# FES 524: Natural Resources Data Analysis

Reading 7.2: Binomial GLMs

## Contents



Here, we will continue to learn about binomial generalized linear models (GLMs) and how to fit them. We will take time to talk about additional complications that can arise with these models. We will root the discussion in an example that goes with this week's example lab data.

## 1 Example data

The example data come from a fictitious experiment in which researchers are interested in the effects of agricultural runoff into streams, particularly the effects of runoff contaminated with aflotoxin, a carcinogenic compound produced by certain fungi that decompose agricultural crops. They are interested in the rate at which fish may develop cancerous tumors due to runoff with different concentrations of aflotoxin. They decide *a priori* that a ten-fold increase in the odds of a fish developing a cancerous tumor relative to the baseline rate would be substantially detrimental to the population.

To explore this, they establish 28 fish tanks and randomly assign each to one of four doses of aflotoxin such that there are 7 tanks that receive each dose. Each tank has between 20 and 40 fish (fingerlings, or juveniles) in it depending on the size of the tank such that the density of fish is the same. The response variable is the proportion of fish that develop cancerous tumors in each tank. Part-way through the experiment, it is discovered that 3 tanks were not cleaned before the experiment, so they are removed from the analysis.

# 2 Overdispersion

Overdispersion is the term that we use when the variance in the observed data is greater than the variance we expect from the distribution we are using to model the data. In other words, our model to not flexible enough to account for the observed variation in the data. Single parameter distributions are especially prone to overdispersion since, once we know the mean, we also know the variance. However, any distribution where the variance is based on the mean (i.e., there is a mean-variance relationship) can have issues where the variance defined by the distribution doesn't capture the variance in the data.

Since the binomial distribution is a single-parameter distribution, the variance is completely defined by the mean without any other *scale* parameters (those that define the scale of variation). The mean of a binomially-distributed random variable $Y$ is, $\mathbb{E}(Y) = np$ and the variance is $\text{Var}(Y) = np(1-p)$, where $n$ is the number of trials and $p$ is the probability of success. When discussing overdispersion and the binomial distribution, you may also see the term *extra-binomial variation* used.

Why do we care about overdispersion? Well, our model results are based on the distribution-defined variance. If the variance from the distribution is smaller than the variance observed in the data, our results will be based on an underestimate of the variance. Underestimating the variance means all standard errors are too small, which means any p-values are too small and the width of the confidence interval is underestimated. Our results will be *anticonservative.*

## 2.1 Causes of overdispersion

Recall that, when we defined the binomial distribution in class, we said it tracks the number of successes in $n$ *independent* Bernoulli trials. A positive correlation among the trials within a study unit is one primary cause of overdispersion in binomial data. For example, consider the motivating example of the number of insects killed by varying pesticide concentrations out of a total of 20 insects per concentration treatment. If it were true that, once once insect dies, the others become even less likely to survive (for example, if the dead turn into zombies and attack the living bugs?), then the individual Bernoulli trials for each insect would not be independent. This can cause overdispersion.

Another potential cause of overdipsersion is "a missing explanatory variable", although this essentially induces positive correlation among trials as discussed above. For example, if we used blocking in our design but leave blocks out of the model, we could see overdispersion. Again, this would be due to trials no longer being conditionally independent because, as you know, if blocking is successful then blocks will cause positive correlation among observations with blocks.

Finally, having excessive 0 values that aren't modeled well by the binomial distribution can lead to overdispersion.

## 2.2 Accounting for overdispersion

There are several alternative approaches we can use if we have overdispersion in a binomial GLM, since having overdispersion means our model does not fit the data well. The list below is not in any particular order, so the first option is not necessarily a better or more useful option than the next.

### 2.2.1 Quasibinomial

When using a quasibinomial model, we estimate the overdispersion and simply multiply all standard errors by the square root of this value. That is, we scale up the standard errors as if we had a separate dispersion parameter for the binomial family. The quasibinomial model is a very simple approach to dealing with overdispersion, which is why this is what we will be using in the analysis example this week. However, since we use a single, overall correction based on a single estimate of overdispersion, this approach may not be

flexible enough for a many situations. For example, maybe we have different amounts of overdispersion in different groups, which we can't model with the quasibinomial distribution.

One complication with this approach is that we can no longer use subsequent statistics that depend on the likelihood, such as AIC or likelihood ratio tests. This is because we have adjusted the binomial model in a post-hoc sort of way. This means we no longer have a true likelihood. The quasibinomial distribution is not actually a distribution (it is a "quasi" distribution) so it has no likelihood associated with it. Because of this, the quasibinomial option is not available in some R packages when fitting GLMMs.

One difference between the binomial model and the quasibinomial model is that we use $F$ and $t$ tests when reporting results from a model fit using a quasilikelihood approach instead of the $\chi^2$ and $t$ tests that we use with a binomial GLM. There is actually no theoretical reason for this, but it is considered to be a logical thing to do and is standard.

