FES 524: Natural Resources Data Analysis

Reading 8.1

**Contents**

[Other readings to do before class 1](#_Toc64530521)

[Generalized linear models for counts 1](#_Toc64530522)

[Negative binomial distribution 2](#_Toc64530523)

[Poisson distribution 2](#_Toc64530524)

[Probability model 2](#_Toc64530525)

[Link function 3](#_Toc64530526)

[Zeros and complete separation 3](#_Toc64530527)

[Count per unit effort 4](#_Toc64530528)

[Fitting a negative binomial GLM 5](#_Toc64530529)

[Check for overdispersion 5](#_Toc64530530)

[Residual plots 5](#_Toc64530531)

[Tests and confidence intervals 6](#_Toc64530532)

[Results 6](#_Toc64530533)

[Practically important change in a continuous explanatory variable 6](#_Toc64530534)

[Plotting results with multiple continuous explanatory variables 7](#_Toc64530535)

[Output from R 8](#_Toc64530536)

# Other readings to do before class

Read Week 8 Handout 1: “Study on colonizing insects and mangrove island size” and come prepared to use this example to talk about the topics covered in reading 8.1. Make sure you’ve had time to think about the answers to the questions in that handout.

# Generalized linear models for counts

We’ll start today by reviewing the count distributions we discussed in week 7 reading 1.

Counts are discrete variables that can include 0. Counts are always integers.

**Examples**:

# seedlings per plot

# inflorescences per plant

It is common to have counts where the effort to take the counts varies between observations. This effort could be the amount of time spent counting or the area counted in, for example. Any analysis should take into account varying effort in some way, often by making inference to a count per unit effort (e.g., count per m2).

**Examples effort**:

Plot size

Amount plant cover present

The most commonly used distributions for analyzing counts in natural resources are the negative binomial and the Poisson.

## Negative binomial distribution

The negative binomial distribution is a two parameter distribution. The variance is related to the mean. In the mean-variance relationship shown below, you can see that we expect the variance to increase *quadratically* with the mean.

**Distribution Mean Variance**

Negative binomial λ λ + λ2/θ

The negative binomial often works well for modeling counts in ecology. This is because the distribution tends to work well for “clustered” counts, where the presence of an organism is related to the presence of other organisms. Such data often have many zero or low counts as well as some very high counts.

The example I gave last week of a situation with clustered counts involves having very high counts of seedlings in areas with many parent trees but having no or few seedlings when there are few parent trees. In this example there are either a lot of seedlings together or there are very few seedlings present.

Since the the variance depends on the mean for this distribution we can technically have problems where the variance in the data is larger or smaller than the variance based on the distribution. However, given this distribution has a scale parameter, θ, overdispersion like in single parameter distributions isn’t possible. Any evidence of overdispersion (i.e., value of overdispersion > 1) will indicate a general lack of fit of the model to the data. We will discuss this a bit more later in this reading.

## Poisson distribution

The Poisson distribution is a one parameter distribution. If we know the mean, we know the variance. In the mean-variance relationship shown below, you can see that we expect the variance to increase *linearly* with the mean.

**Distribution Mean Variance**

Poisson λ λ

Overdispersion, aka *extra-Poisson variation*, is very common in the Poisson distribution. Overdispersion is always common for one parameter distributions.

The Poisson distribution is most often useful to describe count distributions with limited ranges. The distribution can be used for situations with overall very high counts or overall very low counts but will not work well when there are both very low and very high counts. For this reason, the Poisson distribution is rarely useful for ecological count data. As I mentioned last week, I have seen the Poisson distribution work for species richness data when the number of species possible was limited (e.g., 0-5).

The negative binomial and Poisson distributions are related to each other. It turns out the Poisson distribution is a special case of the negative binomial distribution. We will work through a basic proof of this in class.

## Probability model

Let’s write out the probability model for the counts.

The distribution of the observations, conditional on the variables in the model, will depend on which distribution we are working with.

Yt | Xt ~ Negative Binomial(λt, θ) or

Yt | Xt ~ Poisson(λt)

The mean of the distribution is related to the linear predictor via some link function. Here is the equation using a generic link function *g*:

g(λt) = β0 + β1Xt

## Link function

The canonical link for count distributions is the natural logarithm. This is true for both the negative binomial and the Poisson distributions.

The equation to relate the mean to the linear predictor for count data using the canonical link:

log(λt) = β0 + β1Xt

Using the log link leads to two scales: the model scale and the data scale. The *model scale* is on the scale of the log mean count. The model scale is the scale that we do hypothesis tests and calculate confidence interval limits.

The *data scale* is on the scale of mean counts, calculated by using the inverse link. The inverse link of the natural log link is exponentiation. The data scale is what we use for interpretation of model results. All estimated relationships are multiplicative changes in *mean* counts. Note that this is a different interpretation then when we use a log *transformation* and fit a linear model.

