FES 524: Natural Resources Data Analysis

Reading 7.1

**Contents**

[Other readings to do before class 1](#_Toc63924013)

[Introduction to generalized linear models 2](#_Toc63924014)

[Model equation 2](#_Toc63924015)

[Probability model 2](#_Toc63924016)

[Canonical link function 3](#_Toc63924017)

[Generalized linear model overview 3](#_Toc63924018)

[“General” linear model 3](#_Toc63924019)

[Types of response variables 4](#_Toc63924020)

[Continuous and symmetric 4](#_Toc63924021)

[Normal distribution 4](#_Toc63924022)

[Positive continuous and right-skewed 5](#_Toc63924023)

[Gamma distribution 5](#_Toc63924024)

[Counts 6](#_Toc63924025)

[Negative binomial distribution 6](#_Toc63924026)

[Poisson distribution 7](#_Toc63924027)

[Counted proportions 8](#_Toc63924028)

[Binomial distribution 8](#_Toc63924029)

[Presence/absence 9](#_Toc63924030)

[Continuous proportions 9](#_Toc63924031)

[Beta distribution 9](#_Toc63924032)

[Positive continuous with a point mass at 0 10](#_Toc63924033)

[Variance depends on mean 10](#_Toc63924034)

[Error distribution 11](#_Toc63924035)

[Transformation 11](#_Toc63924036)

[Binomial generalized linear model 11](#_Toc63924037)

[Probability model 11](#_Toc63924038)

[Logit link 12](#_Toc63924039)

[Model scale 12](#_Toc63924040)

[Odds scale 13](#_Toc63924041)

[Data scale 14](#_Toc63924042)

We change gears this week to focus on generalized linear models.

# Other readings to do before class

Read Week 7 Handout 1: “Study on effect of aflatoxin on tumors in salmon fingerlings” and come prepared to use this example to talk about the topics covered in reading 7.1. Make sure you’ve had time to think about the answers to the questions in that handout.

# Introduction to generalized linear models

Today we are going to start learning about *generalized* linear models, a big step away from the linear models we have discussed so far in the class and from models you likely learned in your introductory statistics classes. We have to change the way we think about the statistical model in order to understand what is going on in a generalized linear model, so we are going to take time to go through the mathematical notation first.

## Model equation

So far this quarter we’ve seen a single *model equation* when describing linear and linear mixed models with mathematical notation. This is the traditional equation form, which was used originally because such models provided standard justification for using linear model tools (i.e., ANOVA, regression) based on minimizing the sums of squares. Such equations are only relevant, however, if a normal distribution of the errors is plausible.

The single model equation has all terms of interest in it. The linear part of the model is related directly to the observed response variable.

For example, here is the model equation for a linear model with a single, continuous explanatory variable:

Yt = β0 + β1Xt + εt

When using a model equation like this the distribution is defined for the errors:

εt ~ N(0, σ2)

## Probability model

In the *probability model*, which can be used more generally than the model equation, we describe two equations instead of a single equation.

We first define the *distribution* of the observations conditional on X (explanatory/fixed effects) and Z (random effects). This *conditioning* means that the response variable only follows the distribution we define after accounting for terms in the model. Remember that this conditioning is one reason why we do model checks with the residuals; the raw data are observed values without conditioning on the variables in the model.

After defining the conditional distribution of the response, we describe linear predictor of the *mean* response.

Here is an example of the probability model written for the model equation I wrote above.

We start with the conditional distribution of the response variable:

Yt | Xt ~ N(μt, σ2)

We can see the response is not simply listed as Yt but as Yt | Xt. This notation makes it clear that it is the response variable *only after accounting for the explanatory variable* that is expected to follow a normal distribution. We do not expect the raw data to follow a normal distribution.

You can see the mean of the distribution of the response variable is represented by μt.

In a second step we write the *linear predictor* of the mean response variable. The part of the equation that contains the betas, on the right-hand side of this equation, is what we refer to as the linear predictor:

g(μt) = β0 + β1Xt

We can see the mean on the left-hand size of this second equation. The mean is *linked* to the right-hand size of the equation with some function I’m calling *g*.

## Canonical link function

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 of their probability distribution. Skipping over the statistical theory for this, the canonical function is a good candidate for describing how the mean of the distribution is related to the linear predictor.

Models assuming a normal distribution are a special case. As you know, we think things are linear on the same scale as the observed data. The canonical link function for the models we’ve been fitting this quarter so far is the *identity* function. When you write out a probability model when assuming the normal distribution, it is standard to not write out the link function at all:

μt = β0 + β1Xt

I introduced the model with an explicit link function *g* to make things more general, setting us up to talk about models with a conditional response variable from some other distribution.

