FES 524: Natural Resources Data Analysis

Reading 7.1: Generalized linear models

Contents

1 Introduction to generalized linear models

I am going to present some theory below for anyone who is curious, but **feel free to skip ahead to Section 2 if you do not care for the theory aspect and want a more practical introduction**.

1.1 Theoretical background

Generalized linear models (GLMs) are an extension of the linear models we have discussed so far in class. By extension, I mean that the linear models we have discussed are GLMs, but GLMs include a much broader collection of models. To define GLMs more broadly, however, we need to adjust how we write our linear models slightly. So far, we have defined our models most broadly as

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma} + \epsilon_i$$

where $\mathbf{x}_i$ is a $p \times 1$ vector of explanatory variables measured for the i^{th} observation, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\mathbf{z}_i$ is a vector of 0s and 1s to "pull out" the random effects associated with the i^{th} observation from the vector of random effects, $\boldsymbol{\gamma}$, and ϵ_i is a random error from a normal distribution. Most often, we assume that $\epsilon_1, \epsilon_2, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, but we also discussed some commonly used correlation structures for the errors last week.

An alternative way to write this is:

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}$$
$$y_i | \mu_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

where μ_i is the mean or expected value of the the i^{th} observation, $g()$ is some function called a *link* function, and the vertical bar, $y_i|\mu_i$, means *given* or *conditional on*. That is, our response variable is only normally-distributed after accounting for the explanatory variables and random effects and how they affect the mean. The link function "links" the mean of the distribution of the response variable to a *linear predictor* (i.e., the familiar linear model with explanatory variables and regression coefficients). We will discuss what the link function is and why it is important below and in lecture.

1.2 Generalizing to the exponential family

The exponential family (some day, I will make a goofy graphic with all the exponential family distributions as anthropomorphized characters in a family photo...) is a collection of probability distributions that are related by a specific formula. That is, they can all be written as

$$f_Y(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{1}$$

where Y is the random variable, θ is the *canonical parameter*, ϕ is a *dispersion* parameters, and $a()$, $b()$, and $c()$ are all functions that are unique to the specific sub-family of distributions. These have specific names, but the only one we will name here is the *cumulant function*, $b(\theta)$. This has importance to statistical theory, but we will not get into it.

The form in Equation 1 is known as an *exponential dispersion model* (EDM) and is the key to generalizing our familiar linear models to new horizons. As a quick demonstration, let's re-write the density function for a normal distribution in this form.

1.2.1 The Normal distribution as an EDM

Starting with some rearrangement of the normal pdf,

$$\begin{aligned}
f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\left(-\frac{\mu^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2}\right) - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\
&= \exp\left\{\left(\frac{y\mu - \mu^2/2}{\sigma^2}\right) + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\right\}
\end{aligned}$$

now define the following:

- $\mu = \theta$ is the canonical parameter and $b(\theta) = \theta^2/2$ is the cumulant function
- $\sigma^2 = \phi$ is the dispersion parameter and $a(\phi) = \phi$
- $c(y,\phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$.

If we then complete the substitutions, we see we wind up at the exponential dispersion model of Equation 1.

1.3 The canonical link function

The reason I am presenting the above statistical theory is that you may often hear the term *canonical link function* when fitting or learning about GLMs. This idea of the "canonical" link comes directly from the form in Equation 1. For distributions other than the normal distribution, the relationship between the mean and the linear predictor does not happen on the scale of the data. Distributions have natural or canonical link functions based on the specific form of $b(\theta)$ once we re-write the probability distribution as an EDM. Specifically, for any EDM, the mean, $\mathbb{E}(Y) = b'(\theta)$. Thus, the mean can easily be derived in terms of the canonical parameter. From there, we just need the relationship between θ and μ to find a natural or canonical link to link the linear predictor to the mean. For a normal distribution,

Table 1: Components of a GLM

Systematic component	$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$
Random component	$Y_i \vert \mu_i, \phi \sim f_Y(\mu_i, \phi)$

$$b'(\theta) = \frac{d}{d\theta}\frac{1}{2}\theta^2 = \theta$$

and

$$\theta = \mu,$$

so the GLM can be written as

$$\theta_i = g(\mu_i) = \eta_i$$

where I am using η_i to be the linear predictor of the i^{th} observation. Since we know that $\theta = \mu$ for the normal distribution, this tells us that the canonical link for a normal model is the identity link (i.e., the "do nothing" or "multiply by 1" function). It is standard to drop the extra notation and not even write the link function, as we have been doing so far this term, even though it is there in theory (lurking...).

