FES 524: Natural Resources Data Analysis

Reading 8.2

**Contents**

[Other readings to do before class 1](#_Toc64894626)

[Fitting a Poisson GLM 1](#_Toc64894627)

[Checking for overdispersion 1](#_Toc64894628)

[Residual plots 2](#_Toc64894629)

[Tests and confidence intervals 2](#_Toc64894630)

[Results 2](#_Toc64894631)

[Alternatives to the Poisson distribution 2](#_Toc64894632)

[Zero inflation 3](#_Toc64894633)

[Having lots of zeros isn’t necessarily zero inflation 4](#_Toc64894634)

[Modeling options for zero inflated data 4](#_Toc64894635)

[Hurdle model 5](#_Toc64894636)

[Zero inflated model 5](#_Toc64894637)

[Generalized linear mixed models 6](#_Toc64894638)

[Difficulties with GLMM’s 6](#_Toc64894639)

[Residual plots 6](#_Toc64894640)

[Repeated measures and GLMM’s 8](#_Toc64894641)

# Other readings to do before class

Read the description of the study that assignment 8 is based on in Handout 8.2 and come prepared to discuss aspects of the research question and study design.

# Fitting a Poisson GLM

We’ll start by talking about fitting a Poisson generalized linear model.

## Checking for overdispersion

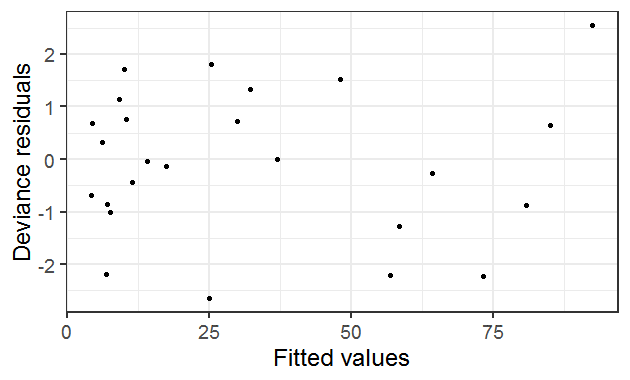
Overdispersion is an especially prevalent issue for one parameter distributions like the Poisson distribution. Much like with the binomial distribution, overdispersion in the Poisson (aka *extra-Poisson variation*) is so common that we might want to simply assume that overdispersion will be an issue and start with a model that addresses the issue. For ecological data it often makes a lot of sense to start with a negative binomial model, since a Poisson model so rarely works. There are other alternative models to a Poisson GLM, as well, which I’ll outline later in this reading.

As with the other generalized linear models we’ve gone over, we check for overdispersion on our fullest model. The estimate of the overall overdispersion is the sum of the squared Pearson residuals divided by the residual degrees of freedom.

If we have overdispersion, we’d report the overdispersion and introduce the alternative model we fit to the data. If we don’t have evidence of overdispersion, report that along with an estimate of overdispersion.

## Residual plots

We use plots of residuals primarily to look for unusual patterns that could indicate a lack of fit. In the plot below from a model fit to the motivating example we don’t see much of interest in terms of patterns.



One thing to note is that the plot shows deviance residuals. We expect ~95% of the points to fall between -2 and 2. The 5 of the 25 total points outside that range in this case is another indication of overdispersion. This matches up with an estimate of overdispersion above 1 when fitting a Poisson GLM to the data from the motivating example.

## Tests and confidence intervals

For any hypothesis tests, a good standard option is to use drop-in-deviance (aka likelihood ratio) tests. These can be done by fitting a full and reduced model and using the anova() function to compare models or with the drop1() or car::Anova() functions. The resulting hypothesis tests are based on the χ2 distribution.

The worst tests to report are the Wald z tests in the summary output. See a more complete discussion of tests for GLM/LMM/GLMM models here, <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#testing-hypotheses>.

For confidence intervals, see the previous discussion in week 7 reading 2 about fitting a binomial GLM.

## Results

We must exponentiate results from a Poisson generalized linear model with a log link in order to interpret them. Estimated differences in means among groups or changes in means across a continuous variable are multiplicative. Unlike a log transformation, however, we still interpret results as *means* and not medians when using a log link.