## Generalized linear model overview

Generalized linear models are a *generalization* of the linear model. The acronym we use for generalized linear models is GLM. Conditional response variables in GLM’s can be from a wide variety of distributions, and we no longer need to focus so much on the normal distribution.

The response variable, conditional on everything in the model, comes from some distribution. The mean of the response variable is described by some relationship with explanatory variables (fixed effects) and/or random effects. If the model contains random effects we refer to the model as a generalized linear *mixed* model (GLMM). We will not have time to learn GLMM’s this quarter.

The models we’ve been fitting all quarter, assuming a normal distribution for the errors, are a special case of a GLM. This is called a linear model (LM). Using this terminology, a linear model is always a type of a generalized linear model but a generalized linear model is not always a linear model.

## “General” linear model

Confusingly, you may see the term “general” linear model in some papers or hear it used by people who use SAS or SPSS for analysis.

The “general” linear model is the old-fashioned term for what we call a linear model today. The term was coined in the 1970’s, when the linear model was “state of the art” even though they are not actually particularly general.

SAS and SPSS unfortunately named their functions for fitting linear models “GLM” during that era. As time has passed this has made things confusing, since we now have more statistical tools like generalized linear models which we appropriately call GLM. You may see the acronym “GLiM” for generalized linear models for people who use the term “general” linear models.

The point of this section is that while you may see the term “general linear model”, I recommend you use the contemporary language of “linear models” when describing analyses where you assumed normality of the errors.

# Types of response variables

In this next section we’ll start talking about different types of response variables and distributions that can be used to describe them. Since everyone learns the special case of LM’s first, assuming normality of the conditional response variable, we’re not used to thinking about what sort of distribution could be used to model different types of variables.

Here are some common kinds of response variables:

Continuous symmetric

Continuous, nonzero, right-skewed

Counts

Counted (discrete) proportions

Presence/absence

Continuous proportions

Continuous with point mass at 0

## Continuous and symmetric

Continuous, symmetric response variables are what we’ve been talking about so far this quarter.

**Examples**:

Height

Weight

Yield

These kinds of variables are when the special case of the normal distribution can make sense.

Remember that the normal distribution is also called a Gaussian distribution, which you might see in the literature or as an option in software.

## Normal distribution

The normal distribution is a two parameter distribution, usually notated with μ and σ2.

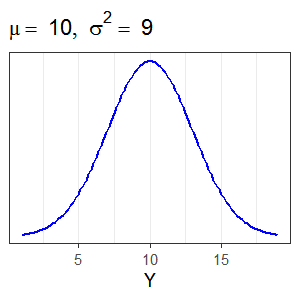
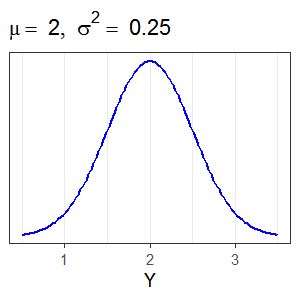
Here’s the mean and variance of the normal distribution, using this parameter notation:

**Mean**: μ

**Variance**: σ2

The variance of the normal distribution is completely independent of the mean.

Here are two examples of normal distributions with varying means and variances. They have the same shape but cover different values (compare the values on the x axes).

## Positive continuous and right-skewed

Positive, continuous, right-skewed variables are what we discussed in class 4.2. These are data with no 0 values.

**Examples**:

Stream discharge (cfs) (no ephemeral streams in sample)

Time to event

The log-normal distribution is one option, where we use a log-transformation on the response variable and assume normality. You learned all about the log-normal distribution in reading 4.2. The gamma distribution is another option, which you may not have seen before.

## Gamma distribution

The gamma distribution has a heavier left tail than the log-normal distribution. Seeing left skew after you log transform a variable is an indication that the gamma distribution might work better (if interested, see here for more discussion on this <https://stats.stackexchange.com/questions/72381/gamma-vs-lognormal-distributions>).

The gamma distribution is a two parameter distribution, usually notated with μ and φ. The μ parameter is called the *shape* parameter and the φ parameter is called the *scale* parameter.