2 GLM structure and overview

In order to communicate the models we will be fitting, we will be using a sort of alternate notation for a linear model. Thus far, we have written our models by adding the error term onto the end of the linear predictor equation. Specifically, we have been writing

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for the simple linear regression model, where we assume the usual iid normal assumptions for the error terms. However, an alternative way to write this, which I will refer to below as the *probability model syntax*, is to say,

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 x_i \\ y_i \vert \mu_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

where μ_i is the mean or expected value of the the i^{th} observation, $g()$ is some function called a *link* function, and the vertical bar, $y_i|\mu_i$, means *given* or *conditional on*. That is, our response variable is only normally-distributed after accounting for the explanatory variables and how they affect the mean. The link function "links" the mean of the distribution of the response variable to a *linear predictor* (i.e., the familiar linear model with explanatory variables and regression coefficients). We will discuss what the link function is and why it is important below and in lecture, and Table 2 gives a list of commonly used link functions (i.e., those that are *natural* or "work well" with a given distribution for the response). The first equation is the *systematic* component, which describes how the mean changes with changes in the explanatory variables. The second component is the *stochastic* or random component, which describes how the "noise" around the mean is distributed.

The *conditional response variable*, $y_i|\boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_i$ is a list of parameters all collected into a vector, can be modeled as originating from a wide variety of distributions. All we need is a link function to link the mean to a linear predictor and an error distribution to describe the stochastic or random component of the model (Table 1). This means we can encompass many types of data since we can use different distributions to describe the data generating process, given a mean value μ.

Table 2: Table listing common distributions for the conditional response in a GLM, their usual parameterizations, the mean-variance relationship, and the usual link function.

Distribution	**Usual parameters**	**Mean**	**Variance**	**Usual link function**
Normal	$-\infty < \mu < \infty$ (mean) $\sigma^2 > 0$ (variance)	μ	σ^2	$g(\mu) = \mu$
Gamma	$\alpha > 0$ (shape) $\beta > 0$ (scale)	$\mu = \alpha\beta$	μ^2/α	$g(\mu) = \mu^{-1}$
Poisson	$\lambda > 0$ (rate)	$\mu = \lambda$	μ	$g(\mu) = \log(\mu)$
Negative Binomial	$\mu > 0$ (mean) $\phi > 0$ (dispersion)	μ	$\mu + \mu^2/\phi$	$g(\mu) = \log(\mu)$
Binomial/Bernoulli	$0 < p < 1$ (probability) $m \in \mathbb{N}$ (size)	$\mu = p$	$\mu(1-\mu)/m$	$g(\mu) = \text{logit}(\mu)$
Beta	$\alpha > 0$ (shape) $\beta > 0$ (shape)	$\mu = \alpha/(\alpha+\beta)$	$\mu(1-\mu)/(1+\psi)$[1]	$g(\mu) = \text{logit}(\mu)$

2.1 Types of response variables

Since everyone learns the special case of linear models with normally-distributed errors first, you may not be used to thinking about what sort of distribution could be used to model different types of variables.

Here are some common categories of response variables:

- Continuous and symmetric
- Continuous, nonzero, potentially right-skewed
- Counts (including or excluding zeros)
- Counts out of a total (discrete proportions)
- Presence / absence
- Continuous proportions, where there is no "total" defined
- Positive, continuous variables with (potentially many) true zeros

2.1.1 Continuous and symmetric

Continuous, symmetric response variables are what we have been talking about so far this quarter. These kinds of variables are when the special case of the normal distribution can make sense.

2.1.2 Positive continuous and right-skewed

Positive, continuous, right-skewed variables are what we discussed reading 4.2. These are data with no 0 values. The log-normal distribution is one option for a probability model for this kind of variable, where we use a log-transformation on the response variable and then assume normality. You learned about the log-normal distribution in reading 4.2. The gamma distribution is another option, which you may or may not have seen before.

The gamma distribution is a two-parameter distribution usually denoted Gamma(α, β), where α is a *shape* parameter and β is a *rate* parameter. The exponential distribution is a special case of the gamma distribution when $\alpha = 1$. You can see how these parameters influence the shape of the curve using this desmos graph: https://www.desmos.com/calculator/vk2tqrxpk5. To recast this as an EDM, we reparameterize in terms of the mean and variance (Table 2). Once we have a mean and variance defined, we can specify the gamma probability model. From Table 2, you can see that the mean and variance of the distribution are related to

[1] $\psi = \alpha + \beta$ is the precision parameter, the inverse of the dipsersion.

one another. This "mean-variance" relationship may or may not be a good model for the heteroskedasticity in a given dataset.

