Generalized Linear Models

response variable and one or more Predictor variables.

↳ used to model relationship b/w

GLM analysis comes into play when the error distribution is not normal

and /or

When a vector of non linear functions of the responses

$$\eta(Y) = (\eta(y_1), \eta(y_2), \ldots \eta(y_n))',$$

and no y itself,

has expectation $X\beta$.

Multiple Linear Regression:

$$Y = X\beta + \epsilon$$

$$E(Y) = X\beta \quad , \quad E(\epsilon) = 0$$

here

$$\boxed{E(\eta(y)) = X\beta.}$$

In GLM, the response variable distribution must be a member of the exponential Family.

The exponential Family of distribution.

A random variable u that belongs to the exponential family with a single parameter θ has a probability density function (pdf)

$$f(u, \theta) = s(u)\, t(\theta)\, e^{a(u) b(\theta)}.$$

Where
s, t, a, b are all known functions

Rewrite:

$$f(u, \theta) = \exp\{a(u)\, b(\theta) + d(u) + c(\theta)\}$$

Where

$$d(u) = \ln(s(u)) \quad c(\theta) = \ln(t(\theta))$$

Where $a(u) = u$, the distribution is said to be in Canonical form & $b(\theta)$ is called natural parameter.

Parameters other than the parameter of interest θ are called nuisance parameters.

Some members of the exponential family

1. **Normal distribution** $N(\mu, \sigma^2)$

pdf of normal distribution

$$f(u;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma^2}\right)}, \quad -\infty \leq u \geq \infty$$

$$= \exp\left\{ u \cdot \frac{\mu}{\sigma^2} + \left[\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2\right] - \frac{u^2}{2\sigma^2}\right\}$$

$$a(u) = u, \quad b(\theta) = \frac{\mu}{\cancel{\sigma^2}} = \mu.$$

$$c(\theta) = \frac{-\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2$$

$$d(u) = \frac{-u^2}{2\sigma^2}$$

2. <u>Binomial Distribution</u> $Bin(n, p)$:

p – parameter of interest.

n – nuisance parameter

$$f(u, p) = \binom{n}{u} p^u (1-p)^{n-u}, \quad u = 0, 1, \ldots n$$

$$= \binom{n}{u} \left(\frac{p}{1-p}\right)^u (1-p)^n$$

$$= \exp\left\{ u \ln\left(\frac{p}{1-p}\right) + n \ln(1-p) + \ln\binom{n}{u} \right\}$$

$a(u) = u$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$ = Natural Parameter

$c(\theta) = n \ln(1-p)$,

$d(u) = \ln\binom{n}{u}$

3. Poisson Distribution $P(\lambda)$

$$f(u,\lambda) = \frac{e^{-\lambda}\lambda^{u}}{u!}, \quad u = 0,1,\ldots$$

$$= \exp\{u \ln \lambda - \lambda - \ln u!\}$$

$a(u) = u$, $b(\theta) = \ln \lambda$ → Natural Parameter

$c(\theta) = -\lambda$

$d(u) = -\ln u!$

4 Gamma distribution

with parameter θ of interest & α as nuisance parameter.

$$f(u,\theta) = \frac{\theta^{\alpha} u^{\alpha-1} e^{-\theta u}}{\Gamma\alpha}, \quad \alpha,\theta > 0, u \geq 0$$

$$= \exp\{-\theta u + \alpha \ln \theta - \ln \Gamma\alpha) + (\alpha-1)\ln u\}$$

$a(u) = u$, $b(\theta) = -\theta$

$c(\theta) = \alpha \ln \theta - \ln \Gamma\alpha$

$d(u) = (\alpha-1)\ln u$

5. Exponential distribution

$$f(u,\theta) = \theta e^{-\theta u} \quad u \geq 0 \quad \theta > 0$$

$$= \exp\{-u\theta + \ln\theta\}.$$

$a(u) = u$ $\qquad b(\theta) = \theta \rightarrow$ natural Parameter

6. Negative Binomial distribution

The variable u is the no. of failures observed to attain r Successes in binomial trials with probability of Success θ

Probability mass function can be written as

$$f(u,\theta) = \binom{r+u-1}{r-1} \theta^r (1-\theta)^u, \quad u = 0,1,\ldots$$

$$= \exp\left\{u\ln(1-\theta) + r\ln\theta + \ln\binom{r+u-1}{r-1}\right\}$$

$a(u) = u$, $\quad b(\theta) = \ln(1-\theta) \rightarrow$ natural Parameter

$c(\theta) = r\ln\theta \qquad d(u) = \ln\binom{r+u-1}{r-1}$

Expected Value & Variance of $a(u)$

$$E(a(u)) = \frac{-c'(\theta)}{b'(\theta)} \quad \& \quad \cancel{V(a(u)) = b''(\theta)}$$

$$V(a(u)) = \frac{b''(\theta)\,c'(\theta) - c''(\theta)\,b'(\theta)}{[b'(\theta)]^3}$$

Example:

