|  |  |
| --- | --- |
| **Biostokastikum** | 2014-06-12 |

Exercise 5 – Zero inflated data

Hemmingsen et al. (2005) examined a large number of cod for counts of a parasite. The investigation was carried out during three years in four areas along the coast of Norway. The count of parasites and the length of each fish were measured. Our dataset, which was taken from Zuur et al. (2009), includes the variables Intensity (i.e. the count of the parasite), Area, Year, and Length.

Build a generalized linear model for the intensity of the parasite. Consider a model with fixed effects of the factors Area and Year. Consider Length as a covariate (i.e. a continuous explanatory variable) and include possible interactions. The dataset includes many zeroes.

## Hints for R users

The dataset Cod.txt is a tab separated text file that can easily be read into R using the read.table function with header = TRUE.

Generalized linear models for count data can be fitted using the glm function with family “poisson”, or using the glm.nb function, which fits a negative binomial distribution. Zero-inflated Poisson and zero-inflated negative binomial models can be fitted using the zeroinfl function of the pscl package. This function handles dist = “poisson” and dist = “negbin”. The formula should include two parts that are separated by a vertical bar. To the left, the model for the counts is given, and to the right the model for the probability of excess zero. These models can, but need not, be the same.

Models fitted to the same data but using different functions can be compared using the AIC criterion, which shall be as small as possible. The anova function is not applicable to fitted objects from the zeroinfl function, but likelihood ratio tests (for comparison of two models, such that one is a special case of the other) can be carried out using the lrtest function of the lmtest package.

## Hints for SAS users

The SAS file Exercise5.sas includes the cod dataset.

Generalized linear models for count data can be fitted using the genmod procedure function with the option dist = poisson or dist = negbin in the model statement. Zero-inflated Poisson and zero-inflated negative binomial models can also be fitted using the genmod procedure, through dist = zip and dist = zinb, respectively. However, for zero inflated models, the zeromodel statement must also be included in the genmod procedure. In the zeromodel statement, the model for the probability of excess zeroes is specified. The model specified in the zeromodel statement can, but need not, be the same as the model specified in the model statement.

Models fitted to the same data but using different functions can be compared using the AIC criterion, which shall be as small as possible. Also study the p-values that are provided by the genmod procedure.

## References

Hemmingsen W., Jansen P. A., MacKenzie, K. (2005). Crabs, leeches and trypanosomes: an unholy trinity? *Marine Pollution Bulletin* 50, 336–339.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). "Mixed effects models and extensions in ecology with R," Springer, New York.