Responses to reviewers

Ms. No. ESPR-D-15-00741
submitted to
Environmental Science and Pollution Research

Eduard Szöcs and Ralf B. Schäfer

March 16, 2015

Dear editor Dr. Schulz and reviewers,

We are thankful for reviewing our manuscript and the comments that helped to improve the paper. We revised the manuscript according to comments of the three reviewers and are re-submitting the manuscript for consideration for publication in Environmental Science and Pollution Research.

In the remainder of this document, we describe the changes that we have made to the paper for resubmission. To assist the assessment of our changes we have submitted two versions of the revised manuscript: one with highlighted changes and another without any highlighting. Note, that changes in titles, citations and figures were not highlighted.

Kind regards,
Eduard Szöcs and Ralf B. Schäfer

ToDo:

1. Now also added poisson GLM. incorporate into text!
2. first t1 then power.
3. recompile supplements
4. Holm also for case-study. Recalculate and rewrite 3-brock.R
5. update sysinfo in readme
6. Comment 62 umsetzen.
7. abstract:

Response to Reviewer 1

Comment 1: *"The purpose of the manuscript is highly commendable. Unfortunately, the present manuscript has some major shortcomings: The simulation results, which constitute the key contribution, are far from being as unequivocal as the title may lead one to think. Also, it is disturbing to see a fairly technical (and for ESPR non-standard) formulation of statistical models and yet the authors seem to lack some understanding of the models in some cases."*

Response: aaaa

Comment 2: *"About the title: Please consider a more informative title such as "Comparison of statistical approaches for analysis of non-normally distributed response in ecotoxicology". Also in view of the fact that the results are not very clearly favouring the generalized linear models (GLMs)."*

Response: We agree and changed the subtitle to
"A comparison of statistical approaches for analysis of non-normally distributed data in ecotoxicology".

Comment 3: *"p. 2, left column, lines 2–12: The issue of LOEC/NOEC versus regression is interesting but not really relevant to the aim of the manuscript. Consider skipping this part. However, you're right that for small sam- ple sizes it is extremely important to choose the most efficient statistical analysis and the better the model describes the data the more efficient."*

Response: We shortened this section. However, we sticked with a short introduction to LOEC/NOEC as inference on parameters is also part of our results.

Comment 4: *"p. 2, left column, lines 56–59: You aren't mentioning the issue with back-transformation So, in particular, βx is not additive changes from control on the original scale. To many ecotoxicologists such estimates on a transformed scale aren't directly interpretable as claimed by the authors. A more detailed explanation or at least discussion on how to back-transform is warranted (this is an issue regardless of whether GLMs or models for transformed responses are used!)."*

Response:

Comment 5: *"p. 2, right column, line 1: The real advantage of using a generalized linear model for count data shows up in case of many small counts and in the presence of ties. For large counts (e.g., corresponding to a large μ_C) there is practically no difference. However, this difference may not show up in simulations when the sample size is kept between 3 and 9."*

Response: We agree that for large counts there is no practical difference between methods. However, our simulation design aimed to mimic data sets frequently encountered in ecotoxicology. See also response to comment 13.

Comment 6: *"p. 2, right column, lines 19–21: It is not the response variable that is linked to the linear predictors, it is the mean of the response variable (McCullagh & Nelder, 1989). This is a basic fact of generalized linear models as you also show in Eqn. (3). Please explain Eqn. (3) in more detail (explain the parameters)."*

Response:

Comment 7: *"p. 2, right column, lines 25–27: There exists no quasi-Poisson distribution!!! You use R terminology, but this is not in this case statistical terminology. Please consult McCullagh & Nelder (1989) to understand that over-dispersion is dealt with by means of an ad hoc scaling of the standard errors that takes place after having fitted the ordinary Poisson GLM. So Eqn. (4) doesn't make sense."*

Response:

Comment 8: *"p. 2, right column, Eqn. (5): Have you explained the parameters in the text?"*

Response:

Comment 9: *"p. 3, left column, lines 26–42: What about over-dispersion for binomial data? Shouldn't that also be mentioned now that you consider models for over-dispersion for count data."*

Response: We added that overdispersion might also occur in binomial data and pointed the reader to Warton and Hui (2011).

