Lecture 16: Overview to generalized linear modeling and link functions

This lecture is based on chapters 9 & 10 of Statistical Rethinking by Richard McElreath.

```
library(rstan)
library(shinystan)
library(car)
library(mvtnorm)
library(rethinking)
library(MASS)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
source("../utilityFunctions.R")
```

Thus far all of our models have worked by assuming a normal distribution over outcome variables y_i. Then, we replaced the scalar parameter μ with a linear model that gave us:

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i.$$

For response variables that are continuous and not close to theoretical maxima or minima, this type of model has maximum entropy. In other words, it is the least informative distribution that satisfies our prior knowledge of outcomes y.

But, if the response variables are discrete or bounded then Gaussian likelihoods are often inadequate.

Consider counts (e.g., the number of individuals of a species)

- These variables are constrained to be ≥ 0. Using a Gaussian model for such data won't result in a summary execution by the stats police, but the model won't be good at estimating more than the average count
- May give predictions that don't make sense (e.g., predicted observations < 0).

However, we can use our noggins and prior knowledge about the natural constraints of our data in picking another distribution that appeals to maximum entropy.

We do this by generalizing our linear regression strategy—replace a parameter describing the shape of the likelihood with a linear model—to non-Gaussian probability distributions.

This is what a *generalized linear model* boils down to. It looks like this:

$$y_i \sim \text{Binomial}(N, p_i)$$
$$f(p_i) = \alpha + \beta x_i$$

There are only two changes here from our familiar Gaussian model:

1. The likelihood function is binomial rather than normal. For a count response y where each observation comes from N trials and with constant expectation Np, the binomial distribution has maximum entropy.

   Most maximum entropy distributions belong to the exponential family (Fig. 1). This family includes:

   (a) *The exponential distribution*: Constrained to be zero or positive. A distribution of distance and duration, it models the displacement from some point of reference in time or space.

   - Described by λ, the rate of events, or λ^{-1}, the average displacement.
   - This is the core of survival analysis

   (b) *The gamma distribution*: Constrained to be zero or positive. Another distribution of distance or duration, but unlike the exponential, can have peaks above zero.

   - If an event can only happen after two or more exponentially distributed events occur, the resulting waiting times are gamma distributed. For example, the onset of cancer is approximately gamma distributed, because multiple events are necessary for onset.
   - Described by two parameters, but there are multiple formulations of those two parameters.

   (c) *The Poisson distribution*: Count distribution and special case of the binomial.

   - If the number of trials N is large (and usually unknown), and the probability of success p is small, then the binomial converges on the Poisson distribution with an expected rate of events $\lambda = Np$.
   - The practical application of the Poisson is for counts that never get close to any theoretical maximum.

2. The $f(p)$ indicates a *link function* is needed, which is determined separately from the likelihood distribution.

   - We need the link function because it is uncommon for there to be a μ parameter describing the average outcome, and rarely are parameters unbounded in both directions.
   - In the binomial example, there are two parameters, but neither is the mean. Instead, the mean outcome is Np—a function of both parameters.
     - We usually know N, so we attach a linear model to the unknown p.
     - p is a probability mass, so it must lie between 0–1. The link function keeps our linear model $\alpha + \beta x_i$ from exceeding those boundaries.

Link functions

To model from any of the exponential family of distributions, we just need to attach one or more linear models to one or more parameters describing the shape of said distribution.

- But, we need a link function to keep our model in bounds and prevent "mathematical accidents". What should that link be?