2.1.3 Counts

Counts are discrete, integer values that can include zero. For example, we could be counting seedlings in a plot or the number of bird nests in a forest stand. The most common distributions used for modeling counts in ecology are the *negative binomial* and the *Poisson* distributions.

2.1.3.1 Negative binomial distribution The negative binomial distribution often works well in ecology because it works when the presence of an organism in an area is more likely if there are other organisms already there. For example, we will likely find a lot of seedlings together when there are many parent trees in the area but there will be no or few seedlings when there are no parent trees in the area. This is a type of *clustering in space*, where we either find large clusters of seedlings or very few seedlings.

When working with count data where the negative binomial may work well you will likely have many observations with no or low counts as well as a few very high counts.

The negative binomial distribution is a two parameter distribution, and is usually parameterized in terms of it's mean, μ, and the *dispersion parameter*, ϕ. From Table 2, you can see that the mean and variance of the negative binomial distribution are also related, like with the gamma distribution (albeit a different relationship).

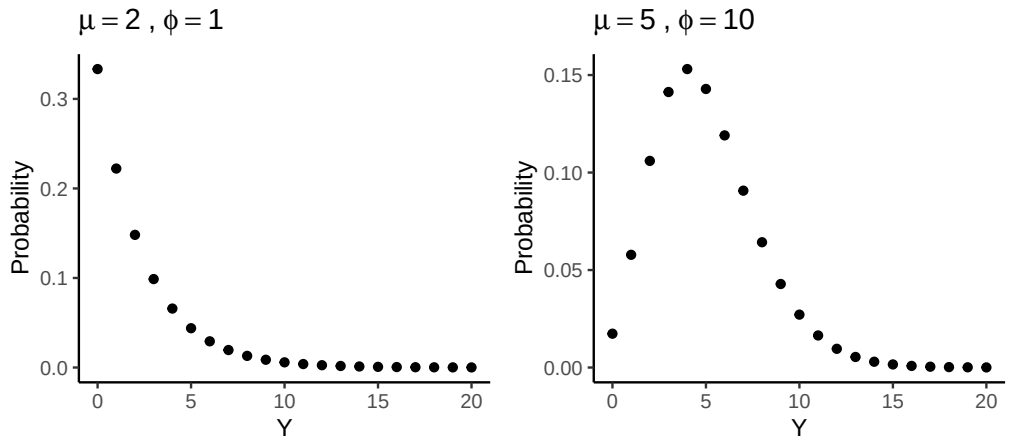

Figure 1: Two examples of negative binomial distributions with different parameter values. Because it is a discrete distribution, the points represent *point masses* of probability.

2.1.3.2 The Poisson distribution The Poisson distribution can describe counts that cover a limited range. It is possible to have overall very high counts or overall very low counts in this distribution, but it does not encompass having both very low and very high counts at the same time. Because of this, it doesn't often work well in ecology.

I have seen the Poisson distribution work for modeling species richness (number of species present) when the number of species is limited, or counts of individuals in very small plots.

The Poisson distribution is the first one parameter distribution we have seen. The single *rate* parameter is usually denoted λ. As shown in Table 2, the variance not only depends on the mean, it is *exactly equal* to the mean. This is one of the reasons the usefulness of this distribution is limited. Figure 2 shows two examples of a Poisson distribution with different rates. Note the limited range on the x-axis.

