

# FES 524: Natural Resources Data Analysis

## Reading 7.1: Generalized linear models

### Contents

<b>1</b>	<b>Introduction to generalized linear models</b>	<b>1</b>
1.1	Generalizing to the exponential family . . . . .	1
1.2	The canonical link function . . . . .	2
<b>2</b>	<b>GLM structure and overview</b>	<b>3</b>

## 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  $i^{\text{th}}$  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  $i^{\text{th}}$  observation from the vector of random effects,  $\boldsymbol{\gamma}$ , and  $\epsilon_i$  is a random error from a normal distribution. Most often, we assume that  $\epsilon_1, \epsilon_2, \dots, \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ , but we also discussed some commonly used correlation structures for the errors last week.

An alternative way to write this is:

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

where  $\mu_i$  is the mean or expected value of the the  $i^{\text{th}}$  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

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

where  $Y$  is the random variable,  $\theta$  is the *canonical parameter*,  $\phi$  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  $\theta$  and  $\mu$  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  $\eta_i$  to be the linear predictor of the  $i^{\text{th}}$  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	$Y_i   \mu_i, \phi \sim f_Y(\mu_i, \phi)$

## 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).