## Alternatives to the Poisson distribution

There are several alternatives to a Poisson GLM, since such models rarely work well due to the strict variance-mean relationships. These are listed below in no particular order.

Quasipoisson

The quasipoisson model was one of the earliest options we had to deal with overdispersion. This is a simple option, where standard errors are inflated based on the estimated overdispersion. This can be overly simple; see the discussion on the quasibinomial in reading 7.2.

Negative binomial

The negative binomial can be a good option to start with ecological count data; see reading 8.1 for a discussion of this distribution.

What we haven’t discussed yet is that there are alternative parameterizations of the negative binomial, available in R in packages such as glmmTMB. You can read about this in the second half of this discussion on models to account for overdispersion, <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#fitting-models-with-overdispersion>.

Generalized Poisson

The generalized Poisson distribution allows for a different variance-mean relationship than the negative binomial distribution. The generalized Poisson is slightly heavier tailed and so can accommodate more zero counts and more very large counts (see Stroup GLMM book section 11.3.2).

Observation-level random effect

As discussed for the binomial distribution in reading 7.2, we can add an observation-level random effect to deal with overdispersion. This would mean fitting a generalized linear mixed model (GLMM). You may see this referred to as a Poisson-log-normal model.

# Zero inflation

We will take a little time in this last reading to discuss *zero inflation*. If you are in a situation where you have zero inflation you will need a more thorough treatment of the issue than what we are going to do in this class. The discussion here is specifically for discrete count distributions, although much of this can technically apply to the binomial distribution, as well. We do not directly address the issue of many zero values along with otherwise continuous variables. The discussion of hurdle models, however, is relevant to this latter issue (also see reading 7.1).

Zero inflation is when there are more zero values than we would expect under a given distribution. Just having a lot of zero values does not indicate zero inflation by itself, as some distributions can have a lot of zero values.

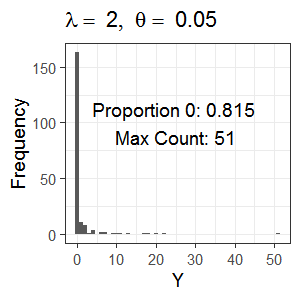
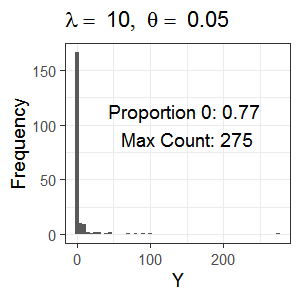
Having more zeros than expected from a distribution will result in overdispersion. This would be overdispersion from a potentially different process than having a missing covariate, which is the most common cause of overdispersion. Instead we would be potentially using the wrong distribution for the model, since the distribution used in the model does not allow for the number of zero values observed in the data. Alternatively, a covariate could explain the excess zero values. In this case, the model with that covariate in it should model the zero values correctly.

If we suspect we have zero inflation, we can compare the number of zeros in the observed data to those expected from a model fit to the data assuming some distribution. You can see examples for checking for excess zero values for negative binomial and Poisson GLM’s here, <https://aosmith.rbind.io/2019/03/06/lots-of-zeros/>.

## Having lots of zeros isn’t necessarily zero inflation

A common mistake is to assume that just because there are a lot of zeros present in the dataset that the data are zero inflated. Zero inflation is *excess* zeros compared to the number in a distribution. Distributions like the negative binomial distribution and the generalized Poisson distribution can have a lot of zeros in them depending on the parameters of the distribution.

