Intro to Generalized Linear Models

BIOS 6611

CU Anschutz

Week 15

Generalized Linear Models

Generalized Linear Models

Ordinary least squares estimation is specific to *linear* regression models. However, we are likely to encounter many other types of data:

- Binary/dichotomous (e.g., 1/0, yes/no)
- Categorical (e.g., green/yellow/red/orange/purple Skittles)
- Ordinal Categories (e.g., low/medium/high)
- Count (e.g., 0, 1, 2, 3, . . .)
- Rates (e.g., 1.3 per 1000 person years)

It is also possible that we have continuous data that does not meet our linear regression assumptions (e.g., linearity, normality, etc.).

In these cases, we can use a very flexible set of models known as **generalized linear models** (GLMs) originally proposed by Nelder and Wedderburn[1].

[1] Nelder, John Ashworth, and Robert WM Wedderburn. "Generalized linear models." *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972): 370-384.

Generalized Linear Models

Consider our linear regression model:

$$E(\mathbf{Y}|\mathbf{X}) \sim N(\mathbf{X}\beta, \sigma^2)$$

The *generalized* in GLM refers to:

- Dropping the normality requirement
- Relaxing the homoscedasticity assumption
- Allowing for some function of $E(Y)$ to be linear in the parameters (e.g., $g(\cdot)$) as a **link**

Specification of the model includes:

- The outcome Y and its distribution
- Covariates $\mathbf{X}$ and how they are linked to the mean of the outcome

We will focus on the linear regression context for GLMs for BIOS 6611, but you will expand this idea greatly next semester in BIOS 6612.

Exponential Families

Exponential Families (Definition)

GLMs are built on the idea of distributions within the class of **exponential families**. These are statistical distributions whose probability density functions can be written in a certain form:

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

- θ is the parameter of interest
- ϕ is a nuisance parameter and represents the dispersion (which is related to the variance)

We will break this down for the normal distribution. However, exponential families are extremely common and include normal, exponential, gamma, Poisson, binomial, and more!

Properties of Exponential Families

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

In this format, we can calculate both the expectation and variance of Y:

- $E(Y) = b'(\theta)$
  - $E(Y)$ depends only on the natural parameter, θ
- $Var(Y) = b''(\theta)a(\phi)$
  - $Var(Y)$ is a function of *both* θ and ϕ
  - In some distributions the variance does not depend on the mean (e.g., normal), others have a dependence between the mean and variance (e.g., binomial with $E(Y) = np$ and variance $Var(Y) = np(1-p)$)

Link Functions

With GLMs, we assume that a known function of the mean, $\mu_i = E(Y_i)$, is related linearly to its covariates $\mathbf{X}$:

$$g(\mu_i) = \mathbf{X}\beta$$

The function we specify, $g(\cdot)$, is referred to as the **link function**.

$g(\cdot)$ links the linear predictor (i.e., $\mathbf{X}\beta$ as our linear combination of the covariates) with the mean of the outcome.

We still assume that $Y_1, ..., Y_n$ are independent and our covariates $(X_1, ..., X_p)$ are fixed. However, the linearity assumption we are familiar with from linear regression now applies to $g(\mu_i)$, which need not equal $E(Y_i)$ (i.e., relaxing our assumption).

GLM Components

For any generalized linear model encountered, we need to specify three components:

1. **Random component**: Y is assumed to follow a distribution from the exponential family
2. **Systematic component**: linear predictor $\eta = \mathbf{X}\beta$
3. **Link function**: $g(\cdot)$:
   - connects $\mathbf{X}$ and μ
   - $g(\mu_i) = \eta_i$
   - should be an analytically tractable, invertible function such that $g^{-1}(\eta_i) = \mu_i$

Canonical Links

The link function, $g(\cdot)$, is **canonical** if $\theta_i = \eta_i$.

In practice, canonical links are preferred for parameter estimation and interpretation:

- The identity link for Gaussian outcomes gives linear regression
- The logit link for binary outcomes gives an *odds ratio* interpretation
- The log link for Poisson outcomes gives a *rate ratio* interpretation

Linear Regression as GLM

Normal Distribution as an Exponential Family

Let's put the normal distribution assuming $\theta = \mu$ and σ^2 is known ($\phi = \sigma^2$) into an exponential family form:

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

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

Normal Distribution as an Exponential Family

Let's put the normal distribution assuming $\theta = \mu$ and σ^2 is known ($\phi = \sigma^2$) into an exponential family form:

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

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

Normal Distribution Mean and Variance

From the previous slide, we have

$$t(y)\theta = y\mu$$
$$b(\theta) = \frac{1}{2}\mu^2$$
$$a(\phi) = \sigma^2$$

We can then calculate the expectation and variance:

$$E(Y) = b'(\theta) = \frac{\partial}{\partial\mu}\frac{1}{2}\mu^2 = \frac{2}{2}\mu = \mu$$

$$Var(Y) = b''(\theta)a(\phi) = \left(\frac{\partial}{\partial\mu}b'(\theta)\right)\sigma^2 = 1\sigma^2 = \sigma^2$$

Linear Regression with GLMs

For our standard linear regression model, we know that $E(Y) = \mathbf{X}\beta$. For the GLM we also know that our systematic component is $\eta_i = \mathbf{X}\beta$. Therefore, we have $E(Y) = \mu = \mathbf{X}\beta = \eta_i$.

For linear regression we use an *identity link function*, because it maps every element in our set to itself: $E(Y) = g(E(Y)) = g(\mu) = \mathbf{X}\beta = \eta$.

To obtain our estimates, $\hat{\beta}$, we will rely on maximum likelihood estimation. For GLMs generally, you will learn about *iteratively re-weighted least squares* (IRLS) and other algorithms that can be applied for any combination of random component, systematic component, and link function.

In the case of linear regression, we can actually derive the closed form estimates (which we did in the last lecture) to use for inference.