Comment 10: *"p. 3, left column, line 33: Why not specify the mean next to the variance in all equations? Or skip the variance where not needed (as in this case)."*

Response:

Comment 11: *"p. 3, right column, Fig. 1: Please be precise in the figure text: By using the Poisson model the width of confidence intervals is underestimated in the presence of overdispersion."*

Response: We agree and stated more precisely that the width of confidence intervals is underestimated.

Comment 12: *"p. 3, right column, line 28: How is multiple testing taken care of? This may also affect the simulation results."*

Response: We adjusted multiple testing using the method of Holm (1979). We added this information to the manuscript.

Comment 13: *"p. 4, left column, line 2: The simulations will not be very informative for such small sample sizes (3–9) as all methods will have low power. Suggestion: Do also consider larger sample sizes: 12, 25, 50, 100. Not completely unrealistic any more as experiments tend to get larger. And then instead keep the μ_C small: 2, 4, 8, 16 to get more ties."*

Response: Our simulation design aimed to mimic data sets frequently encountered in ecotoxicology. Sample sizes between 3 and 9 are very common in regulatory ecotoxicology. For example, mesocosm experiments rarely exceed four replicates per treatment due to logistic constraints (Szöcs et al, 2015). Similarly the *Daphia magna* standard test (OECD, 2004) requires only a minimum of four replicates. We believe that the current simulation design is relevant for ecotoxicologists and did not

simulate larger sample sizes. Moreover, we provide a fully reproducible source code at `https://github.com/EDiLD/usetheglm`, enabling other researchers to easily adapt our simulation design to specific needs.

Comment 14: *"p. 4, left column, lines 10 and 28: In simulation studies it is quite common to use 1000 simulated datasets for each scenario (it would also reduce the sampling variability in the results)."*

Response: We agree and rerun our simulations generating 1000 datasets for all scenarios. Because of the increased computational burden we run the computation in parallel on Amazon EC2 instance (with 15 GB RAM, 2.8 GHz and 8 cores) - taking more then 12 hours for the count data simulations.

Comment 15: *"p. 4, left column, line 36: Please do define the type I error properly."*

Response: We agree and rephrased to:
"[...] Type 1 error (detection of an effect when there is none) [...]"

Comment 16: *"p. 4, right column, lines 26–29: The lack of convergence is not surprising as the simulated datasets are simply too small to fit complex models. Please re-consider the entire concept of the simulation (in particular which scenarios to consider)."*

Response: Our simulations covered scenarios often encountered by ecotoxicologists and therefore, it is like that practicing ecotoxicologists will also encounter such convergence problems. See also response to comment 13.

Comment 17: *"p. 4, right column, lines 31–33: One explanation for the reduced power is certainly the multiple testing issue. Perhaps this point could be addressed?"*

Response:

Comment 18: *"p. 4, right column, lines 57–58: Did LM really maintain the nominal significance level in all cases? This seems to be a subjective assessment (as the level is above in some cases for $N = 3$)."*

Response: We rerun our analyses using 1000 simulations (see also comment 14) and updated this section.

Comment 19: *"p. 6, right column, lines 43–46: Why didn't you also use the transformation suggested by Brock et al. (2015)? Would be useful to see the comparison."*

Response: As pointed out in the introduction, there is little information and advice which transformation to use. The $log(A\ y+1)$ is the most commonly used transformation in ecotoxicological mesocosm experiments. We used only the $log(A\ y+1)$ transformation because our paper did not aim to compare different transformations and there are some inconsistencies in Brock et al (2015). They state:

"For the examples presented in this paper, we followed Van den Brink et al. (2000) using the transformation y(x)=ln(ax+1), where x is the measured abundance and the factor 'a' is selected in such a way that the lowest non-zero abundance of the data set is transformed to 1."

