Use the GLM, Luke!

How the use of proper statistical models can increase statistical power in ecotoxicological experiments.

Eduard Szöcs

September 25, 2014

1 Introduction

In environmental risk assessments statistical tests play an important role to evalute the effects of pesticides, e.g. in finding the Lowest Observed Effect Concentration (LOEC). Ecotoxicologists perform various kinds of experiments yielding to different types of data. Examples are: animal counts in mesocosm experiments (positive and discrete data), proportions of surviving animals (bonded between 0 and 1) or biomass (positive data).

These types of data are inherently not normally distributed. However, in order to use the more familiar and traditional data analysis methods based on normal distribution, ecotoxicologists try to transform their data to met these assumptions. Survivals with binomial distributed data (x out ouf n) are usually transformed using a arcsin square root transformation (Newman, 2012; OECD, 2006), count data using a $\log(Ax + 1)$ transformation (van den Brink et al., 2000). If the transformed data does not met the normality assumptions, non-parametric tests are usually applied (Wang and Riffel, 2011).

However, there is also a third possibility: Using a distribution fitting to the type of data instead of the normal distribution, namely *Generalized Linear Models* (GLM) (Nelder and Wedderburn, 1972). GLMs can handle various types of data distributions, e.g. poisson or negative binomial (counts), binomial (proportions) or gamma (mass); the normal distribution being a special case. Despite that GLMs were available more than 40 years now, ecotoxicologist do not make use of them. One reason might be that they are not mentioned in standard guidance documents for statistical analysis of ecotoxilogical experiments: GLMs are not discussed in the (OECD, 2006) and (EPA, 2002) guidelines and are not included in the respective schemes for hypothesis testing.

Newman (2012) does not cover GLMs in his book. Environment Canada (2005) term GLMS as *useful for toxicological research*, but did not include them in their schemes as *the concept is quite advanced and as yet is not widely used in environmental toxicology.*

Recently studies concluded that data transformations should be avoided and GLMs be used as they have better statistical properties (*Do not log-transform count data*, (O'Hara and Kotze, 2010); *The arcsine is asinine*, (Warton and Hui, 2011)). In this paper we first review what types of analysis are used by ecotoxicologists. Then we demonstrate that GLM provide superior properties compared to data transformation. We used simulated data that mimicked data encountered in ecotoxicology to compare GLMs with two common data types: counts and proportions. Methods were compared in terms of Type I error (maintain a significance level of 0.05 when there is no effect) and power (detect an effect when it is present).

2 Methods

2.1 Review

Literature review of SETAC Journals -¿ NOEC, GLM, log/arcsine transformation? How often, what is applied?

2.2 Simulations

2.2.1 Count data

We simulated count data that mimics count data encountered in mesocosm experiments, with five treatments (T1 - T5) and one control group (C) (e.g. (Brock et al., 2014)). Counts were drawn from a negative binomial distribution with a fixed dispersion parameter for all treatments ($\theta = 3.91$). We simulated datasets with different the number of replicates (N = {3, 6, 9, 12}) and different abundances in control treatments (μ_{C} = {2, 4, 8, 16, 32, 128, 512, }). For each combination we generated 100 datasets. For power estimation mean abundance in treatments T2 - T5 was reduced to half of the control treatment and T1 ($\mu_{\mathrm{T2}} = ... = \mu_{\mathrm{T5}} = 0.5\mu_{\mathrm{C}} = 0.5\mu_{\mathrm{T1}}$). For Type I error estimation mean abundance was equal between all groups.

We fitted two parametric models to this data. First a model assuming a normal distribution after a ln(Ay + 1) transformation the response (eqn 2).

$$y_i^T = log(Ay_i + 1) \tag{1}$$

$$y_i^T \sim N(\mu_i, \sigma^2)$$

$$y_i^T = \alpha + \beta x_i \tag{2}$$

$$var(y_i^T) = \sigma^2$$

where y is the measured abundance and A was selected in such a way that the lowest non-zero abundance of the data set is transformed to 1 (van den Brink et al., 2000). y^T is the transformed abundance and the model assumes a constant variance.

And second GLM assuming that the response follows a negative binomial distribution with a log-link (eqn 3) and the variance is a quadratic function of the mean.

$$y_i \sim NB(\mu_i, \theta)$$

$$log(\mu_i) = \alpha + \beta x_i \tag{3}$$

$$var(y_i) = \mu_i + \mu_i^2/\theta$$

Additionally we also analysed this data using a non-parametric approach (see below).

