Exponential families

Patrick Breheny

January 31

Introduction

- In our first lecture on the framework of generalized linear models, we remarked that the mathematics of generalized linear models "work out nicely only for a special class of distributions called the exponential family of distributions"
- Today we will see what the exponential family is, examine some special cases, and see what it is about members of the exponential family that makes them so attractive to work with

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 positive scale parameter, and b and c are arbitrary functions
- This representation can written different ways and can be slightly generalized, but the above definition is sufficient 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?
- As we have seen, maximum likelihood theory revolves around the score; consider, then, 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

- Recall from our previous lecture that the score has the following properties:

$$\mathrm{E}(U) = 0$$
$$\mathrm{Var}(U) = -\mathrm{E}(U');$$

also recall that the variance of U is referred to as the *information*

- For distributions in 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 (a constant) and on b, a function which controls the relationship between the mean and variance
- Thus, letting $\mu = b'(\theta)$ and writing $b''(\theta)$ as a function of μ with $W(\mu) = b''(\theta)$, we have

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

Examples

- For the normal distribution, $W(\mu) = 1$; the mean and the variance are not related
- For the Poisson distribution, $W(\mu) = \mu$; the variance increases with the mean
- For the Binomial distribution, $W(\mu) = \mu(1 - \mu)$; the variance is largest when $\mu = 1/2$ and decreases as μ approaches 0 or 1

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 have nice properties, both statistically and in terms of mathematical convenience:

- They simplify the derivation of the MLE, as we will see in a week or so
- They ensure that many properties of linear regression still hold, such as the fact that $\sum_i r_i = 0$
- They tend to 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:

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

- 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