Maybe this is just a typographical error, but in the publication of van den Brink et al (2000) it is stated:

"We decided that the factor Ax in the ln(Ax +1) transformation should make 2 by taking the lowest abundance value higher than zero for x."

Moreover, in the supporting material of Brock et al (2015), they used a $log(2\ y+1)$ transformation to which we refer to in our manuscript.

Comment 20: *"p. 6, right column, line 49: Dunnett test is R terminology (yes, it may be found in other recent publications, but it doesn't make it more correct). Please use a term like "comparisons to the control"."*

Response: We agree and changed accordingly.

Comment 21: *"p. 7, right column, lines 7–10: A nice result that is not found in many publication: non-parametric approaches lack power (due to less assumptions being made). However, your results for the parametric models are all quite similar and as it stands the paper does not present a strong case for the use of GLMs; and a more balanced abstract would be appropriate."*

Response:

Comment 22: *"p. 7, Simulations: A suggestion: remove all diverging discussion of LOEC/NOEC, GLMs for multivariate data, GLMMs, sample size cal- culation and concentrate on discussing the simulation results. Or, alter- natively, discuss in much more detail that GLMMs may actually offer a difference approach for handling over-dispersion in binomial and count data (instead of considering completely different*

designs).”

Response:

Response to Reviewer 2

Comment 23: *”One point to consider in a resubmission is that GLMs aren’t for use for all non-normal data - some qualifications are needed of when to use GLMs. Distributionally, GLMs replace the assumption of normality with the assumption that data come from the exponential family, which does not cover all situations. More critically, GLMs replace the assumption of equal variance with an assumption that the mean-variance relationship follows a pre-specified pattern, and violation of this assumption has similar impacts to violations of the equal-variance assumption in a linear model would have. Particular situations where GLMs are useful are binary data (known to be a quadratic mean-variance), count data (especially count data with lots of low counts, which typically cannot be transformed to normality effectively), and proportions constructed from counts. These are the examples highlighted in the text, and are of clear relevance in ecotoxicology, but the argument does not extend much more broadly than these examples - it is not a case of ”your data aren’t normal so you should use GLMs”. It is more accurate to say ”if you have binary, counts or proportions from counts you should use GLMs”. I think this requires a modest adjustment to the phrasing of the abstract and introduction to better qualify the situations where GLMs are appropriate. e.g. where O’Hara and Kotze and Warton & Hui are mentioned, this should be qualified to refer to counts and proportions (of counts) respectively. I don’t think the title needs changing, sufficient qualification elsewhere of the type of non-normality if interest should do the trick.”*

Response:

Comment 24: *”In the first paragraph, ”positive” is used where ”non-negative” is needed (i.e. zero is a possible value), also change bonded to bounded”*

Response: We fully agree and changed accordingly.

Comment 25: *”Page 2, second column - the parameterisation isn’t equal to the Poisson model - the mean model is the same, I think that is what was meant.”*

Response: We edited this section (see also comment 6) and corrected this.

Comment 26: *"Page 2, second column - also, strictly speaking, not all the methods mentioned here are GLMs. Negative binomial regression ("NB2") is a GLM if the overdispersion parameter is fixed, otherwise it is a generalisation of GLM. Similarly, quasi-Poisson isn't really a GLM, but a related method."*

Response:

Comment 27: *"Section 2.2 - a description of what is meant by binomial data would be helpful at the start of this section. Perhaps for completeness a description of what is meant by counts would help at the start of section 2.1"*

Response:

Comment 28: *"Section 2.2.2 - it would be useful to also mention GLMMs with a random intercept term. Often with this sort of data there are extra sources of variation from one sample to the next that are not explained by the binomial assumption, and this is the most natural way to account for it."*

Response:

Comment 29: *"Sections 2.1, 2.2: references to relevant articles or texts would be useful for GLMs, where the reader can find more details at a level appropriate for readership (e.g. a Zuur text? Quinn & Keough?). Also point readers to supp 2 for code and a worked example."*

Response:

Comment 30: *"page 3, line 58, the parametric bootstrap"*

Response: Changed accordingly.