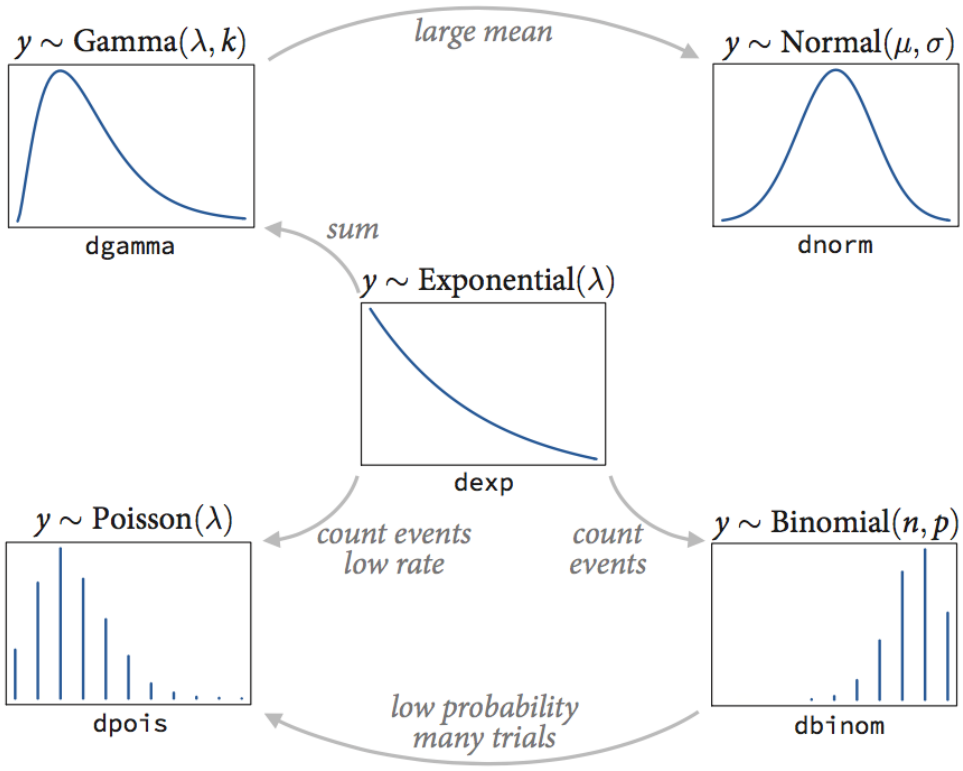

Figure 1: The exponential family of distributions (figure 9.6 of McElreath 2016).

The link function's job is to map the linear space of a model like $\alpha + \beta x_i$ onto the nonlinear space of our parameters. Usually, we will go with either a *logit* or a *log* link.

Logit link: maps a parameter defined as a probability mass—thus constrained between zero & one—onto an unconstrained linear model. The model definition will look like this:

$$\begin{aligned} y_i &\sim \text{Binomial}(n, p_i) \\ \text{logit}(p_i) &= \alpha + \beta x_i \end{aligned}$$

The logit function defines the *log-odds*:

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

The odds of an event are the probability that the event happens divided by the probability the event doesn't happen. So:

$$\log \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

With a bit of algebra, we get the *logistic* or the *inverse-logit* because it inverts the logit transformation:

$$p_i = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}.$$

This is illustrated in Fig. 2. On the left is a linear model ($y_i = 2x_i$) in log-odds land—thus it is unconstrained from $-\infty$ to $+\infty$.