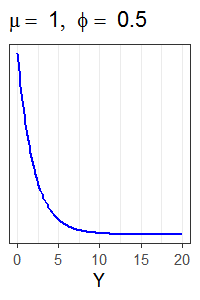
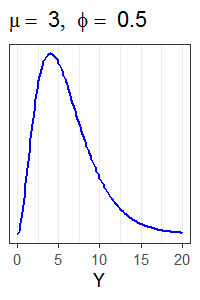
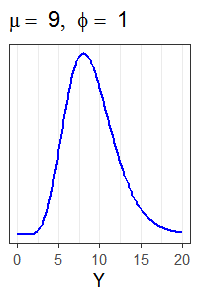
Here’s the mean and variance of the gamma distribution using these parameters:

**Mean**: μ

**Variance**: μ/φ

You can see that the variance of the gamma distribution depends on the value of the mean since the mean is present in the calculation of the variance.

Here are three examples of gamma distributions with varying shape and scale parameters. The distributions have very different shapes.

## Counts

Counts are a discrete variable that can include 0. Counts are integers.

**Examples**:

# seedlings per plot

# inflorescences per plant

It is common to have counts where the counting *effort* varies. The effort could be the amount of time spent counting or the size of the area counted in. This kind of variable can still be treated as a count if effort is used as an *offset*. We will talk about offsets more next week.

**Examples effort**:

Plot size

Amount cover present of plant species

The most common distributions used for modeling counts in ecology are the negative binomial and the Poisson.

## 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, usually notated with λ and θ. The λ parameter is called the *shape* parameter and the θ parameter is called the *scale* parameter.

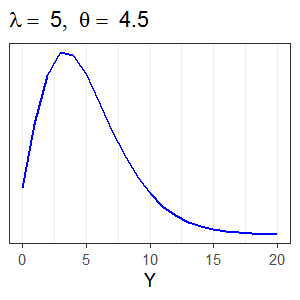
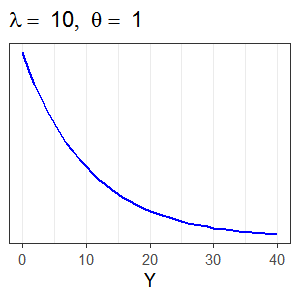
Here’s the mean and variance of the negative binomial distribution using these parameters:

**Mean**: λ

**Variance**: λ + λ2/θ

You can see that variance of the negative binomial distribution depends on the value of the mean.

Here are two examples of negative binomial distributions with varying shape and scale parameters. The distributions are both right-skewed but have different shapes. You can see the discrete nature of the underlying data in the first plot; the curve is no longer as smooth as the examples from continuous distributions.

## 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 tends to rarely work well in ecology

I have seen the Poisson distribution work for modeling species richness (# species present) when the number of species is limited.

This Poisson distribution is the first one parameter distribution we’ve seen. The parameter is usually notated with λ.

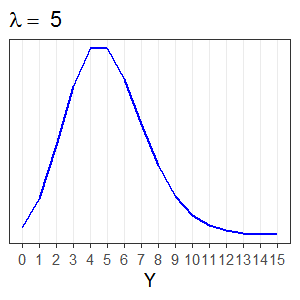
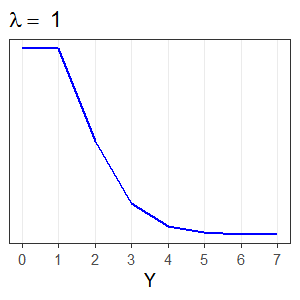
Here’s the mean and variance of the Poisson distribution:

**Mean**: λ

**Variance**: λ

The variance not only depends on the mean, but is exactly equal to the mean. This is one of the reasons the usefulness of this distribution is limited.

Here are two examples of Poisson distributions using two different means. You can see the discrete nature of the underlying data because the curves are no longer as smooth. You should note the limited range of the Y axes.

## 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.

**Examples**:

Proportion dead trees in a plot

Proportion damaged fruit on a tree

The binomial distribution is for counted proportions. Be careful to recognize that the binomial distribution is not the same as the *negative* binomial distribution we discussed for counts.

## Binomial distribution

The binomial distribution is for counted proportions. It includes both 0 and 1 values. You must know the total number of trials in order to use the binomial.

This is another one parameter distribution. We will use p for the parameter, but you will also see π.

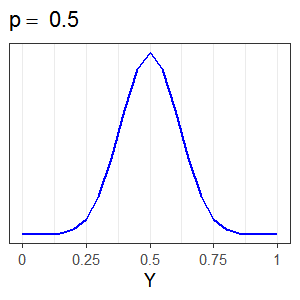
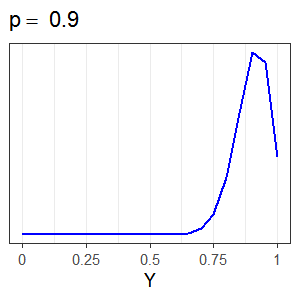
Here’s the mean and variance of the proportion from a binomial distribution:

**Mean**: p

**Variance**: p(1-p)/m

The variance of the binomial distribution depends on the mean. The m in the variance equation kind of looks like a parameter. However, it is actually the observed number of trials for a particular proportion. It is collected data, not a parameter.