Comment 31: *"Page 3, second column, line 30: Wald tests should be used with caution, when a mean estimate is zero (or close to it) they can have strange behaviour (related to the issue of separable data - parameter estimates and ses diverging, test statistics going to zero)."*

Response:

Comment 32: *"Page 4, second column, lower means: it is noted that the means are lower when data are transformed. This is known as transformation bias, and occurs*

because the transform linear model is actually estimating something different - it is no longer trying to estimate mean response. It would be helpful to say this and maybe include a useful reference. This also gets at the issue of interpretability - a disadvantage of the transform approach is a loss of interpretability, because we are no longer modelling mean response."

Response:

Comment 33: *"page 7 first column, line 7: bounds not bonds"*

Response: Thanks for pointing to this typo.

Comment 34: *"Section 4.1 the case study discussion was a little confusing. So is the point that you can get different results by linear models according to transformation, but that GLMs can resolve the problem of choice of transformation? (by replacing it with a decision about mean-variance relationship?) I think a stronger point is that the GLM is usually a better fit to count data, no transformation can make data Gaussian when it has lots of zeros and small counts, but GLM is designed for such data. If diagnostic tools support this argument for your case study they could be included."*

Response:

Comment 35: *"page 7 column 1 line 27: it is not so much unreliable and biased - they are estimating a different parameter (in an unbiased fashion) - they estimate the mean of log(Ax+1) data, and a problem is that this has no natural interpretation in terms of the original data. It is fair to make the point that it is biased as an estimate of the mean response, but it would be helpful to explain why and how the estimators have problematic interpretations."*

Response:

Comment 36: *"page 7 column 1 line 50: a priori not a priory*
page 7 column 2 line 37: data are (data being the plural of datum)
page 7 column 2 line 42: GLMs have (or GLM has)"

Response: We are thankful for pointing to this errors and changed accordingly.

Comment 37: *"page 7 column 2 line 44: the higher power in Fig 5 seemed to be through*

the GLMs being more conservative, rather than due to greater efficiency. This issue could likely be fixed using a parametric bootstrap, as was done previously in figure 2. A related point is that the linear model had the advantage that it seemed to be better at maintaining nominal significance levels at small sample sizes, which should be mentioned. As the authors say however it did not have as good power (and also loses something in interpretability and biased estimation of means). Hence while the GLM has a number of advantages, combining it with param bootstrap at small sample sizes might be worth considering to address the Type I error issue."

Response:

Comment 38: *"Supplement 2 is a useful addition, and it might be worth giving it a little more emphasis in the text to ensure ecotoxicologists wanting to try out new analyses give it a look. A couple of suggestions though: - there is no parametric bootstrap code there at the moment. This would be easy enough to do by using the mvabund package, calling manyglm (even on a univariate response) then anova(..., resamp="monte.carlo") will do a parametric bootstrap. - you can construct residual plots for GLMs on this package, using the plot function on a fitted manyglm object."*

Response:

Response to Reviewer 3

Comment 39: *"With its simulations in an ecotox. context, the paper is a fine contribution to make researchers aware of the possibilities and advantages of a generalized linear models (GLM). What is lacking is pointing to the disaster of using GLM in an improper way, notably without adjustment for overdispersion. The paper and abstract can still be made more useful by given not just the recommendation for GLM but also to give advice and warnings of proper usage. For example, (almost) never use loglinear/Poisson regression without adjustment for overdispersion, be aware that a nicer looking model like the negative binomial may give inflated type I error, a case that is overcome in this paper by bootstrapping. The same applies to binomial/logit regression unless the experiment really consist of independent observations for each 0/1 result (and is not just a count with a predetermined maximum). In this light it is strange that the logistic/binomial case is treated differently from the loglinear/Poisson case! Should this not be repaired or at least given more warning.*
(1) Also, if the authors stick to the binomial/logistic without overdispersion, the nor-