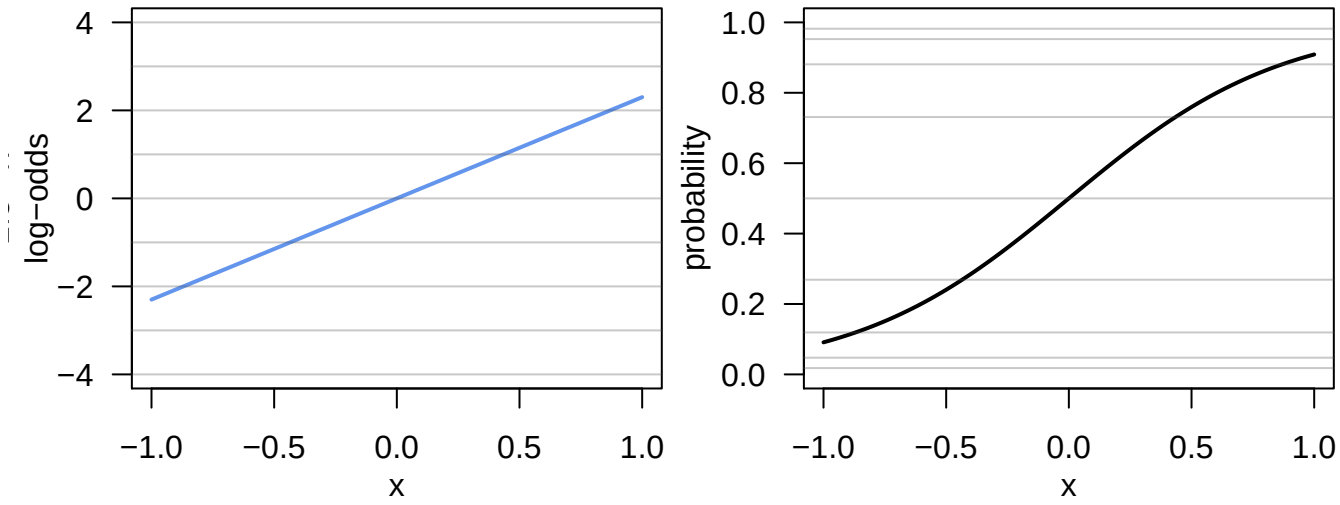

Figure 2: Logit link transforms a linear model into a probability. In doing so, it compresses geometries far from zero, so that unit changes mean less and less (code not shown).

- The horizontal lines indicate unit changes in the predicted values as x increases.

On the right, the linear model has been transformed to a probability mass and is now bounded between 0–1. The horizontal lines have been compressed near 0–1 to fit the linear model to the geometry of a probability.

- The compression produces the characteristic sinusoidal logistic curve.

This compression makes it more tricky to interpret parameter estimates because a unit change in x no longer results in a unit change in y. Instead a unit change in x will produce a larger or smaller change in p depending on how far the log-odds are from zero.

In our model above, when $x = 0$ the log-odds equal 0. Going up to $x = 0.5$ results in a ≈ 0.25 increase in probability.

- But increasing another half-unit to $x = 1$ results in only a 15% increase in probability. And another half unit increase only increases the probability by 7%.
- Each additional half unit produces less and less of a probability increase until the increases are vanishingly small.

This makes intuitive sense when you think about it. If an event is almost guaranteed to happen, its probability cannot increase very much regardless, irregardless even and more importantly, of how important a parameter might be.

Log link: maps a parameter that is defined over positive values onto a linear model. For example, we might want to model abundance counts of newts as a function of x. The model might look something like:

$$y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \alpha + \beta x_i.$$

The log link ensures that λ_i will always be positive, as is required of the expected value of count outcomes.

