Unifying theory of GLMs

02.14.22

STA 310

[Click here for PDF of slides]

Announcements

- Fill out mini-project 01 team evaluation by **Thu, Feb 17 at 11:59pm**
- Quiz 02: Wed, Feb 16 (after class) - Fri, Feb 18 at 11:59pm

STA 310

Quiz 02

- Open Feb 16 after class (check for announcement) and is due Fri, Feb 18 at 11:59pm
  - The quiz is not timed and will be administered in Gradescope.
- Covers content since Quiz 01
  - Jan 26 - Feb 16 lectures
  - BMLR Chapters 4 - 5 (Use Chapter 3 for reference)
- Fill in the blank, multiple choice, short answer questions
- Open book, open note, open internet (not crowd sourcing sites). You <u>cannot</u> discuss the quiz with anyone else. Please email me if you have questions.

STA 310

Learning goals

- Identify the components common to all generalized linear models
- Find the canonical link based on the distribution of the response variable
- Explain how coefficients are estimated using iteratively reweighted least squares (IWLS)

STA 310

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

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

2️⃣ A linear combination of predictors, $\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ (**systematic component**)

3️⃣ A **link** function $g(\theta)$ that connects θ to η

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

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

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

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 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

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

STA 310

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.

STA 310

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 $\eta = \sum_{j=1}^{p} \beta_j x_j = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

A **link function**, $g()$, is a monotonic and differentiable function that connects θ to η

The **canonical link** is a link function such that $g(\theta) = \eta$

- When working with a member of the one-parameter exponential family, the canonical link is $b(\theta)$

STA 310

Canonical link for Poisson

Recall

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

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

STA 310

GLM framework: Poisson response variable

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

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

2️⃣ A linear combination of predictors, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

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

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

STA 310

Activity: Identifying canonical link

For the distribution

- Describe an example of a setting where this random variable may be used.
- Identify the parameter.
- Write the pmf or pdf in one-parameter exponential form.
- Identify the canonical link function
- One person from each group: Write your response on the board.

Activity

Distributions

1. Binary
2. Exponential
3. Negative binomial (with fixed r)
4. Geometric
5. Normal (with fixed σ)

- If your group finishes early, try identifying the canonical link for the other distributions.
- See [BMLR - Section 3.6] for details on the distributions.

10:00

STA 310

Iteratively reweighted least squares (IWLS)

Data: Noisy Miners

The dataset `nminer` contains information about the number of noisy miners (small Australian bird) detected in two woodland patches within the Wimmera Plains of Victoria, Australia. It was obtained from the GLMsdata R package. We will use the following variables:

- **`Minerab`**: The number of noisy miners (abundance) observed in three 20 minute surveys
- **`Eucs`**: The number of eucalyptus trees in each 2 hectare area (about 4.94 acres)

STA 310

Noisy Miner Model

Eucs	Minerab
2	0
10	0
16	3
20	2
19	8

Our goal is to use a Poisson regression model to predict the number of noisy miners observed in three 20 minute surveys based on the number of eucalyptus trees.

$$\log(Minearab) = \beta_0 + \beta_1\ Euc$$

What are the best estimates of β_0 and β_1?

Iteratively reweighted least squares (IWLS)

- The estimates of β_0 and β_1 are found using maximum likelihood estimation.

- **Iteratively reweighted least-squares (IWLS)** is used to find the MLEs
  - Nelder and Wedderburn (1972) show that under certain specifications of the weights and a modified response variable, the estimates found using IWLS are equivalent to the MLEs.

STA 310

IWLS Set up

Working response: Modified response variable at each step of the iteration.

$$z_i = g(\theta) + g'(\theta)(y_i - \theta_i)$$

For Poisson regression, this is

$$z_i = \log(\lambda) + \frac{(y_i - \lambda_i)}{\lambda_i}$$

Working Weights: Weights applied to the observations at each step of the iteration

$$W_i = \frac{\theta^2}{Var(Y)} \quad \Rightarrow \quad W_i = \frac{\lambda^2}{\lambda} = \lambda \text{ for Poisson regression}$$

STA 310

IWLS procedure

1. Find initial starting values $\hat{\theta}_i$.

2. Calculate the working response values z_i.

3. Calculate the working weights W_i.

4. Find the coefficient estimates of the weighted least squares model.

$$z_i = \beta_0 + \beta_1 x \quad \textbf{with weights } W_i$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates for the model coefficients.

Use $\hat{\beta}_0$ and $\hat{\beta}_1$ to calculate updated values of $\hat{\theta}_i$ and repeat steps 2 - 5 until convergence.

STA 310

Demo in `lecture-11` repo

Acknowledgements

These slides are based on content in

- BMLR: Chapter 5 - Generalized Linear Models: A Unifying Theory
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.
- Generalized Linear Models with Examples in R
  - Chapter 5 - Generalized Linear Models: Structure
  - Chapter 6 - Generalized Linear Models: Estimation