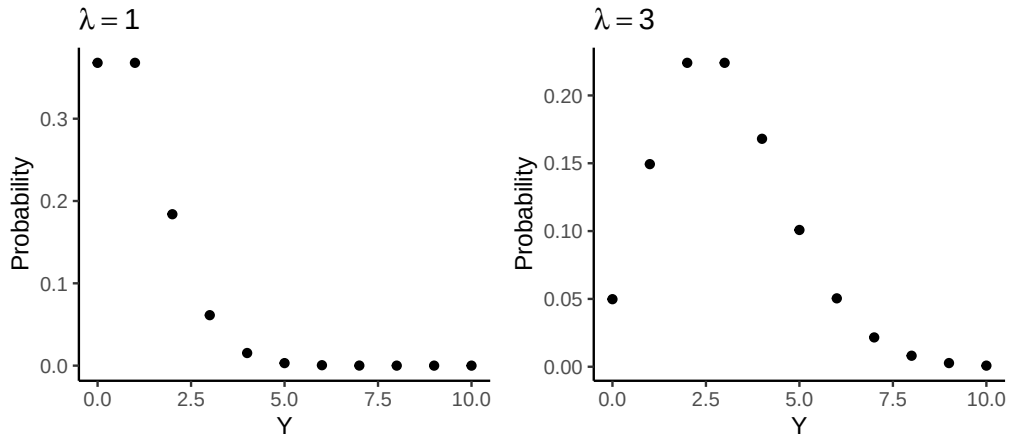

Figure 2: Two examples of Poisson distributions with different parameter values. Because it is a discrete distribution, the points represent *point masses* of probability.

2.1.4 Counted proportions

Counted proportion variables are discrete variables that include 0. They are the counts of the number of subjects that meet some criteria (i.e., successes or "yeses") out of the total number of subjects counted (i.e., number of trials). These are conditionally independent trials; after accounting for known sources of variation, the outcome of one trial doesn't affect the outcome of another trial. Some examples include the number of trees that die out of a counted number of trees infested by a pest, or the number of successfully fledged birds out of the number of eggs lain.

2.1.4.1 Binomial distribution Be careful to recognize that the binomial distribution is not the same as the *negative* binomial distribution we discussed for counts. This is because, for the binomial distribution, we *fix* the number of trials and count the number of successes out of the total trials, while for the negative binomial distribution, we fix the number of successes we want to see and then count the number of trials it takes to reach that many successes. You must therefore know the total number of trials in order to use the binomial distribution.

This is another one parameter distribution, and includes values of 0 and 1. We will use p for the parameter, but you will also see π pretty often. Table 2 shows the mean variance relationship for the binomial distribution. Notice that, while the m is the variance equation looks like another parameter, that is the number of trials, which we already said we need to know. So, like the Poisson distribution, if we know the mean, we know the variance. This can be limiting in some cases.

2.1.4.2 Presence / absence (Bernoulli) Presence / absence is a counted proportion when the number of trials is 1. Such data can only have values of 0 or 1 (i.e., yes/no or success/failure). For example, we could be assessing which snags are used as nests for woodpeckers. Either the snag is or it isn't used for a nest in a given season.

2.1.5 Continuous proportions

When proportion values do not come from a count per trial, they are continuous proportions instead of counted proportions. For example, a proportion of leaf area infected by a rust or the cover of invasive grasses in a quadrat are not counts out of a total, but continuous quantities out of a continuous total.

2.1.5.1 Beta distribution The beta distribution has classically been used for continuous proportions. However, this distribution does not include 0 or 1 values. This makes it less useful than it otherwise would be, and often folks end up using zero, one, or zero-and-one *inflated beta distributions*. These are called "inflated" distributions but are really a type of hurdle model (hurdle models discussed more below and in reading 8.2).

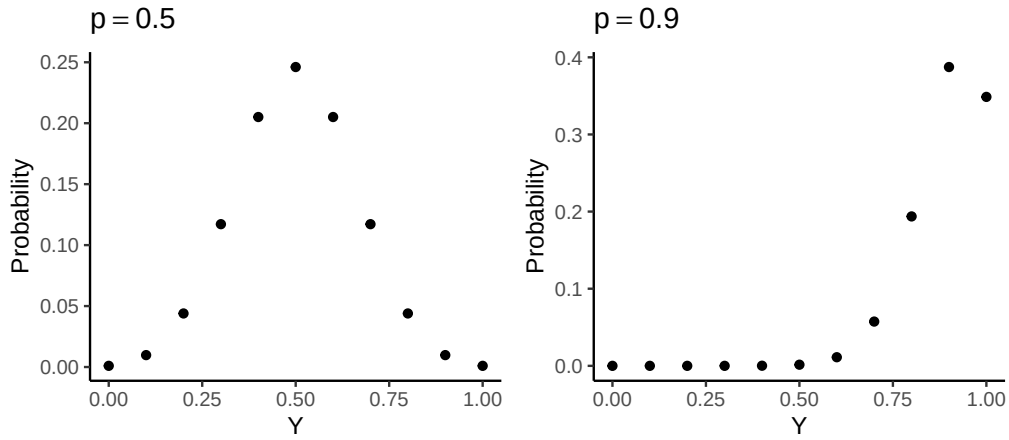

Figure 3: Two examples of Binomial distributions with different parameter values. Because it is a discrete distribution, the points represent *point masses* of probability.

