Unifying Theory of GLMs

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 5 of Roback and Legler text.

Computing set up

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(viridis)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "70%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Topics

- Identify the components common to all generalized linear models
- Find the canonical link based on the distribution of the response variable
- Properties of GLMs

Notes based on Chapter 5 Roback and Legler (2021) unless noted otherwise.

Unifying theory of GLMs

Many models; one family

We have studied models for a variety of response variables

- Least squares (Normal)
- Logistic (Bernoulli, Binomial, Multinomial)
- Log-linear (Poisson, Negative Binomial)

These models are all examples of **generalized linear models**.

GLMs have a similar structure for their likelihoods, MLEs, variances, so we can use a generalized approach to find the model estimates and associated uncertainty.

Components of a GLM

Nelder and Wedderburn (1972) defines a broad class of models called **generalized linear models** that generalizes multiple linear regression. GLMs are characterized by three components:

Components of a GLM

Nelder and Wedderburn (1972) defines a broad class of models called **generalized linear models** that generalizes multiple linear regression. GLMs are characterized by three components:

1. Response variable with parameter θ whose probability function can be written in exponential family form (**random component**)

Components of a GLM

Nelder and Wedderburn (1972) defines a broad class of models called **generalized linear models** that generalizes multiple linear regression. GLMs are characterized by three components:

1. Response variable with parameter θ whose probability function can be written in exponential family form (**random component**)
2. A linear combination of predictors, $\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ (**systematic component**)

Components of a GLM

Nelder and Wedderburn (1972) defines a broad class of models called **generalized linear models** that generalizes multiple linear regression. GLMs are characterized by three components:

1. Response variable with parameter θ whose probability function can be written in exponential family form (**random component**)
2. A linear combination of predictors, $\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ (**systematic component**)
3. A **link** function $g(\theta)$ that connects θ to η

One-parameter exponential family form

Suppose a probability (mass or density) function has a parameter θ. It is said to have a **one-parameter exponential family form** if

- The support (set of possible values) does not depend on θ, and
- The probability function can be written in the following form

$$f(y; \theta) = e^{[a(y)b(\theta) + c(\theta) + d(y)]}$$

Mean and variance

One-parameter exponential family form

$$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$$

Using this form:

$$E(Y) = -\frac{c'(\theta)}{b'(\theta)} \qquad Var(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

Poisson in one-parameter exponential family form

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0, 1, 2, \ldots, \infty$$

$$\begin{aligned} P(Y = y) &= e^{-\lambda} e^{y \log(\lambda)} e^{-\log(y!)} \\ &= e^{y \log(\lambda) - \lambda - \log(y!)} \end{aligned}$$

Recall the form: $f(y; \theta) = e^{[a(y)b(\theta) + c(\theta) + d(y)]}$, where the parameter $\theta = \lambda$ for the Poisson distribution

- $a(y) = y$
- $b(\lambda) = \log(\lambda)$
- $c(\lambda) = -\lambda$
- $d(y) = -\log(y!)$

Poisson in exponential family form

- The support for the Poisson distribution is $y = 0, 1, 2, \ldots, \infty$. This does not depend on the parameter λ.
- The probability mass function can be written in the form $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

The Poisson distribution can be written in one-parameter exponential family form.

Canonical link

Suppose there is a response variable Y from a distribution with parameter θ and a set of predictors that can be written as a linear combination $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

- A **link function**, $g()$, is a monotonic and differentiable function that connects θ to η
- When working with a member of the one-parameter exponential family, $b(\theta)$ is called the **canonical link**
- Most commonly used link function

Canonical link for Poisson

Recall the exponential family form:

$$P(Y = y) = e^{y \log(\lambda) - \lambda - \log(y!)}$$

then the canonical link is $b(\lambda) = \log(\lambda)$

GLM framework: Poisson response variable

1. Response variable with parameter θ whose probability function can be written in exponential family form

$$P(Y = y) = e^{y \log(\lambda) - \lambda - \log(y!)}$$

GLM framework: Poisson response variable

1. Response variable with parameter θ whose probability function can be written in exponential family form

$$P(Y = y) = e^{y \log(\lambda) - \lambda - \log(y!)}$$

2. A linear combination of predictors,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

GLM framework: Poisson response variable

1. Response variable with parameter θ whose probability function can be written in exponential family form

$$P(Y = y) = e^{y\log(\lambda) - \lambda - \log(y!)}$$

2. A linear combination of predictors,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

3. A function $g(\lambda)$ that connects λ and η

$$\log(\lambda) = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Activity: Generalized linear models

For your group's distribution

- Write the pmf or pdf in one-parameter exponential form.
- Describe an example of a setting where this random variable may be used.
- Identify the canonical link function.

Activity: Generalized linear models

Distributions

1. Exponential
2. Gamma (with fixed r)
3. Geometric
4. Binary

See BMLR - Section 3.6 for details on the distributions.

