FES 524: Natural Resources Data Analysis

Reading 7.1: Generalized linear models

Contents

1 Introduction to generalized linear models

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:

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

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.

1.1 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.

1.1.1 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.

1.2 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,

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

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…).

Table 1: Components of a GLM

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

2 GLM structure and overview

The *conditional response variable*, $Y_i|\mu_i, \phi$, can be modeled as originating from a wide variety of distributions (anything that can be written as an EDM). 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 error distributions to describe the data generating process.

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 all 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 the desmos graph: https://www.desmos.com/calculator/vk2tqrxpk5. To recast this as an EDM, we reparameterize in terms of the mean and dispersion (Table **??**).