Generalized Linear Models

Dr. Kiah Wah Ong

ILLINOIS TECH

Introduction

Up to now, we have discussed three types of regression models, namely:

linear, logistic and Poisson

These approaches share many common features that we would like to point out.

Common Features

- Notice that each approach uses predictors $X_1, \cdots, X_p$ to predict a response Y.
- We assume that Y belongs to a certain family of distributions.
  - Linear regression: Y follows a normal distribution.
  - Logistic regression: Y follows a Bernoulli distribution.
  - Poisson regression: Y follows a Poisson distribution.

Remark: The normal, Bernoulli and Poisson distributions are part of what we call an *exponential family*.

Common Features

- Each approach models the mean of Y as a function of the predictors.
  - In linear regression, the mean of Y takes the form

    $$E(Y|X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

    i.e. it is a linear function of the predictors.
  - In logistic regression, the mean takes the form

    $$\begin{aligned} E(Y|X_1, \cdots, X_p) &= \Pr(Y = 1|X_1, \cdots, X_p) \\ &= \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_P}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} \end{aligned}$$

  - In Poisson regression, the mean takes the form

    $$\begin{aligned} E(Y|X_1, \cdots, X_p) &= \lambda(X_1, \cdots, X_p) \\ &= e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p} \end{aligned}$$

Link Function

Also, notice that the equations we have seen below

$$E(Y|X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\ln\left(\frac{E(Y|X_1, \cdots, X_p)}{1 - E(Y|X_1, \cdots, X_p)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\ln(E(Y|X_1, \cdots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

can be expressed using a link function, η, such that if

$$\eta\left(E(Y|X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

then we have $\eta(\mu) = \mu$, $\eta(\mu) = \ln((\mu/1 - \mu))$, and $\eta(\mu) = \ln(\mu)$ for linear, logistic and Poisson respectively.

Note: Notice that with Y follows a Bernoulli distribution, we have $E(Y|X_1, \cdots, X_p) = \mathrm{Pr}(Y = 1|X_1, \cdots, X_p)$.

Generalized Linear Model

Any regression approach that has the features mentioned above is known as a *generalized linear model* (GLM).

Hence, linear regression, logistic regression, and Poisson regression are three examples of GLMs.

That is, GLM extend linear regression by allowing the response variable to have:

- a general distribution in exponential family, and
- a mean that depends on the predictors through a link function η, with

$$\eta(E(Y|X_1, \cdots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Exponential Family

Let us look into these so called exponential family.

Exponential family is a set of probability distributions that has pdfs of the form

$$f(x|\theta) = h(x)\exp(\theta T(x) - A(\theta))$$

Where θ is a parameter of the distribution. (Note: It can take a vector form.)

Exponential Family

For example: the probability density function for a Bernoulli r.v. with $\Pr(x = 1) = \pi$ and $\Pr(x = 0) = 1 - \pi$ is

$$\begin{aligned} f(x|\pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\} \end{aligned}$$

By comparing to

$$f(x|\theta) = h(x) \exp(\theta T(x) - A(\theta))$$

we see that

$$\begin{aligned} \theta &= \log\left(\frac{\pi}{1-\pi}\right) \\ T(x) &= x \\ A(\theta) &= -\log(1-\pi) = \log(1+e^\theta) \\ h(x) &= 1 \end{aligned}$$

Exponential Family

For example: the probability density function for a Poisson r.v. is given by

$$\begin{aligned} f(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!} \exp\{x \log \lambda - \lambda\} \end{aligned}$$

By comparing to

$$f(x|\theta) = h(x) \exp(\theta T(x) - A(\theta))$$

we see that

$$\begin{aligned} \theta &= \log \lambda \\ T(x) &= x \\ A(\theta) &= \lambda = e^{\theta} \\ h(x) &= \frac{1}{x!} \end{aligned}$$

Exponential Family

For example: the probability density function for a Normal r.v. is given by

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

By comparing to

$$f(x|\boldsymbol{\theta}) = h(x)\exp(\boldsymbol{\theta}^T\mathbf{T}(x) - A(\boldsymbol{\theta}))$$

we see that

$$\boldsymbol{\theta} = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}, \quad \mathbf{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$\begin{aligned} A(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) \\ h(x) &= \frac{1}{\sqrt{2\pi}} \end{aligned}$$

Generalized Linear Models

Response distribution	Link function	$\eta(\mu)$
Normal	Identity	μ
Bernoulli	Logit	$\log\left(\frac{\mu}{1-\mu}\right)$
Poisson	Log	$\log(\mu)$
Exponential	Inverse	$-1/\mu$
Gamma	Inverse	$-1/\mu$

There are other choices of response distribution in the exponential family which can be used to fit the generalized linear model.

For example, exponential distribution can be used to model survival time of patients in a clinical study as a function of age, gender, disease, type of treatment etc.

Generalized Linear Models

To use GLM model in *R* we use the following nomenclature:

$$\mathrm{glm}(\mathrm{formula}, \mathrm{family}, ...)$$

where the family argument can come from the following:

- $\mathrm{binomial}(\mathrm{link} = ''\mathrm{logit}'')$
- $\mathrm{gaussian}(\mathrm{link} = ''\mathrm{indentity}'')$
- $\mathrm{gamma}(\mathrm{link} = ''\mathrm{inverse}'')$
- $\mathrm{poisson}(\mathrm{link} = ''\mathrm{log}'')$

and so on.