If your group finishes early, try completing the exercise for another distribution.

Using the exponential family form

The one-parameter exponential family form is utilized for

- Calculating MLEs of coefficients (recall iteratively re-weighted least squares)
- Inference for coefficients
- Likelihood ratio and drop-in-deviance tests

The specific calculations are beyond the scope of this course. See Section 4.6 of Dunn, Smyth, et al. (2018) for more detail (available at Duke library).

Exponential Distribution

Let Y $=$ time spent waiting for the first event in a Poisson process with an average rate λ events per time unit.

$$f(y, \lambda) = \lambda \exp\{-\lambda y\}.$$

Exponential Distribution

a. Write the pmf or pdf in one-parameter exponential form and
c. give the canonical link.

$$f(y, \lambda) = \lambda \exp\{-\lambda y\} \tag{1}$$
$$= \exp\{\log(\lambda \exp\{-\lambda y\})\} \tag{2}$$
$$= \exp\{\log \lambda - \lambda y\} \tag{3}$$

Recall the form: $f(y; \theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$, where the parameter $\theta = \lambda$ for the Exponential distribution.

- $a(y) = -y$
- $b(\lambda) = \lambda =$ canonical link
- $c(\lambda) = \log(\lambda)$
- $d(y) = 0$

Exponential Distribution

b. Example: The exponential distribution can be used to model the number of miles traveled until encountering the first pothole on a North Carolina road.

Gamma distribution (with fixed r)

Y $=$ time spent waiting for the rth event in a Poisson process with an average rate of λ events per unit of time.

$$f(y, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} \exp\{-\lambda y\}.$$

Gamma distribution (with fixed r)

a. Write the pmf or pdf in one-parameter exponential form and
c. give the canonical link.

$$
\begin{aligned}
f(y, \lambda) &= \frac{\lambda^r}{\Gamma(r)} y^{r-1} \exp\{-\lambda y\} && (4) \\
&= \exp\left[\log\left(\frac{\lambda^r}{\Gamma(r)} y^{r-1} \exp\{-\lambda y\}\right)\right] && (5) \\
&= \exp\left[r \log \lambda - \log(\Gamma(r)) + (r-1)\log y - \lambda y\right] && (6) \\
&\propto \alpha \exp\left[-\lambda y + r \log \lambda + (r-1)\log y\right] && (7)
\end{aligned}
$$

Recall the form: $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$, where the parameter $\theta = \lambda$ for the Gamma distribution.

- $a(y) = -y$
- $b(\lambda) = \lambda =$ canonical link
- $c(\lambda) = r\log(\lambda)$
- $d(y) = (r-1)\log y$

Gamma distribution (with fixed r)

b. Example: The gamma distribution can be used to model the number of miles traveled until encountering 10 potholes on a North Carolina road.

Geometric distribution

Y = number of failures before the first success in a Bernoulli process

$$f(y, p) = (1 - p)^y p.$$

Geometric distribution

a. Write the pmf or pdf in one-parameter exponential form and
c. give the canonical link.

$$f(y, \lambda) = (1 - p)^y p \tag{8}$$
$$= \exp\{\log[(1 - p)^y p]\} \tag{9}$$
$$= \exp\{y \log(1 - p) + \log(p)\} \tag{10}$$

Recall the form: $f(y; \theta) = e^{[a(y)b(\theta) + c(\theta) + d(y)]}$, where the parameter $\theta = \log(1 - p)$ for the Geometric distribution.

- $a(y) = y$
- $b(p) = \log(1 - p) = \text{canonical link}$
- $c(p) = \log(p)$
- $d(y) = 0$

Geometric distribution

b. Example: A geometric distribution can be used to model the number of random people you call who decline before someone agrees to complete a survey.

Binary distribution

$$f(y,p) = p^y(1-p)^{1-y}$$

Binary distribution

a. Write the pmf or pdf in one-parameter exponential form and
c. give the canonical link.

$$
\begin{aligned}
f(y,p) &= p^y(1-p)^{1-y} && (11)\\
&= \exp\{\log[p^y(1-p)^{1-y}]\} && (12)\\
&= \exp\{y\log p + (1-y)\log(1-p)\} && (13)\\
&= \exp\{y\log(\frac{p}{1-p}) + \log(1-p)\} && (14)
\end{aligned}
$$

Recall the form: $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$, where the parameter $\theta = \log(\frac{p}{1-p})$ for the Binary distribution.

- $a(y) = y$
- $b(p) = \log(\frac{p}{1-p}) =$ canonical link
- $c(p) = \log(1-p)$
- $d(y) = 0$

References

Dunn, Peter K, Gordon K Smyth, et al. 2018. *Generalized Linear Models with Examples in r*. Vol. 53. Springer.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society Series A: Statistics in Society* 135 (3): 370–84.

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.