The beta distribution is a two parameter distribution, usually denoted with two *shape* parameters, α and β. The beta distribution is extremely flexible, and can model continuous proportions with many different shapes. It is a continuous distribution, so Figure 4 shows a the "density" curve, which is a smooth curve, unlike the point masses of the distributions discussed above.

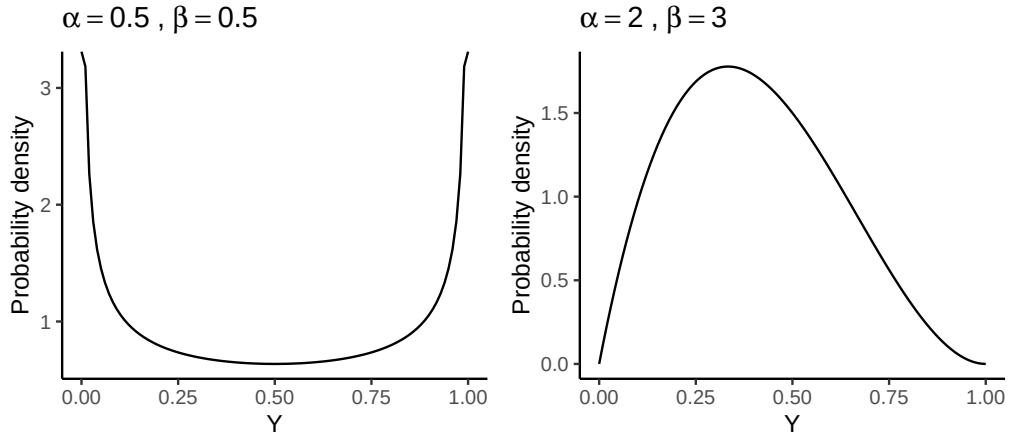

Figure 4: Two examples of beta distributions with different parameter values. Because it is a continuous distribution, the y-axis is the *probability density*.

2.1.6 Positive continuous with a point mass at 0

Positive, continuous data with a point mass at 0 (i.e., many truly zero values) are notoriously difficult to analyze. An example of these data would be measure plant species cover in plots where you can have true zero cover, but otherwise continuous values. Historically we didn't have a distribution to deal with such data.

One option for analyzing such data is to do two analyses instead of one. In the first, you can model presence/absence using all the data with a binomial distribution. In the second model you analyze only the continuous data using a log-normal or gamma distribution. This approach is called a *hurdle* model. You could do a log-normal hurdle model or a gamma hurdle model, depending on the distribution used for the second model.

A newer option is the Tweedie distribution. This distribution has a wide variety of shapes, but when it is what we call a *compound-Poisson-gamma*, then it has a point mass at 0 as well as positive, continuous values

that are usually right-skewed. The Tweedie has worked well for difficult situations involving total cover and is worth considering if you are working with such data.

2.2 Transformation

Where does transformation fall in all of this?

First, make sure you recognize that using a GLM is not the same as transforming the raw data. If using a GLM you will not be using a transformed response variable.

We discussed the general issues with transformation back in unit 4.2. Something new to understand this week is that transformations of data for use in a LM when the variable actually comes from a different distribution can fail to give correct estimates and/or correct estimates of the standard errors. Stroup (2015) has a good discussion on this, which is definitely worth a read.

3 Binomial GLM

Now that we have covered generalized linear models in general, we can focus in on the topic of the week: binomial generalized linear models.

3.1 The probability model

Since we are working with GLMs from now on, let's practice writing out the probability model. We can start by first defining the conditional distribution of the response variable. For this, we write

$$Y_i|\mu_i \sim \text{Binomial}(p_i, m_i)$$

where μ_i is the mean or expected value for the response variable, which is defined by the linear predictor, which we will define next. p_i is the proportion or probability of success, and m_i is the number of trials for the i^{th} observation. Notice that each observation gets its own parameters, p_i, μ_i, and m_i, since the mean changes with each counted proprotion based on the explanatory variables. Similarly, the total count could change with each observation, so we write m_i. For example, at each plot, I could have different numbers of snags in which I count the number that have woodpecker nests.

