**	**Model 4**
habitat	1	3.09 (0.078#)	2.58 (0.108)	3.27 (0.070)	2.66 (0.103)
age	1	31.50 (<0.001)	26.23 (<0.001)	26.58 (<0.001)	26.97 (<0.001)
habitat*age interaction	1	4.79 (0.028)	3.99 (0.045)	5.01 (0.025)	5.80 (0.016)
temperature	1	0.09 (0.758)	0.08 (0.779)	0.12 (0.731)	0.13 (0.718)
snow	1	0.24 (0.626)	0.20 (0.657)	0.32 (0.572)	0.37 (0.540)

(*p*–values in brackets)

These results for all models are similar. There are significant differences of age groups as well as the interaction of experimental factors. There are no significant differences between habitats that are burned and unburned forest. Considered climatic variables – the temperature and the number of days of snow cover in January and February are not significant so the influence of these weather conditions on roe deer antler asymmetry was not confirmed.

We obtained the same results for all considered models but finally we chose the Poisson model with adjustments for overdispersion (model 2) as the best model for presented data. Our decision was determined mainly by the fact that the Poisson distribution is the most common distribution for modelling count data and the negative binomial distribution is applied generally in these situations where the Poisson model is a poor fit.

Table 3 presents roe deer antler asymmetry incidence scores and their significance for main and interaction effects of the model (2.2.1). There is a significant difference in roe deer antler asymmetry incidence between age classes for both considered habitats. The incidence of antler asymmetry for yearlings was lower than for 3 –years–old ones in high as well as for the low quality deer habitat. Younger bucks with asymmetric antlers were more rarely observed at the areas of the forest regeneration covered with young stands than in unburned forest, while 3 –years–old ones with asymmetric antlers lived in both types of habitats at the comparable level.

Table 3. Roe deer antler asymmetry incidence scores for main and interaction effects (ns – not significant

	Age category			
habitat	**yearlings**	**3–years–old bucks**	**significance**	**total**
unburned forest	19.4%	44.8%	***	32.6%
burned forest	6.8%	50.0%	***	20.4%
significance	*	ns		ns
total	14.6%	45.8%	***	29.0%

* – p<0.05, ** – p<0.01, *** – p<0.001)

4. Conclusions

1) Three of four considered generalized linear models: the Poisson model adjusted due to slight overdispersion, two negative binomial models with different log–link functions and fitted presented count data well.
2) Roe deer antler asymmetry was significantly lower for yearlings than for 3–years–old males, while it was observed on similar level in unburned and burned forest. The incidence of antler asymmetry for younger bucks was lower in burned than in unburned forest, but for 3–years–old bucks this incidence was similar in both considered habitats.
3) The weather conditions such as average temperature and average number of days with snow cover in January and February haven't affected roe deer antler asymmetry.

References

Cameron A.C., P.K. Trivedi (1998). *Regression analysis of count data*. Cambridge University Press.

Dean C., Lawless J.F. (1989). Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association* 84 (406), 467–472.

Littell R.C., Stroup W.W., Freund R.J. (2002). *SAS for Linear Models*. 4d ed. SAS Institute Inc., Cary, NC.

McCullagh P., Nelder J.A. (1989). *Generalized Linear Models*. 2d ed. Chapman and Hall, London.

Nelder J.A., Weddenburn R.W.M. (1972). Generalized Linear Models. *J. R. Statist. Soc*. A 135, 370– 384.

SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Stokes Maura E., Charles S. Davis, Gary G. Koch. (2000). *Categorical data analysis using the SAS System* 2d ed. SAS Institute Inc., Cary, NC.