## Zeros and complete separation

We discussed the issue of zeros and the log transformation back in week 4. Count distributions can, and often do, contain 0 values. Why aren’t zero values a problem for GLM’s for counts when using the log link?

To understand why this isn’t a problem we have to be sure we understand what a link is. We are not transforming the response variable when we use a link. Instead we are linking the *mean* of the distribution to the linear predictor. This is easiest to see by writing out the mathematical notation.

When we log transform the response, it is log(Y) that is the response variable. If Y contains 0 we will have problems:

log(Yt) = β0 + β1Xt + εt

But we don’t use transformation when fitting a GLM. Instead we relate the log of mean Y (i.e., λ) to the linear predictor. We won’t have problems if Y contains 0 because we don’t take the log of Y:

log(λt) = β0 + β1Xt

Issues can arise with 0 values and counts, though, due to complete separation. Complete separation is not something that is commonly discussed for GLM’s for counts. However, remember the definition of complete separation we learned with binomial GLM’s: when an outcome is perfectly determined by the explanatory variable.

One way for the outcome to be perfectly determined for counts is if all the observations in a group are 0. Much like we saw with the binomial GLM and the logit link, if one group has all 0 counts then the *mean* of that group is 0. The model tries to estimate log(0) and so runs into problems. Take another look at the probability model for counts written above and make sure you can see why problems will arise when the mean is 0.

GLM’s with a log link fit to count data where a group has all 0 values will show large standard errors and/or warnings and errors. In R, GLMM’s will likely give errors but GLM’s may only have large standard errors in the output.

You should have a good idea if you are going to have complete separation after exploring your data, since one group or combination of groups will be all 0 values. Seeing that will give you a chance to think about the problem and how you want to approach it. The issues and options are similar as those discussed in week 7 reading 2. However, penalized options like Firth’s regression are not common for Poisson and may not exist at all for the negative binomial distribution. It is possible to use a Bayesian approach for penalization based on priors, though, and this is a good option.

## Count per unit effort

Everything covered so far is specific to counts, where the response is a discrete integer and the negative binomial and Poisson distributions are viable options. We have yet to address how to address count per unit effort data.

**Effort doesn’t vary**

In some cases, the actual effort didn’t vary. For example, if seedlings were counted in plots that were all the same size then the effort was all the same. What I’ve seen, though, is that in some fields it is standard to convert the observed counts from a small area to counts per area (like trees per acre or trees per hectare). This is done by extrapolating the observed counts per small area to counts for some larger area through multiplication. These counts per area, of course, are often not integers. In addition, the resulting distribution of values usually has many 0 values along with some extremely large values. Modeling the extrapolated data then becomes extremely difficult, even if the researcher forces things to be integers via rounding (which isn’t a great idea, anyway).

If the effort doesn’t vary, don’t extrapolate the observed values to larger areas prior to analysis. Analyze the observed counts with a count distribution. Then convert the results from the model (i.e., mean counts) to mean counts per area. For example, convert the estimated differences in mean number of trees to mean number of trees per hectare. That way you can take advantage of count distributions for analysis but present the results on the expected scale.

**Effort varies**

There are cases, though, where the effort expended to get the counts does vary.

Some more specific examples of varying effort:

Time spent counting (e.g., net hours when running a net line)

Plot size when counting number of individuals

Amount cover of the species present when counting inflorescences

In such cases, investigators are often tempted to make a *count per unit effort* variable by dividing the count by the effort; they calculate a *density* or *rate*. Since this variable is now continuous with a minimum at 0, the investigator attempts to analyze these data assuming normality or transforming to assume normality.

However, densities share a lot of the same features with count distributions, including variance heterogeneity and the potential presence of many 0 values. Trying to use a linear model, assuming constant variance and normality of errors, can therefore fail for similar reasons to why such models fail for counts.

Luckily, there is an option that can allow us to continue to work with count distributions and GLM’s but make inference to densities. This option is called using effort as an *offset*. Note that using an offset is specific to models that use the log link. It is most common for analyses of counts. When using an offset, the effort variable must be also put on the log scale. We will work through the math that demonstrates why we can do this and why effort is on the log scale in class together.

You can see more discussion of offsets here, <https://stats.stackexchange.com/questions/11182/when-to-use-an-offset-in-a-poisson-regression>, and here, <https://stats.stackexchange.com/questions/66791/where-does-the-offset-go-in-poisson-negative-binomial-regression>.

# Fitting a negative binomial GLM

The process of fitting a negative binomial generalized linear model and checking model fit is similar to what we did with the binomial GLM last week.