The systematic component, the linear predictor, is written as

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

since the mean, μ_i, is p_i, the probability of success for any one trial out of the total. We have written this using the generic function g to represent the link function, however, we need to choose an appropriate link function that relates the mean to the linear predictor.

3.2 Logit link

For proportions, whether continuous or counted, the canonical link function and the most-used link function is the logit link (Table 2). The logit is also called the *log odds*. Recall that the odds of an event is the probability of that event divided by the probability that the event does not take place. Thus, the logit function is the log of the odds,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where log() is the natural log (sometimes written as ln()). Thus, the systematic component of the model can be written as

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i. \tag{2}$$

Notice that this sort of looks like we are transforming the response, but recall that p_i is the mean of the distribution, not the observed response.

3.3 Thinking on different scales

One of the reasons we are starting our work with GLMs with a binomial GLM is because they really highlight the importance of the interpretation of different components and parameters in the model. Using a logit link means we need to understand three different scales: the model scale or the scale of the linear predictor (log odds), the scale of interpretation (odds), and the data scale (proportions/probabilities).

3.3.1 Model scale

The model scale is the scale at which we do hypothesis tests and calculate confidence interval limits. The interpretation at this scale is to log odds, which we can see in the equation to model the mean (2). The linear predictor is additive on this scale. However, we are pretty much never going to be interested in talking about results for additive differences in estimated log odds. It simply doesn't mean much.

Below is a plot on the scale of the true (not estimated) relationship from a hypothetical model. I'm pretending the response variable is the proportion of insect pests that die in a trial with increasing pesticide concentration. I made a plot with a continuous X because it is easiest to see additivity (linearity) on the model scale when using a continuous explanatory variable. Additivity still holds true for categorical variables.

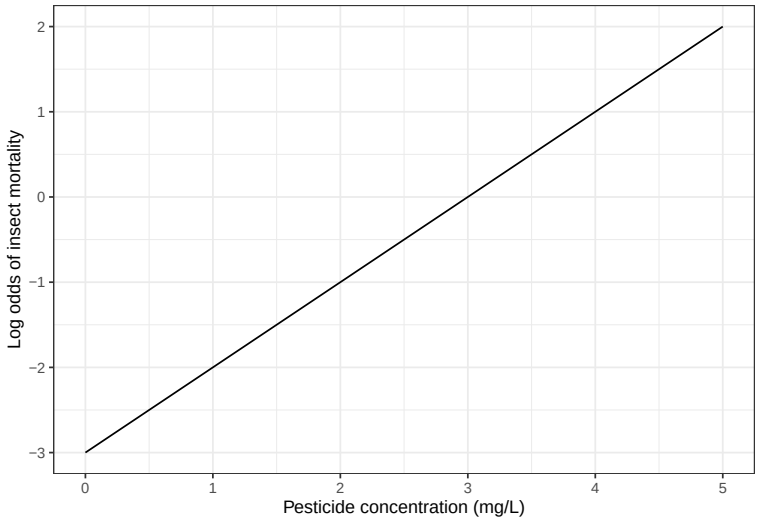

Figure 5: Example relationship on the model scale for a hypothetical example of the proportion of insect pests that are killed with increasing pesticide concentrations.

The relationship in Figure 5 shows that for every 1 unit change in concentration we get a 1 unit increase in the log odds of mortality. This change is the same if X goes from 0 to 1 or from 3 to 4.

3.3.2 Odds scale

The binomial GLM with a logit link is unusual because we interpret results on neither the link/model scale nor the data scale. If we exponentiate both sides of the model we are able to interpret results as odds. All estimated relationships are multiplicative changes because of the exponentiation. We practiced our language for multiplicative relationships in week 4; you may want to review that material this week.

This is what the model looks like on the odds scale. You can see that odds are now on the left-hand side of the equation:

$$\frac{p_i}{(1\breve{}p_i)} = \exp(\beta_0) \times \exp(\beta_1 x_i) \tag{3}$$

Figure 6 shows the same model as in Figure 5, but on the odds scale.

