Generalized linear models and exponential families

Patrick Breheny

March 3

Introduction

In the second half of this course, we will focus on modeling data which do not necessarily follow a $\mathrm{N}(\mu_i, \sigma^2)$ distribution, including:

- Outcomes with unequal variance
- Binary and categorical outcomes
- Discrete and count outcomes
- Outcomes with skewed distributions

This generalization does come at a cost, however – we can no longer derive closed form solutions for regression coefficients and inference is only approximate

Generalized linear models

- The basic structure of a generalized linear model (GLM) is as follows:

$$\begin{aligned} Y_i &\sim \text{some distribution with mean } \mu_i\text{, where} \\ g(\mu_i) &= \mathbf{x}_i^T\boldsymbol{\beta} \end{aligned}$$

- A GLM therefore consists of three components:
  - The *systematic component*, $\mathbf{x}_i^T\boldsymbol{\beta}$
  - The *random component*: the specified distribution for Y
  - The *link* function g

The systematic component

- Because the systematic component is specified in terms of $\mathbf{x}_i^T\boldsymbol{\beta}$, the general ideas and concepts that we have learned so far with respect to linear modeling carry over to generalized linear modeling
- This means that model specification and interpretation is the same, with the exception that we now have to think about the link and distribution of the outcome
- The quantity $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ is referred to as the *linear predictor* for observation i

The link

- In principle, g could be any function linking the linear predictor to the distribution of the outcome variable
- In practice, we also place the following restrictions on g
  - g must be smooth (*i.e.*, differentiable)
  - g must be monotonic (*i.e.*, invertible)

The random component

- Again in principle, we could specify any distribution for the outcome variable
- However, the mathematics of generalized linear models work out nicely only for a special class of distributions called the *exponential family* of distributions
- This is not as big a restriction as it sounds, however, as most common statistical distributions fall into this family, such as the normal, binomial, Poisson, gamma, and others

Example #1

- Most of today's lecture will involve working out the properties, terminology, and notation of exponential families, but before we do so, let's explore two examples of problems that can be cast into the GLM framework
- In the early stages of a disease epidemic, the rate at which new cases occur increases exponentially through time
- Thus, if μ_i is the expected number of new cases on day t_i, a model of the form

  $$\mu_i = \gamma \exp(\delta t_i)$$

  might be appropriate

Example #1 (cont'd)

- If we take the log of both sides,

$$\begin{aligned}\log(\mu_i) &= \log(\gamma) + \delta t_i \\ &= \beta_0 + \beta_1 t_i\end{aligned}$$

- Furthermore, since the outcome is a count, the Poisson distribution seems reasonable
- Thus, this model fits into the GLM framework with a Poisson outcome distribution, a log link, and a linear predictor of $\beta_0 + \beta_1 t_i$

Example #2

- The rate of capture of prey, y_i, by a hunting animal increases as the density of prey, x_i, increases, but will eventually level off as the predator has as much food as it can eat
- A suitable model is
$$\mu_i = \frac{\alpha x_i}{h + x_i}$$
- This model is not linear, but taking the reciprocal of both sides,
$$\begin{aligned}\frac{1}{\mu_i} &= \frac{h + x_i}{\alpha x_i} \\ &= \beta_0 + \beta_1 \frac{1}{x_i}\end{aligned}$$
- Because the variability in prey capture likely increases with the mean, we might use a GLM with a reciprocal link and a gamma distribution

Definition

- A distribution falls into the exponential family if its distribution function can be written as

$$f(y|\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right\},$$

  where the parameter of interest $\theta=h(\mu)$ depends on the expected value of y, ϕ is a scale parameter, and b and c are arbitrary functions
- This representation can be slightly generalized, but the above definition is sufficiently general for all commonly used GLMs
- As we will see, if a distribution can be written in this manner, maximum likelihood estimation and inference are greatly simplified and can be handled in a unified framework

Example: Poisson distribution

- To get a sense of how the exponential family works, let's work out the representation of a few common families, starting with the Poisson:

$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

- This can be rewritten as

$$f(y|\mu) = \exp\{y \log \mu - \mu - \log y!\},$$

thereby falling into the exponential family with $\theta = \log \mu$ and $b(\theta) = e^\theta$

- Note that the Poisson does not have a scale parameter ($\phi = 1$); for the Poisson distribution, the variance is determined entirely by the mean

Example: Normal distribution

Other distributions such as the normal, however, require a scale parameter:

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

which is in the exponential family with $\theta = \mu$, $b(\theta) = \frac{1}{2}\theta^2$, and $\phi = \sigma^2$

Example: Binomial distribution

- Finally, let's consider the binomial distribution with $n = 1$:

$$\begin{aligned} f(y|\mu) &= \mu^y (1-\mu)^{1-y} \\ &= \exp\left\{ y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu) \right\}, \end{aligned}$$

  which is in the exponential family with

$$\begin{aligned} \theta &= \log\left(\frac{\mu}{1-\mu}\right) \\ b(\theta) &= \log(1+e^\theta) \end{aligned}$$

- Note that, like the Poisson, the binomial distribution does not require a scale parameter
- The more general $n > 1$ case is also in the exponential family

Score statistic for exponential families

- What is so special about exponential families?
- Much of maximum likelihood estimation revolves around the derivative of the log-likelihood, called the *score*
- Consider the score for a distribution in the exponential family:

$$\begin{aligned} U &= \frac{\partial}{\partial \theta} \ell(\theta, \phi | y) \\ &= \frac{y - b'(\theta)}{\phi} \end{aligned}$$

Properties of the score statistic

- The score has the following properties, which you proved in Biometrics II:

$$\begin{aligned} \mathrm{E}(U) &= 0 \\ \mathrm{Var}(U) &= -\mathrm{E}(U') \end{aligned}$$

- Recall that the variance of U is also called the *information* and denoted J
- For the exponential family,

$$\mathrm{Var}(U) = \phi^{-1} b''(\theta)$$

Mean and variance for exponential families

- Thus, for the exponential family,

$$\begin{aligned}\mathrm{E}(Y) &= b'(\theta)\\ \mathrm{Var}(Y) &= \phi b''(\theta)\end{aligned}$$

- Note that the variance of Y depends on both the scale parameter and on a function of the mean (because θ is a function of μ), with b controlling the relationship between mean and variance
- Thus, if we write $b''(\theta)$ as a function of μ, with $V(\mu) = b''(\theta)$, we have

$$\begin{aligned}\mathrm{Var}(Y) &= \phi V(\mu)\\ \mathrm{Var}(U) &= \phi^{-1} V(\mu)\end{aligned}$$

The canonical link

- Although in principle, we can arbitrarily specify the distribution and link function g, note that if we choose $g = h$ (recall that h was defined as $\theta = h(\mu)$), then

$$\theta_i = h(\mu_i) = h(h^{-1}(\eta_i)) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- In other words, it ensures that the systematic component of our model is modeling the parameter of interest (sometimes called the *natural parameter*) in the distribution
- There is, therefore, a reason to prefer this link (the *canonical link*) when specifying the model

Benefits of canonical links

Although one is not required to use the canonical link, they tend to have nice properties, both statistically and in terms of mathematical convenience:

- They simplify the derivation of the MLE, as we will see in the next lecture
- They ensure that many properties of linear regression still hold, such as the fact that $\sum_i r_i = 0$
- They ensure that μ stays within the range of the outcome variable

Example: Binomial distribution

- As an example of this last point, consider the canonical link for the binomial distribution:

$$\begin{aligned} g(x) &= \log\left(\frac{x}{1-x}\right) \\ \mu &= g^{-1}(\eta) \\ &= \frac{e^\eta}{1+e^\eta} \end{aligned}$$

- As $\eta \to -\infty, \mu \to 0$, while as $\eta \to \infty, \mu \to 1$
- On the other hand, if we had chosen, say, the identity link, μ could lie below 0 or above 1, clearly impossible for the binomial distribution