#### 2.2.2 Beta-binomial model

In the beta-binomial model, we model the mean of the binomial distribution with the beta distribution and then the observed data with a binomial distribution. This sort of model allows for more variation than the binomial GLM does. It also allows us to use covariates in both the parts of the model, as needed. The beta-binomial model is a type of *compound* or *mixture model*, which can be fit in R using packages `VGAM` or `glmmTMB`.

#### 2.2.3 Observation-level random effect

Since we are missing the observation-level random effect in our binomial GLM (i.e., the error term), we could add this to account for observation-level noise. This would mean fitting a generalized linear mixed model (GLMM).

None of the three listed options are useful if excessive 0 values are causing the overdispersion. In that case, other tools like *zero-inflated models* will likely be needed.

### 2.3 Underdispersion

Underdispersion, where the variation in the data is smaller than what we expect given the distribution, is also a possibility. For example, we can get underdispersion if we had negative correlation of trials within any study units. Historically, underdispersion has usually been ignored and there aren't modeling approaches that have been designed to deal with it. Certainly, underdispersion indicates a lack of fit of the model to the data, but it is "safer" to ignore it because you are being overly conservative and understating the results. P-values are too large and confidence intervals will be too wide.

### 2.4 Binary data and overdispersion

Truly binary data cannot be overdispersed. However, any variables known to cause variation in the response still need to be in model (fixed effects, blocks, etc.). Switching to working with binary data, such as working on the fish level instead of tank level in the example data this week, is not a way around the issue of overdispersion. The model still won't fit the data well even if you can no longer calculate the overdispersion, and an alternative model would still be needed.

## 3 Complete separation

*Complete separation*, also called *perfect separation*, is when the outcome is perfectly determined by an explanatory variable. Complete separation is a concept that is most often associated with binomial or

binary data since it is commonly a problem with models using those distributions, but, it can occur in other situations as well.

In this week's example data, the control group demonstrates complete separation for a categorical response variable. The outcome is perfectly determined if we know a tank is in the control group because no fingerlings in that group developed tumors. The estimated probability of developing a tumor for the control group is zero.

Complete separation can happen with continuous variables as well. Figure 1 is a plot showing a binary response variable versus some continuous variable, "distance." You can see that distance perfectly explains presence/absence in this case: all observations below 2 are 0 and all observations above 2 are 1. These data are completely separated at a distance of 2.

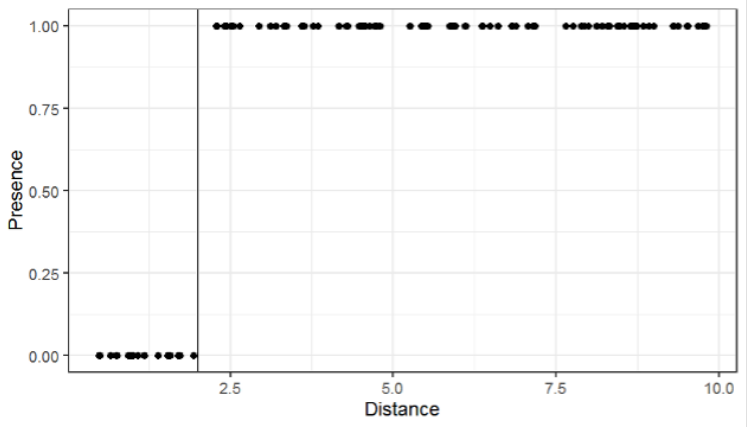


Figure 1: Example of complete separation with a continuous variable.

This can cause issues with the numerical fitting routines that computers use. The "slope" coefficient associated with distance is mathematically defined, but the maximum likelihood estimate is $+\infty$. This is obviously problematic for a computer that can't understand infinity, much less carry out computations up to infinity. A symptom of complete separation will be *extremely* large standard errors along with estimated values of coefficients greater than 4, which on the odds scale are multiplicative effects of $> e^4 \approx 55$. You may also see warning messages about

```
fitted probabilities numerically equal to 0 or 1
```

A group of all 0 values is likely going to be extremely interesting scientifically. What we need to ask ourselves is if it interesting statistically. When a group is all 0, do we need to estimate differences from that group? This may be a case where the option of simply stating this very strong result is sufficient and statistical results won't strengthen the evidence at all.

Having complete separation isn't necessarily a problem. Tests based on the likelihood are still valid (these are discussed more below). Estimates and CI's made from profiling the likelihood will be fine. Wald-based tests and confidence intervals, though, such as those used by package `emmeans` or output in the model summary in R, will not be valid.

### 3.0.1 Alternatives when there is complete separation

Like so many decisions in statistics, if you are in a situation where complete separation needs to be addressed there are many options and the way you should approach modeling will take careful thought. Any choice needs to be justified and well described in your methods. While some solutions are fairly ad-hoc, penalized approaches are generally a good option. There are now Bayesian-based penalized approaches for more complicated models like GLMMs that keep models constrained while working with the original data. Note that doing nothing can also be a fine option in many cases.