Below are two examples of binomial distributions using two different proportions and 20 trials (m = 20). You can see the discrete nature of the underlying data because the curves are not smooth. Note the y axis always goes from 0 to 1.

## Presence/absence

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).

**Example**:

Tree is a bat roost or not

The binomial distribution is also used for presence/absence data. It is a special case of the binomial distribution when the number of trials is 1. Presence/absence data can also be referred to as *Bernoulli* or *binary* data.

## Continuous proportions

When proportion values do not come from a count per trial they are continuous proportions instead of counted proportions.

**Examples**:

Leaf area with lesions

Cover of shrubs in plot

The beta distribution can be used for continuous proportions.

## 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).

The beta distribution is a two parameter distribution, with μ as the shape parameter and φ as the scale parameter.

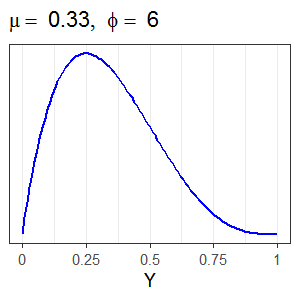
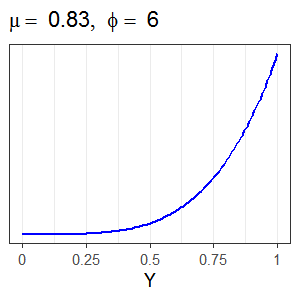
Here’s the mean and variance of the beta distribution:

**Mean**: μ

**Variance**: μ(1 - μ)/(1 + φ)

The variance of the beta distribution depends on the mean.

The beta distribution is extremely flexible, and can model continuous proportions with many different shapes. It is a continuous distribution, and the plots showing examples of different beta distributions are smooth.

## Positive continuous with a point mass at 0

Positive, continuous data with a point mass at 0 are notoriously difficult to analyze. Historically we didn’t have a distribution to deal with such data.

**Example**:

This issue has come up a lot with certain kind of measurements of plant cover, in particular when measuring the total cover for all vegetation layers in a plot.

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.

## Variance depends on mean

As you look back through all the distributions we just covered, you should note the variance-mean relationships again. In all cases except for the normal distribution, the variance is directly related to the mean. This means that for many, many types of response variables it doesn’t make sense to assume constant variance. The assumption of constant variance from a LM shouldn’t be met if the conditional response variable doesn’t reasonably come from the normal distribution.

**Distribution Mean Variance**

Normal μ σ2

Gamma μ μ/φ

Negative binomial λ λ + λ2/θ

Poisson λ λ

Binomial p p(1 - p)/m

Beta μ μ(1 - μ)/(1 + φ)

## Error distribution

Going back to our old model equation notation, we always defined the distribution of the errors:

εt ~ N(0, σ2)

The fact that the variance and mean are independent of each for the normal distribution is the reason we could define the distribution of the errors like this. Once the variance is related to the mean, the errors do not have a known distribution. The error distribution is only defined for LM’s. This means there are no assumptions to check about the distribution of εt when talking about GLM’s that don’t assume normality of the conditional response variable. In addition, we no longer include the observation-level random effect in the model by default once we start working with GLM’s.

## 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 class 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 makes this general argument in his 2015 paper “Rethinking the analysis of non-normal data in plant and soil science”, which is definitely worth a read. He shows a concrete example of how transformations fail when working with count data in Table 11.1 of his 2013 book, “Generalized Linear Mixed Models: Modern Concepts, Methods and Applications”.

# Binomial generalized linear model

Now that we’ve covered generalized linear models in general, we can focus in for the topic of the week: binomial generalized linear models.

Remember that the variance depends on the mean when working with the binomial distribution.

**Distribution Mean Variance**

Binomial p p(1 - p)/m

In what situation might it make sense to assume constant variance?

We will discuss this more in class. Be sure to bring a calculator for calculating variances.

## Probability model

Since we’re working with generalized linear models we need to practice writing out the probability model.

Remember that we first define the conditional distribution of the response variable.

For the binomial this is written:

Yt | Xt ~ Binomial(pt, mt)

If the number of trials did not vary we could put in a constant m instead of mt.