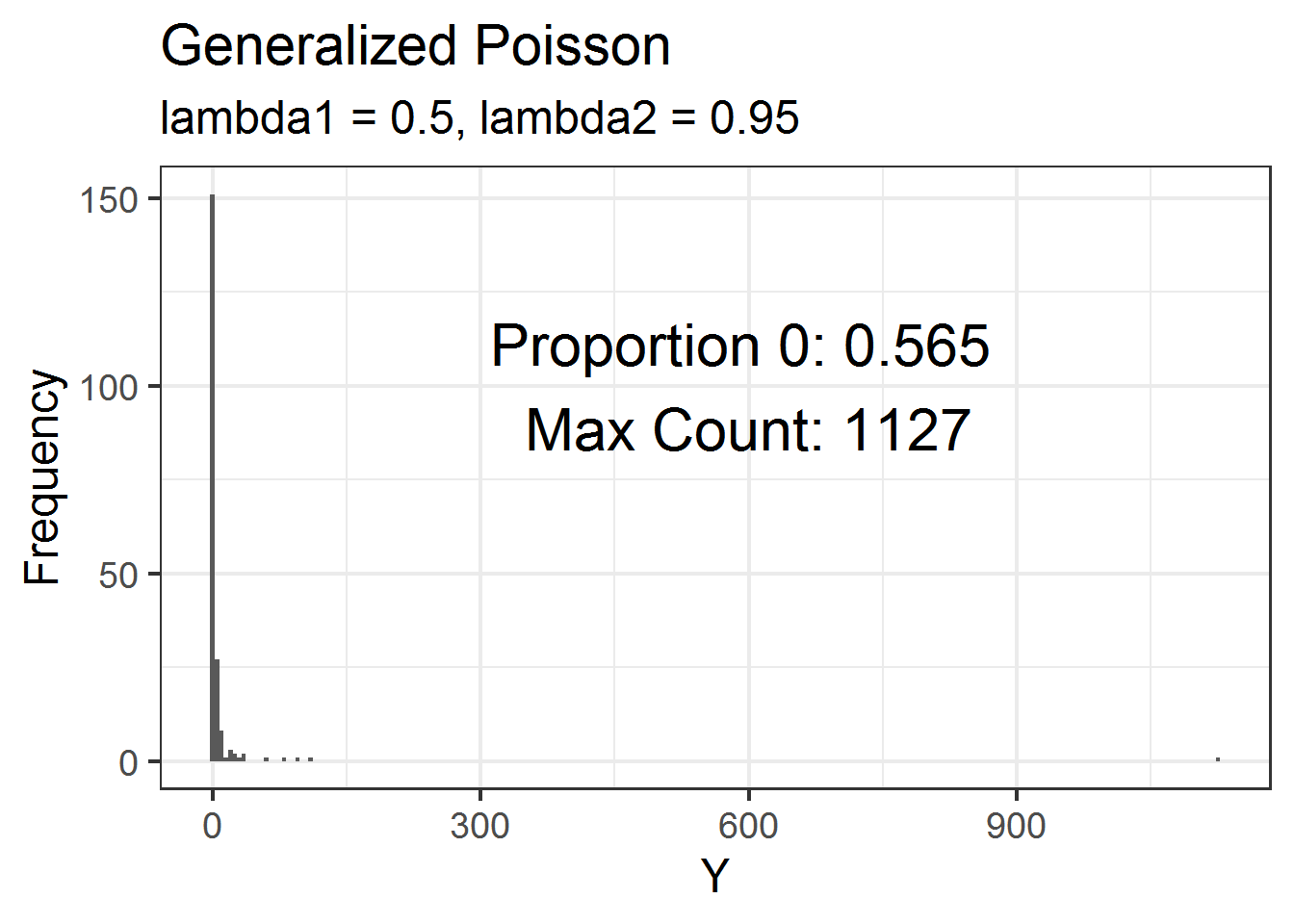
Here are two examples of simulated negative binomial data with many zero values as well as a few very high counts.

Th

The proportion of zero values in the data are printed within the plot. In both cases, more than 75% of the data are zeros. Even with 77% 0 values, the negative binomial distribution can have a count as high as 275.

Next, here’s an example showing data simulated from the generalized Poisson distribution. More than 50% of the values are zero but the maximum count is over 1000.



## Modeling options for zero inflated data

If you are in a situation where you have more zeros than expected, you don’t have a covariate to explain the extra zero values, and other alternative distributions don’t help, you have two main analysis options: a hurdle model or a zero-inflated model.

For both of these models we assume the zero inflation observed in the count data was created through a mixture of a binary, on-off process and a discrete counting distribution. The discrete counting distribution is usually negative binomial or Poisson. In the “off” phase of the binary process, only zero counts are possible. During the “on” phase, nonzero counts may occur. What we assume is happening during the “on” phase differs between the two types of models.