You can see a nice discussion and list of options to deal with complete separation on Cross Validated here, https://stats.stackexchange.com/a/68917/29350. The last option in this list is by far the least attractive option.

# 4 Fitting a binomial GLM

Now that we have covered the nuances that go with binomial GLMss, letss go through the process of how we would fit a binomial generalized linear model with a logit link and check that the model fits. This is what we will practice in lab.

## 4.1 Response variable

When fitting a binomial GLM, the response variable is not just the observed proportion of successes for a study unit. You must also have information the number of trials. This is because the number of trials is directly relevant to the variance and because the number of trials in a binomial GLM is a type of *weighting* variable. Observations with more trials are given more weight in analysis because we have more information about those observations.

The response variable can therefore be stored in two ways.

1. Have one column for the number of successes and another for the number of failures or the total number of trials per observation.
2. Have one column for the proportion of successes and another for the total number of trials.

If you have only the proportion and not the number of trials and the number of trials isn't a fixed value, you cannot use a binomial GLM.

## 4.2 Check for overdispersion

Overdispersion is the first thing to check for when fitting a binomial GLM.

Since we know that overdispersion can be a symptom of a missing covariate, we will have anticipated this and collected and included covariates that we believe cause variation in our response in the model. We should be checking for overdispersion on the fullest reasonable model we can fit. Sometimes, as in the case of the motivating example, we have no additional information to use so the fullest model is fairly simple.

An estimate of the overall overdispersion is the sum of the squared Pearson residuals divided by the residual degrees of freedom. Values greater than 1 indicate overdispersion. There are no hard and fast rules for how much overdispersion is too much. An overdispersion value of 2 is definitely considered large, but a value of 1.1 is small and likely doesn't matter too much. The researcher has to decide when the value is large enough to be problematic. If we see evidence of overdispersion (or even if we don't), we would report that we checked for overdispersion and the estimated value for overdispersion.

For binomial GLMs, we can often operate under the assumption that we probably have overdispersion. Since extra-binomial variation is so common, it is likely okay to simply assume that you have overdispersion and start with a model to address the issue.

## 4.3 Complete separation

We should already have a pretty good idea if we have complete separation because we will have already looked at our dataset prior to analysis. If we know this is an issue we will have already been thinking about what we want to do with a group that causes complete separation and have anticipated any modeling approaches that we might want to use.

If we have quasi-complete separation, we usually only become aware of it because of warning messages and *giant* SE in the summary output. Again, this is not necessarily a problem, but quasi-complete separation can indicate overfitting (i.e., our model is too complex for our data) and we need to rethink our model. Alternatively, it may mean that some complex combinations of groups are all 0.

## 4.4 Role of residuals

While we no longer have additive errors in our model, we will still use plots of residuals to help check model fit. There are different types of residuals we can get, which we use for different things.

- The Pearson residuals are what we use in the calculation for checking for overdispersion.
- Deviance residuals are considered to be "well behaved", and so can be good for making residual plots. Most deviance residuals should fall between -2 and 2, allowing us to see any points that would be considered unusual under the current model. In addition, the way deviance residuals are calculated we expect to see reasonably similar variance among groups in any residual plots.

Note there is no assumption of normality of errors in a binomial GLM, so don't check any assumptions of normality using residuals. While the deviance residuals are technically asymptotically normal, we can't expect this to be true for small samples. You can see an example where the deviance residuals will never be approximately normally distributed here, https://github.com/florianhartig/DHARMa/blob/master/Code/DHARMaExamples/DevianceResiduals.md. I find that checking for normality of the residuals is more confusing than helpful, since it makes people think that normality is some assumption of all GLMs when it's not. We will discuss quantile residuals in class.

## 4.5 Hypothesis tests and confidence intervals

Of course no hypothesis test or confidence interval should be calculated prior to addressing any issues you found in the model, such as overdispersion. Once you have a model that you are happy with, you can get estimates with confidence intervals and any tests you want to report.

For any hypothesis tests, a good standard option is to use *likelihood ratio tests*, also known as *drop-in-deviance tests*. The "worst" tests we can use are the Wald $z$ tests in the model summary output, so use those as a last resort. Among other issues, it is the Wald tests that are problematic when we have complete separation. We can also get Wald $\chi^2$ tests from the `car` package `Anova()` function, but those are even less reliable than likelihood ratio tests. See a more complete discussion of hypothesis tests for GLM/LMM/GLMM models here, https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#testing-hypotheses.

If you decide to use a quasilikelihood model, it is standard to use an $F$-based likelihood ratio test instead of a $\chi^2$-based likelihood ratio tests. $F$ tests are more conservative. You will see an example of how to calculate these tests in lab this week.

The gold standard method to calculate confidence intervals from a GLM is by profiling the likelihood. This is true for any GLM and can be done with `confint()` on the object if the model was fit in R with the `glm()` function. Add-on packages will likely report Wald confidence intervals. These tend to be okay if there is no complete separation, particularly if we have a lot of data. Note if using quasilikelihood methods, any Wald confidence intervals should be based on the $t$-distribution and not the $z$-distribution.