2.2.2 Binomial data

We simulated binomial data that mimics data encountered in survival experiments, with five treatments (T1 - T5) and one control group (C) (e.g. (Newman, 2012), example 5.1 therein). Each group consisted of 10 animals and was sampled from a Bin(10, p) distribution. For power simulation p in the control and T1 was held constant at p = 0.9 and varied for T2 - T5 (p = {0.5, 0.55, 0.60, ... , 0.90}). For Type I error simulations p was equal between groups ($p_c = p_{T1} = ... = p_{T6}$; p = {0.5, 0.55, 0.60, ... 0.90}). In both simulations we also varied the number of replicates (N = {3, 6, 9, 12}). For each combination we generated 100 datasets.

We fitted two parametric models to this data. First a model assuming a normal distribution after arcsin transforming the response (eqn 4 + 5, (EPA, 2002)).

original reference Bartlett 1937!

$$y_i^T = \begin{cases} arcsin\sqrt{1/4n}, & \text{if } y_i = 0 \\ arcsin(1) - arcsin\sqrt{1/4n}, & \text{if } y_i = 1 \\ arcsin\sqrt{y/n}, & otherwise \end{cases} \tag{4}$$

where y are the number of dead animals, y^T are the transformed proportions and n is the number of animals per replicate (n = 10).

$$\begin{aligned} y_i^T &\sim N(\mu_i, \sigma^2) \\ y_i^T &= \alpha + \beta x_i \\ var(y_i^T) &= \sigma^2 \end{aligned} \tag{5}$$

And second a model assuming a binomial distribution of the response with a logit link (eqn 6).

$$\begin{aligned} y_i &\sim Bin(\pi_i, 10) \\ logit(\pi_i) &= \alpha + \beta x_i \\ var(y_i) &= 10 \times \pi_i \times (1 - \pi_i) \end{aligned} \tag{6}$$

where y is the number of dead animals (out of ten) and π the probabilty of encountering a dead animal.

2.2.3 Comparing methods

A global treatment effect was investigated using Likelihood-Ratio-Tests. Moreover, we applied a non-parametric Kruskal-Wallis-Rank-Sum-Test on the untransformed data.

We also investigated power and Type I error for detecting the LOEC (T1 in the simulations). For the parametric models we used the parametrisation of contrasting to the control group (equivalent to Dunnett contrast). We used a pairwise Wilcoxon-Rank-Sum-Test as a anon- Wald!-te parametric method to detect the LOEC. P-values were adjusted for multiple tesing using Holm's procedure.

All simulations have be done in R (Version 3.1.1) on a 64-bit Linux machine with 8 GB and 2.2 GHz. Exemplary analysis of real data (Count data (Brock et al., 2014), survial data (EPA, 2002)) can be found in the supplement. Source code for this work is available online at `https://github.com/EDiLD/usetheglm`.

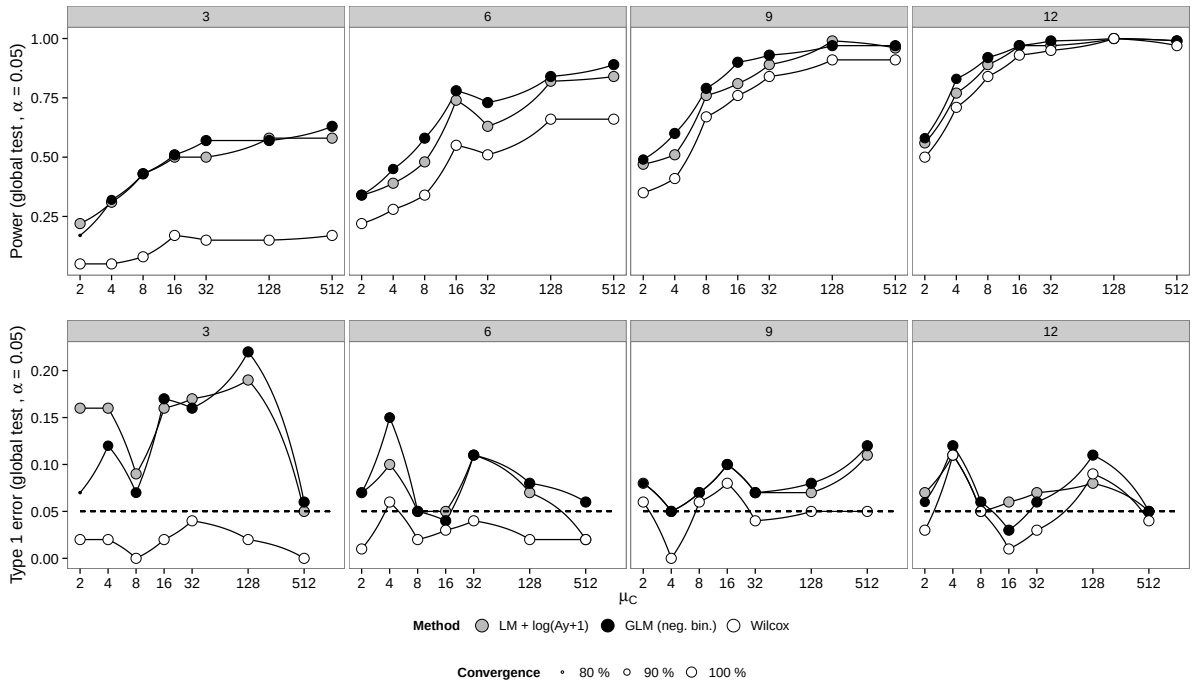

Figure 1: Simulation results from power (upper) and Type I error simulations (lower) for count data.

