Generalized linear models

Patrick Breheny

January 24

Introduction

- Previously, we discussed the topic of transforming the data to make linear regression assumptions hold
- Let us now take up the question of building models that do not make those assumptions in the first place – specifically, allowing distributions such as:
  - Outcomes with unequal variance
  - Binary and categorical outcomes
  - Discrete and count outcomes
  - Outcomes with skewed distributions

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

Linear regression

Thus, linear regression is one example of a GLM:

- Systematic component: $\mathbf{x}_i^T\boldsymbol{\beta}$
- Random component: $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$
- Link: $g(\mu) = \mu$, the *identity link*

Epidemic infection rates

- As a more interesting example, let's consider modeling the outbreak of disease cases in an epidemic
- In the early stages of an 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

Epidemic infection rates (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 $\beta_0 + \beta_1 t_i$

Predator-prey model

- 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

Summary

- This framework provides two important extensions of linear regression modeling: the ability to allow for nonlinear relationships between explanatory variables and the outcome, and the ability to allow non-normal distributions
- This generalization does come at a cost, however – as we will see, we can no longer derive closed form solutions for regression coefficients and inference is only approximate
- Estimation and inference regarding those regression coefficients is driven by a statistical idea known as *likelihood theory*; for the next two weeks we will be discussing likelihood theory and establishing results that will allow us to study the properties of GLMs