I find understanding how the models differ in the “on” phase to be clearer with examples, which I go through below. These examples are taken from Stroup’s 2013 GLMM book section 11.5 (this book is on your class reading list). See more discussion and examples in Zuur et al. 2009 chapter 11 (also on your class reading list and available as a PDF from the OSU library). I find the Zuur et al. discussion on sources of zeros starting in section 11.3.1 to be useful when thinking about why a dataset may contain many zero values and whether those zero values are “true” zeros.

## Hurdle model

Using a hurdle model makes sense in situations where we only have nonzero counts during the “on” phase.

For example, suppose we are counting the number of times individuals visited the doctor over the past year. There are two types of individuals: those who never visited the doctor over the past year and those who did. For those who did not visit the doctor, the number of visits must be zero. This is the “off” phase, where we always observe a zero count.

For those individuals who did visit the doctor in the past year, the number of visits must be at least one. This is the “on” phase, and in that phase we can only have nonzero counts.

A hurdle model is equivalent to fitting two separate models, one with a binary response variable (e.g., visited or not) and one with the positive counts as the response variable. The model for the positive counts, which contain no zeros, will be based on zero-truncated distributions such as the zero-truncated negative binomial or zero-truncated Poisson. There are packages that can fit both pieces of the hurdle model simultaneously.

Hurdle models can also be useful for positive, continuous data with many zeros as discussed briefly in reading 7.1. Zero inflated beta models are also actually a type of hurdle model.

## Zero inflated model

For zero inflated models, we can have both zero counts or nonzero counts during the “on” phase.

For example, suppose we are counting cars going through a controlled intersection (i.e., one with a traffic light). Assuming people are obeying the law, no cars will go through the intersection when the light is red. This is the “off” phase: when the light is red, we always observe a zero count.

When the light is green, and cars are free to move through the intersection, we can have positive counts. However, if traffic is very sparse, as we might expect during certain times of the day, we may have no cars go through the intersection even though the light was green. During the “on” phase, when the light is green, we can have both zero and nonzero counts.

A zero inflated model is a single model, and you will need a software package to model the two parts of the model (binary process and count distribution) simultaneously. The negative binomial and the Poisson distributions are the most common distributions to use for the count distribution part of the model.

# Generalized linear mixed models

Finally, to end the quarter we’ll go through an even briefer introduction to generalized linear mixed models (GLMM’s).

GLMM’s are when we add random effects to GLM’s. The response variable, conditional on the explanatory variables and random effects, come from non-normal distributions. However, the random effects still come from a normal distribution. There are some models out there that allow for non-normal distributions for random effects but they are still comparatively rare.

A good introduction to GLMM’s with examples is Stroup’s 2015 paper in Agronomy, “Rethinking the analysis of non-normal data in plant and soil science”. See his 2013 book from the class reading list, as well, although it is written more for statisticians-in-training than applied scientists.

Packages for fitting classic GLMM’s in R include lme4 and, newer, glmmTMB and GLMMadaptive. There are also Bayesian-based packages like brms and MCMCglmm.

## Difficulties with GLMM’s

Difficulties can arise when fitting GLMM’s. Some of these are related to general issues in GLM’s, such as having limited information in the response variable to fit many parameters. This issue is even more pronounced in GLMM’s.

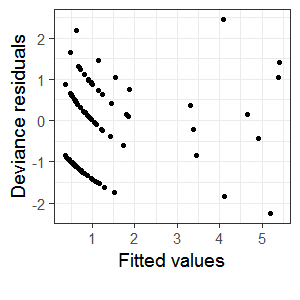
The optimization itself can be difficult for GLMM’s. We often get warnings and messages that take effort to navigate to decide if they are actual problems or ignorable warnings. For example, we often have to standardize continuous variables in order to get models to converge when working with GLMM’s. Then we have an extra step to unstandardized coefficients in order to interpret them.

If working with package lme4 in R, you may find yourself changing optimizers to achieve convergence. Alternatively, you may find you need to fit multiple models, each with a different optimizer, to see if you get similar answers from each and so can conclude any warnings are likely spurious. This is a learning curve that you will likely run into when you start fitting GLMM’s in R.