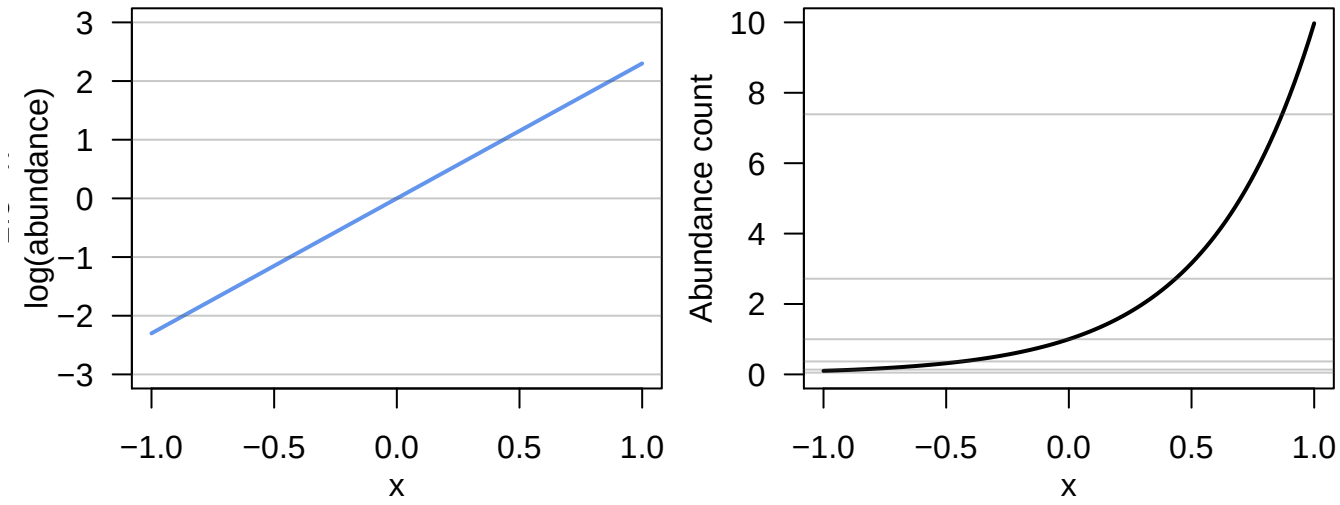

Figure 3: The log link transforms a linear model into strictly positive measurements. This results in exponential mapping of the linear model. A unit change on the linear log-count scale results in increasingly large changes on the outcome scale (code not shown).

The log link assumes that a parameter's value is the exponentiation of the linear model.

- With some algebra, we can solve $\log(\lambda_i) = \alpha + \beta x_i$ for λ:

$$\lambda_i = e^{\alpha + beta * x_i}$$

giving us the inverse link.

The implication of using a log link is that the outcome scales exponentially with the predictor variable(s) (Fig. 3).

- A half-unit increase in x from -0.5 to 0 results in a y increase of ≈ 0.68. Another half unit increase in x from 0 to 0.5 increases y by 2.16. Yet another half-unit increase in x from 0.5 to 1 increases y by ≈ 6.82.

Another way to think about it is that an increase of one unit on the log scale increases the outcome by an order of magnitude on the untransformed scale.

- This is apparent from the widening intervals in the right plot of Fig. 3.

Exponential relationships grow exponentially, but most biological processes have finite upper limits.

- Therefore, we need to be careful when using log links because problems may arise when we try to predict outside of the range of the data used to fit our model.

The take-home message here is that in GLM land, no regression coefficient β ever produces a constant change on the response-variable scale. When we covered multiple regression, we defined interactions as when the effect of one predictor was dependent on another predictor.

In GLMs, every predictor interacts with itself—and thus with every other predictor—because the impact of each predictor depends upon that predictor's value before the change.

Mathematically, this can be shown by computing the rate of change in y for a given change in x. In a normally distributed model, the mean is modeled as:

$$\mu = \alpha + \beta x$$

so the rate of change in μ with respect to x is

$$\frac{\partial \mu}{\partial x} = \beta.$$

No matter what x is, the change is constant.

But consider a binomial probability p:

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

Taking the derivative of p with respect to x gives us:

$$\frac{\partial p}{\partial x} = \frac{\beta}{2(1 + \cosh(\alpha + \beta x))}.$$

The predictor x appears in the derivative, so the rate of change in x depends on the value of x—an interaction of x with itself!

Practically, this means that parameter estimates of GLMs do not—by themselves—tell you this importance of predictor variables on responses because each parameter represents a *relative* difference on the scale of the linear model, ignoring all other parameters.

- We want *absolute* differences in responses so we need to incorporate all parameters.

Also note that we cannot use WAIC or other information criteria to decide on a likelihood function. Luckily, the maximum entropy distribution is usually the best choice of likelihood function, so we don't have to do an immoderate amount of hand-wringing over our choices.