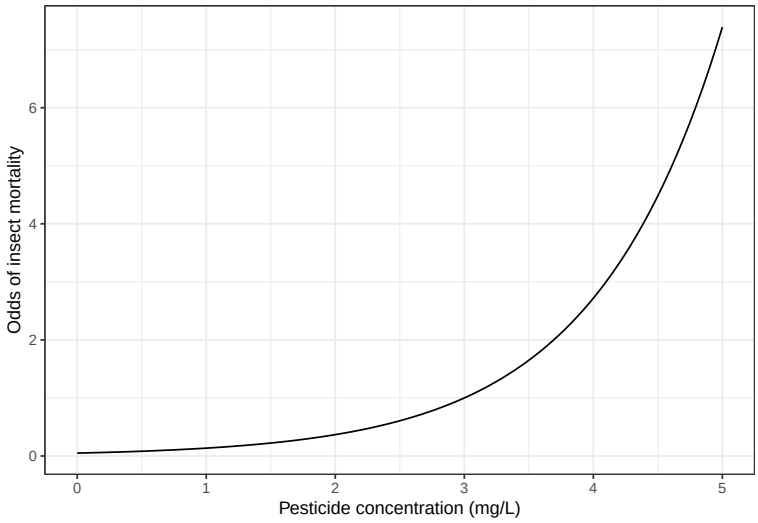

Figure 6: Example relationship on the odds scale for a hypothetical example of the proportion of insect pests that are killed with increasing pesticide concentrations.

Working through an example with some actual values, the odds of mortality when $x = 2$ is 0.37. The odds of getting a tumor when x is 3 is 1. $1/0.37 \approx 2.7$. This multiplicative increase is the same for any other 1 unit increase in x, even though the linear increase in odds will not be the same.

3.4 Data scale

We can use the inverse of the link to go back to the data scale, the scale of proportions or probabilities. This is most useful for making graphics. Note that this is not back transforming since we never transformed the response variable. Instead we use language about inverses, taking the inverse of the link.

The function for the inverse of the logit link in R is `plogis()`, and is equal to

$$\text{logit}^{-1}(p_i) = \frac{e^{p_i}}{1 + e^{p_i}}.$$

After using the inverse link, the left-hand size of the model now only contains the mean. The right-hand side of the model, however, is a bit:

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}. \tag{4}$$

Due to the nature of the logit link and its inverse, there is no clean interpretation of changes in or differences among the proportions. The change depends on the value of x and the interpretation of the estimated coefficients are neither multiplicative or additive. This is true when we have a categorical explanatory variable as well as this example with a continuous explanatory variable. Plan on sticking to odds for interpretation and using plots and tables of estimated proportions to help your reader understand what the odds mean practically. Figure 7 shows the same model we have been working with on the probability scale.

The amount the probability of insect mortality for a 1-unit increase in X depends on the starting value of X. If X goes from 0 to 1 (an increase of 1), the probability of mortality increases by about 0.07 (additive). If X goes from 3 to 4 (an increase of 1), the probability of mortality increases by about 0.23 (additive).

Binomial GLMs will be the focus of week 7.

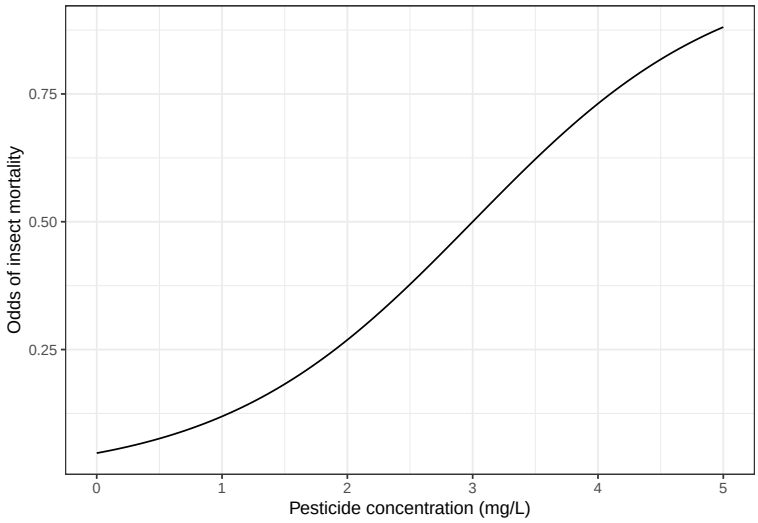

Figure 7: Example relationship on the probability scale for a hypothetical example of the proportion of insect pests that are killed with increasing pesticide concentrations.

References

Stroup, Walter W. 2015. "Rethinking the Analysis of Non-Normal Data in Plant and Soil Science." *Agronomy Journal* 107 (2): 811–27. https://doi.org/10.2134/agronj2013.0342.