FES 524: Natural Resources Data Analysis

Reading 8.1: Generalized linear models for count data

Contents

1 Generalized linear models for counts

In this reading, we start by reviewing the count distributions we discussed in week 7 reading 1.

Counts are discrete, integer values that can include 0. For example, we could count the number of seedlings in a plot or the number of inflorescences on a plant. In practice, 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 square meter).

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

Table 1: Mean-variance relationships for distributions commonly used to model counts.

Distribution	**Mean**	**Variance**
Negative binomial	μ	$\mu + \mu^2/\psi$
Poisson	μ	μ

1.1 Negative binomial distribution

The negative binomial distribution is a two parameter distribution, but unlike the normal distribution, the variance is related to the mean. In the mean-variance relationship shown in Table 1, you can see that we expect the variance to increase quadratically with the mean.

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. This results in high *dispersion* that distributions like the Poisson may not account for.

The example in reading 7.1 was a situation with clustered counts with 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, ψ, 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.

1.2 Poisson distribution

The Poisson distribution is a one parameter distribution. If we know the mean, we know the variance. In Table 1, you can see that we expect the variance to increase linearly with the mean since the variance is *equal* to the mean. Overdispersion, or *extra-Poisson variation*, is very common when using Poisson models in practice since it is another one parameter distribution (as was the binomial distribution).

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.

1.3 Probability model

The systematic component of the probability model for count data will look similar for both types of error distributions, and should look familiar from last week:

$$g(\mu_i) = \beta_0 + \beta_1 x_i$$

where x_i is the i^{th} measurement for a single covariate, x, β_0 is the intercept (on the link scale), and β_1 is the slope parameter (on the link scale).

Now, all we need to do is specify the stochastic component of the model, which will read

$$y_i|\mu_i \sim \text{NegBinom}(\mu_i, \psi)$$

for the negative binomial model, and

$$y_i|\mu_i \sim \text{Poisson}(\mu_i)$$

for the Poisson model. Notice that the negative binomial specification includes a second parameter while the Poisson is a single-parameter distribution.

1.4 Link function

The canonical link for count distributions is the natural logarithm, $g(\mu_i) = \ln(\mu_i)$. This is true for both the negative binomial and the Poisson distributions.

Given this information, we can write the systematic component of the model specifically as

$$\ln(\mu_i) = \beta_0 + \beta_1 x_i.$$

Using the log link leads to two scales: the link scale and the data scale. The link scale is on the scale of the log mean count. As with all GLMs, the link scale is the scale on which 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 than when we use a log transformation and fit a linear model.** That is because the link function links the *mean* to the linear predictor.

2 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 GLMs 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 $\log(y_i)$ that is the response variable.

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

If one or more of the y_i's is 0, we will have problems. But, with a GLM, we don't actually transform the data. Instead, we relate the log of the mean of Y_i (i.e., μ_i) to the linear predictor. We won't have problems if some of the y_i's are 0 because we don't take the log of the data:

$$\log(\mu_i) = \beta_0 + \beta_1 x_i.$$

Issues can still arise with 0 values and counts, though, due to complete separation. Complete separation is not something that is commonly discussed for GLMs for counts; however, remember the definition of complete separation we learned with binomial GLMs:

> When an outcome is perfectly determined by the explanatory variable, we have complete separation.

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 estimate of the *mean* is 0.

GLMs 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, GLMMs will likely give errors but GLMs 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 to those discussed in reading 7.2; but, 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.

3 Counts 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 counts per unit effort data. For example, we might be interested in modeling the mean counts of bees caught in hand-nets, but commonly get stormed off the mountain, cutting the time spent netting short. Clearly, the more time one spends netting, the more bees one will catch, so we need some way of accounting for different efforts used to generate each count.

3.1 When effort doesn't vary

In some cases, the actual effort might not vary. For example, if seedlings were counted in plots that were all the same size, then the effort was all the same. What you may see, 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 may not be a great idea).

If the effort doesn't vary, I recommend you 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, extrapolating both the means and compatibility interval limits.

3.2 When effort varies

There are cases, though, where the effort expended to get the counts does vary. For example, time spent netting bees, or number of plants counted 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 and 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.

4 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.

4.1 Check for overdispersion

Unlike with single parameter distributions such as the binomial distribution, overdispersion isn't common with the negative binomial distribution since it has a scale parameter. 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 a model to account for the zero inflation. Note the negative binomial distribution can have a lot of zeros and not be zero inflated. We will discuss this and zero inflation more in reading 8.2.

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 variables, 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.

4.2 Residual plots

Residuals versus fitted and residual versus explanatory 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 or the quantile residuals that would indicate the model does not fit well based on the distribution we assumed.

We will practice looking at deviance and quantile residuals plots in class.

4.3 Tests and confidence intervals

For any hypothesis tests, a good standard option is to use drop in deviance (i.e., 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 χ^2 distribution. Again, using the `anova()` function is not "doing an ANOVA" when working with GLM's so be sure not to use such terminology when reporting results.

For many types of GLMs, 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 GLMs is to fit two models, full and reduced, and use `anova()`.

The "last resort" 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 and using profile likelihood confidence intervals.

4.4 Reporting results

Recall that the default results will be on the link scale, which is the log scale, so we must exponentiate results from a 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.

4.4.1 Practically important change in a continuous explanatory variable

Throughout 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 the mean of 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 the area of a watershed in square meters. Is a 1-unit change in watershed area (e.g., 1190 to 1191 m^2) 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 will need to do when working with continuous explanatory variables is to decide what a meaningful change in your explanatory continuous variables is. One thing that can be helpful for when you have many continuous variables with many different scales is to include a column to show what change in each X_j was defined as practically interesting in results tables 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: $\hat{\beta}$
- Five-unit change in X: $\hat{\beta} \times 5$

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.

4.5 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. Figure 1 shows the estimated relationship between the mean number of species and the size of an island with all other variables in the model fixed at their medians. 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.

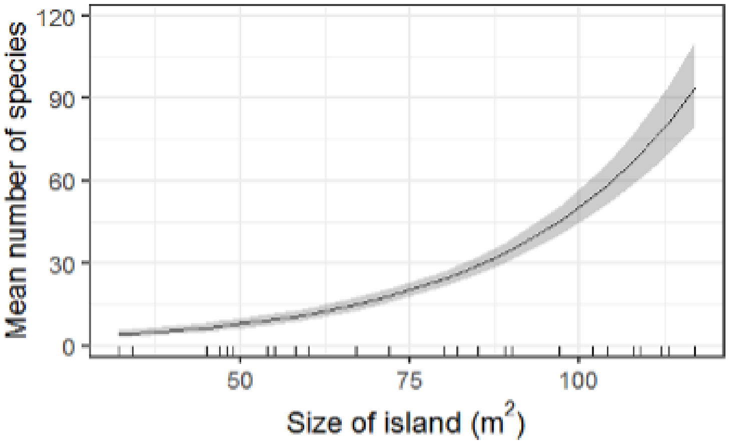

Figure 1: Example added variable plot. As discussed in the week 5 readings, I used a rug plot instead of plotting the estimated line on top of the raw data. Review reasons for this and ask questions about it, as needed.