Every value of Yt has a mean associated with it, which we are calling pt. This mean is related to the explanatory variables (the *linear predictor*):

g(pt) = β0 + β1Xt

I have written this using the generic function *g* to represent the link function. We need to choose an actual link function that relates the mean to the linear predictor.

## Logit link

For proportions, whether continuous or counted, the canonical link function is the logit link. The logit is also called the *log odds*.

You will see the logit link written a couple of different ways. You may see it written as logit(pt).

This is equivalent to writing log( pt/(1 – pt) ) where pt/(1 – pt) is the *odds*.

The equation to relate the mean to the linear predictor for binomial data using the canonical link can be written as either:

logit(pt) = β0 + β1Xt or

log( pt/(1 – pt) ) = β0 + β1Xt

These two ways of writing the model are equivalent in their meaning.

The logit link is not the only link option for binomial data but it is very common, particularly in natural resources. It is so common, in fact, that people came up with a name for such analyses. The term *logistic regression* is one you may have seen used to describe an analysis using a binomial generalized linear model with a logit link.

One of the reasons we are starting our work with GLM’s with a binomial GLM is because these are among the most complicated GLM’s. Using a logit link means we need to understand three different scales: the *model scale* (log odds), the *scale of interpretation* (odds), and the *data scale* (proportions/probabilities).

## Model scale

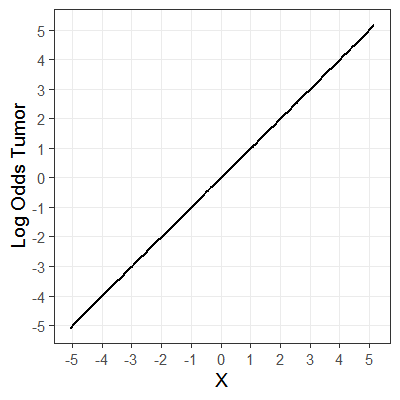
The model scale is the scale at which we do hypothesis tests and calculate confidence interval limits.

The interpretation at this scale is to log odds, which we can see in the equation to model the mean.

log( pt/(1 – pt) ) = β0 + β1Xt

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.

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 fish with tumors in a tank and that there is a single, continuous X variable of interest. 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.



The relationship in the plot shows that for every 1 unit change in X we get a 1 unit change in the log odds of getting a tumor. This change is the same if X goes from -2 to -1 or from 3 to 4.

## Odds scale

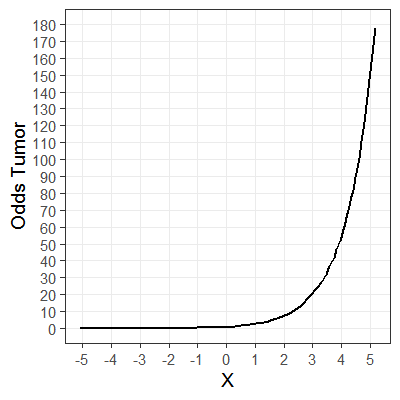
The odds scale is that scale of interpretation.

The binomial GLM with a logit link is unusual because we interpret results on neither the link 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. 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:

pt/(1 – pt) = exp(β0)\*exp(β1Xt)

Here is a plot of the true relationship from above after exponentiating to odds:



For a 1 unit increase in X, the odds of getting a tumor increases by a factor of ~2.7.

Working through an example with some actual values, the odds of getting a tumor when X is 3 is ~20.1. The odds of getting a tumor when X is 4 is ~54.6. 54.6/20.1 = ~2.7. This multiplicative increase is the same for any other 1 unit increase in X.

## 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.

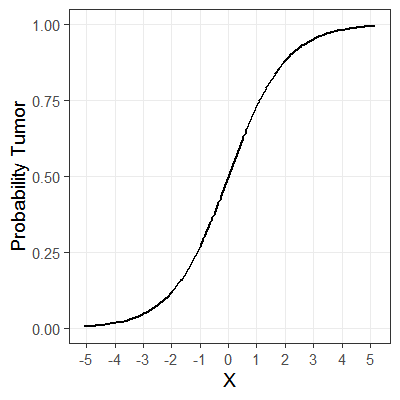
The function for the inverse of the logit link in R is plogis().

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 complicated:

pt = exp(β0 + β1Xt)/(1 + exp(β0 + β1Xt) )

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.

Here is the plot of the same true relationship after taking the inverse link:



The amount the probability of getting a tumor changes 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 getting a tumor increases by ~0.23 (additive).

If X goes from 3 to 4 (an increase of 1), the probability of getting a tumor increases by ~0.029 (additive).

We’ll continue to talk about binomial generalized linear models with the logit link in the second reading of the week.