3 Results

3.1 Review

3.2 Simulations

4 Discussion

5 Conclusion

References

Brock, T. C. M., Hammers-Wirtz, M., Hommen, U., Preuss, T. G., Ratte, H.-T., Roessink, I., Strauss, T., and Van den Brink, P. J. (2014). The minimum detectable difference (MDD)

Check le end -¿ wilcox is not cor- rect!

Double Check results! Line by Line!

Why did we find increased Type I e ror and Warton did not?

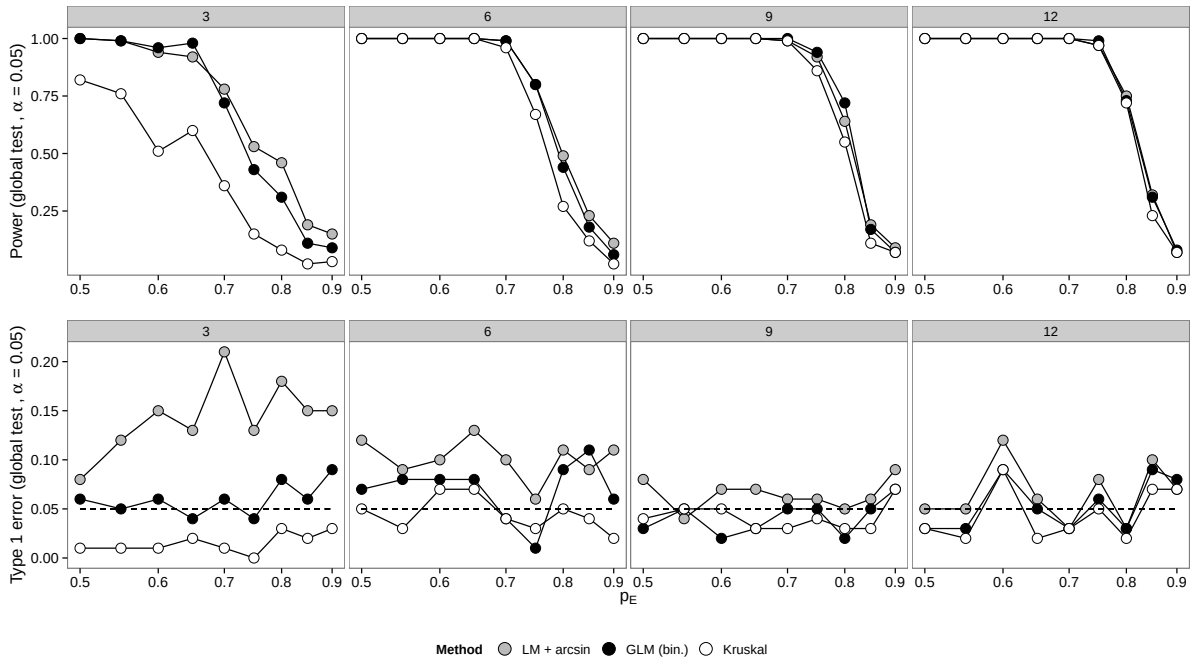

Figure 2: Simulation results from power (upper) and Type I error simulations (lower) for binomial data.

and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environmental Science and Pollution Research*.

Environment Canada (2005). *Guidance document on statistical methods for environmental toxicity tests*. Environment Canada, Ottawa.

EPA (2002). *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Newman, M. C. (2012). *Quantitative ecotoxicology*. Taylor & Francis, Boca Raton, FL.

OECD (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application*. Number 54 in Series on Testing and Assessment. OECD, Paris.

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122.

van den Brink, P. J., Hattink, J., Brock, T. C. M., Bransen, F., and van Donk, E. (2000). Impact

of the fungicide carbendazim in freshwater microcosms. II. zooplankton, primary producers and final conclusions. *Aquatic Toxicology*, 48(2-3):251–264.

Wang, M. and Riffel, M. (2011). Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology. *Ecotoxicology and Environmental Safety*, 74(4):684–92.

Warton, D. I. and Hui, F. K. C. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10. pdf RS.