mal model could be made powerful in the simulations by treating the error variance as known, because it is a known constant for the arcsine transformation.
(2) Also, the simulation now use slight overdispersion whereas the example /typical data has large overdispersion. Add simulations with large overdispersion. The current ones can go to supplementary.
(3) Without the use of a GLM equivalent of the Williams test all the advantage of the use of GLM in terms of power are gone. See the example. Discuss this ambiguity. You can perhaps use a bootstrap test based on (GLM?) monotonic regression or similar. I know some cues/leads in this direction."

Response:

add warton 2005: almost never use glm without counting for overdispersion...

Comment 40: *"Page 1 Line 22 right column: a continuous proportion is fine for area cover and so on, but not for proportion of surviving animals where is just "k out of n" turn into a proportion. Discrete therefore. "*

Response: Thank you for pointing to this error. Of course, we meant discrete proportions and changed accordingly.

Comment 41: *"Line 58 right add ref to (Warton 2005)."*

Response: This reference fits very well to our paper and we added it.

Comment 42: *"Page 2 Line 40 : , for y¿0 should for y¿= 0 and be on the next line without an indent after formula (if it is not new paragraph)."*

Response: We clarified this equation and moved this part of eqn. 1 to the text: *"The factor A was chosen in such way that Ay equals 2 for the lowest non-zero abundance value (y)."*

Take this paper also into discussion.

Comment 43: *"Line 51 x_i is undefined. The model appear to specify a simple linear regression or a control-one_treatment model. Neither is used in the paper!!! Modify the notation therefore.."*

Response:

Comment 44: *"Line 59 Mention backtransformation here, which works fine for the confidence intervals, and not the backtransformed is the median on the original scale. If you want to backtransform to the original mean, use exp(mean+ 0.5*sigma2).*

See also the emphasized discussion on

(Aitchison & Brown 1969) page 8 (Jongman et al. 1995)page 19. "

Response:

Comment 45: *"Page 3 line 17 n= 4.10 = 40 what does the mean here/ what is the purpose?"*

Response: We are sorry for this error. It was a remnant of a early version of the manuscript - removed.

Comment 46: *"Line 41 : give more attention to the condition for the binomial: independent observations of each of the successes (k) out of the number of experiments (n). And: thus give attention to overdispersion in this case as well, particularly as you found nasty things for the LR of NB."*

Response:

Comment 47: *"Line 28 right. Ref for Dunnett contrasts."*

Response: We agree and added reference to the original description of Dunnett (1955).

Comment 48: *"Line 59 right. Only slight overdispersion, whereas the example data have large overdispersion. Add large overdispersion."*

Response: We clarified this section. We now specify the overdisperion in terms of κ in both, the case study and the simulation design. The simulated data mimicked the case study and such data is common for mesocosm studies in ecotoxicology.

Comment 49: *"Page 4 L 19 (missing."*

Response: We fixed the wrong citation type.

Comment 50: *"Line 56 left to 2 right. Four p-values: what to believe? Your simulations (although with small overdispersion) show it. The LR of NB cannot be trusted. So the 0.016 is out. The remaining tests all show about the same p-value if you interpret statistics properly. Please note that the distinction between significant and non-signicant (here p = 0.061 and 0.042) in itself is statistically insignificant. (Gelman & Stern 2006)."*

Response: We fully agree with the reviewer and cautioned that p-values should

not be overinterpreted by ecotoxicologists (citing Gelman & Stern 2006). Parameter estimates (and their uncertainty) give much more information - on which we also emphasized in Figure 1.

Comment 51: *"Line 4 right : the backtransformed mean gives a median on the original scale, a value that is more interesting than the mean (generally for skew data), and for skew right always smaller than the mean."*

Response:

Comment 52: *"Section 3.2 Always start with the Type I error results because the power results are without meaning if the Type I error is inflated. This also applies the order of the figs: interchange the two rows of subfigs. "*

Response: We also changed the order of the subfigures.

reorder in text

Comment 53: *"21 right: remove GLM_nb in this sentence as GLM_nb must be discarded due to inflated Type I error."*

Response: We agree and changed accordingly.