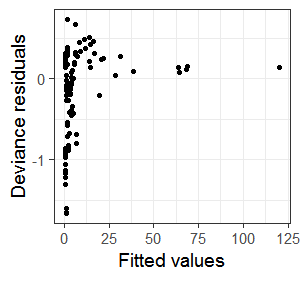
## Residual plots

One really big, practical difficulty in GLMM’s is in interpreting residual plots. Unlike linear mixed models (LMM’s), we don’t know what “OK” residuals plots should look like. This makes it hard to figure out if our model fits our data or if we need to find an alternative.

Here is an example of a residual plot from a Poisson GLMM. The model is fit to simulated data so I know the model fits, but it is difficult to tell that from the plot.



Here is another example, this time from a Poisson-log-normal GLMM (i.e., a Poisson model with an observation-level random effect in it). Again, these data are simulated from the model that I used for the analysis and so I know the model fits fine. Given our understanding of residuals plots from LMM’s, though, we would not think this plot looked OK.

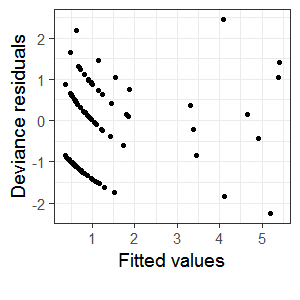
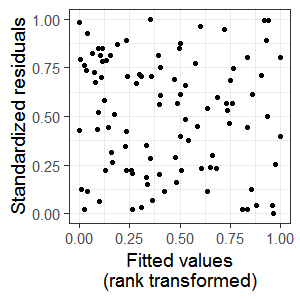


Given this difficulty, we need alternative ways to look at residual plots. The main tool we have to help us with this problem is simulations. For example, if we simulate data from a fitted model and then fit a new model to that data and make residual plots we can see if our actual plots jive with the plots based on simulated data. We know the model fits the simulated data, so any issues in the residual plots there would be spurious. This is a nice idea, but still leaves us with a problem: how can we simulate many plots and see if our original plot falls in the distribution of possible “good” plots in some efficient way?

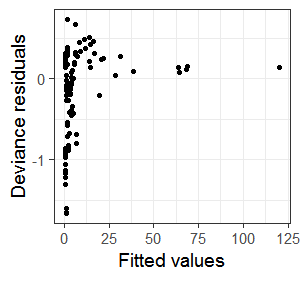
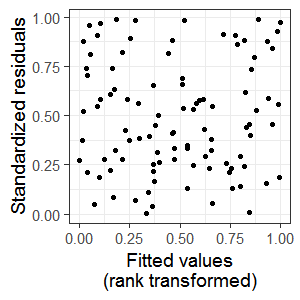
The answer comes with package DHARMa. This package does simulation-based residual checks for GLMM’s. While the package has been adding more and more hypothesis tests for checking assumptions, as well, you know how and why these fail. I recommend focusing only on the plots you make with the simulated residuals from DHARMa.

If you are fitting GLMM’s in R you should use DHARMa for making residual plots. Make sure you read the DHARMa vignette, <https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html>, before starting to use it. The vignette is filled with good information on why simulated residuals are needed and exactly what the package does. It also gives examples of problematic simulated residual plots vs OK simulated residual plots, which of course is always a difficult call to make (especially for small sample sizes).

Below is an example of the original Poisson GLMM residual plot that looked problematic and the DHARMa residual plot based on simulated residuals that shows the model fits fine.

Here is the same thing for the Poisson-log-normal GLMM example.

## Repeated measures and GLMM’s

If you have repeated measures on top of a GLMM, you currently have relatively few options for fitting models in R. Such models are extremely complicated and can be difficult to fit.

Package glmmTMB has a few basic correlation structures available. Package spaMM is relatively new but allows for some complex spatio-temporal correlation structures. If your situation is really complex, see package INLA. This package is pretty different than others to use but is apparently very powerful.

There is a lot to learn when using GLMM’s. You will need resources to help you when you hit snags. One good resource is the GLMM FAQ that Ben Bolker maintains, <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>. It has a lot of information on many different GLMM topics that can come up in a real analysis, and is worth bookmarking and revisiting often.