## Check for overdispersion

Unlike with single parameter distributions such as the binomial distribution, overdispersion isn’t common in the negative binomial distribution since it has an extra scale parameter, θ. It is possible to have what looks like overdispersion, though, where the mean-variance relationship defined be the negative binomial does not correctly estimate the variance of the observed data. For example, we could see apparent overdispersion if we have more 0 values than can be modeled by the negative binomial distribution. This is a type of *zero inflation*, and we could switch to using model to account for the zero inflation. Note the negative binomial distribution can have a lot of zeros and not be zero inflated; we’ll discuss this and zero inflation more in week 8 reading 2.

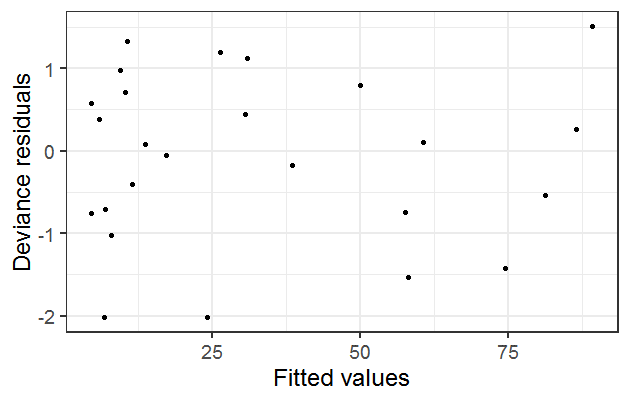
Once a model is fit, we check for problems with the mean-variance relationship of the negative binomial using the sum of the squared Pearson residuals divided by the residual degrees of freedom. Getting a value >1 would indicate that the negative binomial model doesn’t fit correctly. In this case we likely would need to find an alternative distribution.

Since apparent overdispersion can be caused by missing variable, make sure to collect and include covariates that might cause variation in the response variable to the model. Check for overdispersion on a full model.

## Residual plots

Residuals vs fitted and residual vs each variable plots are still used to check model fit even though we no longer assume a distribution for the errors. We are primarily looking for unusual patterns in the deviance residuals that would indicate the model does not fit well based on the distribution we’ve assumed.

Here is a plot based on the deviance residuals from an analysis of the motivating example data. There is nothing interesting to note here.



I have seen odd patterns in residual plots from a negative binomial GLMM that I eventually decided indicated a lack of fit of the model to the data. In that case I saw clear decreasing variation in the residuals with increasing fitted values, the reverse of the horn shape that we saw in week 5 when we did a log transformation. In my case, I ended up fitting an alternative model that fit the data better.

## 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. The resulting hypothesis tests are based on the χ2 distribution. Again, using the anova() function is not “doing an ANOVA” when working with GLM’s so be sure never to use such terminology when reporting results.

For many types of GLM’s the drop1() function or car::Anova() can be useful for these marginal tests. However, these options do not work correctly for negative binomial models fit in R. The primary option for negative binomial GLM’s is to fit two models, full and reduced, and use anova().

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 binomial GLM.

## Results

We must exponentiate results from a negative binomial generalized linear model with a log link in order to interpret them. Estimated differences in means among in 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.

## Practically important change in a continuous explanatory variable

Throughout examples 1 through 7 this quarter, research questions all involved interest in differences in groups. This week, in example 8, the research question finally involves a continuous explanatory variable. This brings up some new issues we need to think about.

First, when we talk about some estimated change in mean Y, we need to know the change in X it is associated with. We will want to report the estimated change in mean Y for a *meaningful* change in X. As you know, by default software will report results for a *1-unit* change in X. However, a 1-unit change is often uninteresting or even misleading, depending on the scale of the X variable.

For example, picture a variable measured in the 1000’s, like area of a watershed in square meters. Is a 1-unit change in watershed area (e.g., 1190 to 1191 m2) even measurable? Would we want to know how something is estimated to change, on average, for an increase in watershed area that is that small?

Now picture something measured between 0 and 1, like the gradient of a stream. A 1-unit change would be a change across the entire possible range of gradient. What if you measured gradient only from 0 to 0.25? Does talking about the estimated change in mean response for a 1-unit change in gradient make sense?

One more step you’ll need to do when working with continuous explanatory variables is to decide what a meaningful change in your explanatory continuous variables is. When I’ve been situations where I had many continuous variables with many different scales, I’ve included a column to show what change in each X was defined as practically interesting in the table of coefficients so the reader could more easily interpret the estimated coefficients.