Comment 54: *"L32 right. ", but this." Although you should reorder the results, I mention that the "but this" should at least be "This could fortunately be . . . "[and we could not have known for sure (apart from the general warnings about LR) without the simulations] This inflated error for GLM_nb must receive more attention. It is an interesting result of this study: you must to bootstrap when using NB."*

Response:

Comment 55: *"Page 5 Fig. 2 legend. Add that n =100 simulations. Add the interpretation that GLM_nb has inflated type I error, and that its line for power is just added for completeness and should not be interpreted as an estimate of the true power."*

Response: We unified the simulations and generated simulated 1000 data sets for each type of scenario (see comment 14). Moreover, we added
"Power levels for models with inflated type I error are shown for completeness."
to figure captions 2 and 3.

Comment 56: *"Page 6 Fig.5 LM for 0.8 has a high Type I error. Does it deviate*

significantly from 0.05? [same question for the high Type I for GLM_nb in fig 2]. Use confidence interval for binomial p = 0.05 n = 100."

Response:

Comment 57: *"Case study : You can check whether the difference between log(2y+1) and log(Ay+1) with A = 0.182 already causes the difference. (it probably does not). The important difference is the use of the Willems test! Without the use of a GLM equivalent of the Williams test all the advantage of the use of GLM in terms of power are gone! See the report of the Williams test below."*

Response:

Comment 58: *"Page 7 L35 Move ref to Warton to after data."*

Response: We changed accordingly.

Comment 59: *"L45 Add a reference to (ter Braak & Šmilauer 2014) who advocate the use of the transformation approach in a multivariate (dimension reduction) context."*

Response:

Comment 60: *"Line 1 right (like GLMs) ->(like GLMs and the Williams test)"*

Response: We did not investigate the differences between multiple comparison procedures (MCP). We used Dunnett constrast for in all simulations, because only these allow individual comparisons between treatments and the control. For a comparison of MCP see Jaki and Hothorn (2013), who recommend a combination of both, Dunnett and Williams constrasts.

Comment 61: *"Line 10-40 Again: discuss Type I error first. Then no mention of GLM_nb in power."*

Response:

Comment 62: *"Line 18 right. Add after "non-normal data' the point the importance of proper accounting for under and overdispersion when using GLM. What is lacking is pointing to the disaster of using GLM in an improper way, notably without adjustment for overdispersion."*

Response: We rerun our simulations and also included the Poisson GLM. The results for Poisson GLM are now also display in Figures 2+3 and we discuss the consequences of failing to account for overdisperion.

changes to text

Comment 63: *"Line 41 right size of 9 ->size of 6-9 (?)"*

Response:

Comment 64: *"Line 44 right the transformation ->the LM"*

Response: We changed accordingly.

References

van den Brink PJ, Hattink J, Brock TCM, Bransen F, 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

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T, Van den Brink PJ (2015) The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. Environmental Science and Pollution Research 22(2):1160–1174

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6(2):65–70

Jaki T, Hothorn LA (2013) Statistical evaluation of toxicological assays: Dunnett or Williams test—take both. Archives of Toxicology 87(11):1901–1910

OECD (2004) Test No. 202: Daphnia sp. Acute Immobilisation Test. Organisation for Economic Co-operation and Development, Paris, URL `http://www.oecd-ilibrary.org/content/book/9789264069947-en`

Szöcs E, Van Den Brink PJ, Lagadic L, Caquet T, Roucaute M, Auber A, Bayona Y, Liess M, Ebke P, Ippolito A, Ter Braak CJ, Brock CM, Schäfer RB (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: A comparison of methods. Ecotoxicology DOI 10.1007/s10646-015-1421-0

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92(1):3–10