Getting a coefficient for a practically meaningful change in an X involves multiplication on the *model scale*:

One-unit change in X: 1

Five-unit change in X: 1\*5

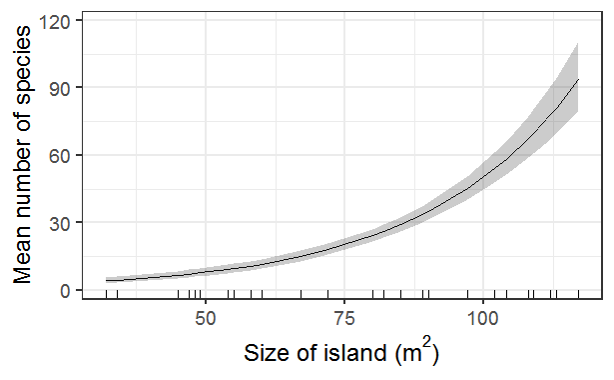
Note that this multiplication must be done on the model scale. If using a log link, do the multiplication prior to exponentiation. The same process can be done for confidence interval limits. Again, do this on the model scale and then exponentiate the results to the data scale.

## Plotting results with multiple continuous explanatory variables

We discussed how to interpret results from models with multiple, continuous explanatory variables in week 5. This is relevant for the analysis for the motivating example this week.

We can use an added variable plot to show the estimated relationship of the variable of interest to the mean response on the original scale with the other variable(s) held fixed. It is most common to hold all other variables to their means and medians; I used median in the plot below. You will need to report the value you hold other variables to in the plot caption. You may also want to report this elsewhere in the text.

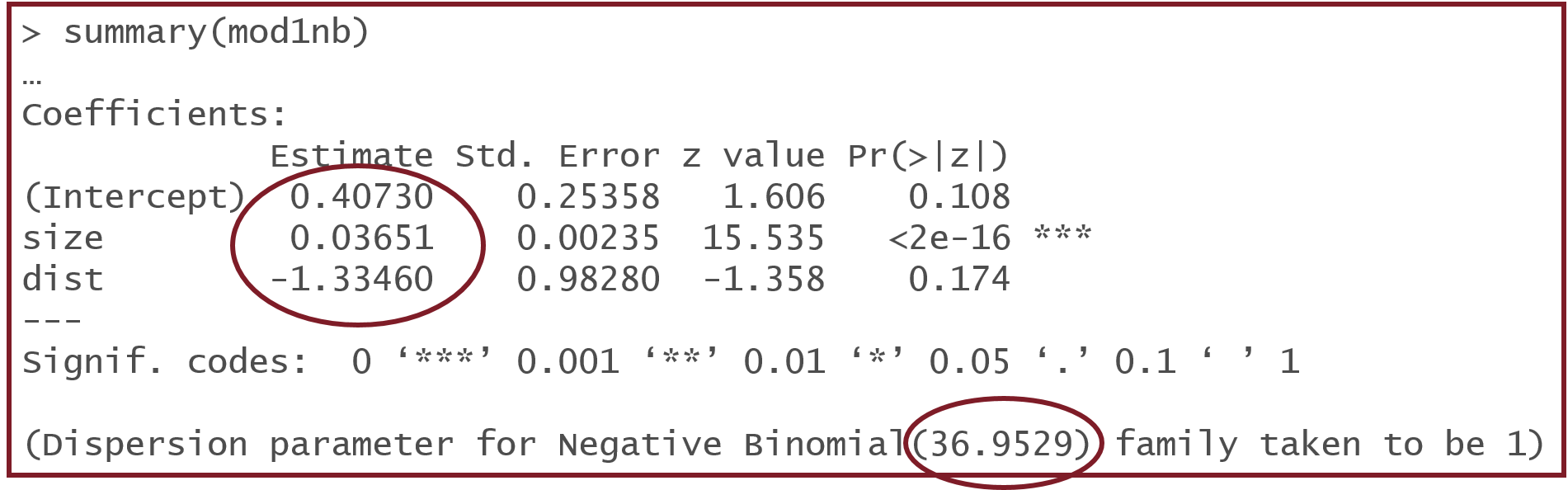
Here is an added variable plot I made based on the motivating example. As discussed in week 5, I used a rug plot instead of plotting the estimated line on top of the raw data. Review reasons for this and ask questions about it, as needed.



In the caption I included information about the value distance to shore, the second variable in the model, was fixed to. Excerpt: “…when distance from shore is fixed at the median value (0.104 km).”

# Output from R

For a model like the one I fit to the motivating example this week, the summary of the model shows the estimated intercept and slopes.



The results above are on the model scale (log mean count). To make inference to mean counts on the original scale I would to exponentiate the estimated slopes shown in “Estimates.

The estimate of the negative binomial θ parameter is given at the bottom of the output. A very large value of θ (in the thousands or larger) might indicate the model doesn’t fit particularly well.

The drop-in-deviance/likelihood ratio tests can be done by fitting a full (all variables) and a reduced (with the variable you are testing removed) model and comparing them with the anova() function. The result is a χ2 test with some degrees of freedom. In this example, the degrees of freedom (shown in the df column) is 1. If relevant, you would report this as χ21 = 63.3, p < 0.0001.