Binomial. $\cancel{a(u)(u)}$

$$a(u) = u$$

$$b(\theta) = \ln\left(\frac{p}{1-p}\right) \qquad c(\theta) = n\ln(1-p)$$

$$b'(\theta) = \frac{1}{p(1-p)} \qquad c'(\theta) = \frac{-n}{1-p}$$

$$b''(\theta) = \frac{2p-1}{[p(1-p)]^2} \qquad c''(\theta) = \frac{-n}{(1-p)^2}$$

$$E(a(u)) = E(u) = \frac{-c'(\theta)}{b'(\theta)} = \frac{n}{1-p} * p(1-p)$$

$$= np$$

$$V(a(u)) = V(u) = np(1-p)$$

Suppose we have a set of independent observations, response variable

$$(y_i, \tilde{x}_i), \quad i = 1(1)n, \quad \tilde{x}_i' = (x_{i1}, x_{i2}, \ldots x_{ip})$$

from some exponential type of distribution of canonical form

[ie $a(y) = y$]. Then the joint probability density function is

$$f(y_1, y_2 \ldots y_n, \theta, \phi) = \exp\left\{\sum_{i=1}^{n} y_i\, b(\theta) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right\}$$

$$= \exp\left\{\sum_{i=1}^{n} y_i\, b(\theta) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right\}$$

$$= \prod \exp\left\{y_i\, b(\theta_i) + c(\theta_i) + d(y_i)\right\}$$

where ϕ is a vector of nuisance parameters that occur with in $b(\cdot), c(\cdot)$ and $d(\cdot)$.

$\theta = (\theta_1, \theta_2, \ldots \theta_n)'$ vector of parameters of interest

The variance in response variable y_i can be explained in terms of $\hat{x}_i$ values

$$\tilde{x}_i = (x_{i1}, x_{i2}, \ldots x_{ip})'$$

Consider the set of parameters

$$\beta = (\beta_1, \beta_2, \ldots \beta_p)'$$

We find some suitable link function $g(\cdot)$ such that

$$g(\mu_i) = \tilde{x}_i' \beta$$

μ_i is $E(y_i)$

y_i follow binomial

$$\ln\left(\frac{P}{1-P}\right) = \tilde{x}_i' \beta$$

y_i is Normal $E(y_i) = \tilde{x}_i' \beta$

$$y_i = \tilde{x}_i' \beta + \epsilon$$

$$E(y_i) = \tilde{x}_i' \beta$$

A link function that is often regarded as sensible one is natural parameter.

GLM analysis comes into play when the error distribution is not normal but the error distribution / the response variable distribution must be a member of exponential family.

Given $(Y_i, \tilde{x}_i')$

We would hope that variance in Y_i or $E(Y_i)$ values could be explained in terms of the $\tilde{x}_i'$ values. We would hope that we could find suitable link function $g(\theta_i)$ such that the model

$$g(\theta_i) = \tilde{x}_i' \beta$$

held.

Where

$\beta = (\beta_1, \beta_2, \ldots \beta_p)'$ is vector of regression coefficient

Link Function is often the natural parameter.

Binomial(n, θ_i)

Natural Parameter $\ln \frac{\theta_i}{1-\theta_i}$

$$g(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$$

We go for model

$$g(\theta_i) = \tilde{x}_i' \beta$$

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \tilde{x}_i' \beta$$

If y_i from Normal distribution

$$y_i = \tilde{x}_i' \beta + \epsilon$$

$$\theta_i = E(y_i) = \tilde{x}_i' \beta$$

$$\theta_i = \tilde{x}_i' \beta$$

Natural parameter for normal distribution is θ

$$g(\theta) = \theta$$

$$\theta_i = \frac{\exp(\tilde{x}_i' \beta)}{1+\exp(\tilde{x}_i' \beta)}$$

$$E(y_i) = \frac{\exp(\tilde{x}_i' \beta)}{1+\exp(\tilde{x}_i' \beta)}$$ Model for Binomial distribution.

In detail about Binomial Distribution

Suppose we have data $(y_i, \tilde{x}_i')$ from a binomial distribution $Bin(n_i, P_i)$

The single observation y_i is of the form $\frac{r_i}{n_i}$, where r_i is the no of the successes in n_i trials, each having probability P_i of success. and $\tilde{x}_i' = (x_{i1}, x_{i2}, \ldots x_{ip})$ is a set of observations of p regressions associated with y_i.

Binomial distribution is a member of the exponential family

joint Pdf = $f(y_1, y_2, \ldots y_n)$

$$= \prod_{i=1}^{n} \binom{n_i}{y_i} (P_i)^{y_i} (1-P_i)^{n_i - y_i}$$

$$= \prod_{i=1}^{n} \exp\left\{ y_i \ln\left(\frac{P_i}{1-P_i}\right) + n_i \ln(1-P_i) + \ln\binom{n_i}{y_i} \right\}$$

$$= \exp\left\{ \sum_{i=1}^{n} y_i \ln\left(\frac{P_i}{1-P_i}\right) + \sum_{i=1}^{n} n_i \ln(1-P_i) + \sum_{i=1}^{n} \ln\binom{n_i}{y_i} \right\}$$

Given y_i $\tilde{x}_i'$ try to explain the variability of x_i in y_i

We would hope that the variation in the y_i | $E(y_i) = P_i$ $\quad y_i = \frac{r_i}{n_i}$

could be explained in terms of the $\tilde{x}_i'$ values,

the

i.e, we would hope that we could find a suitable link function $g(\cdot)$ such that

$$g(P_i) = \tilde{x}_i' \beta$$

Binomial distribution the

$$\text{Natural parameter} = \ln\left(\frac{P_i}{1-P_i}\right)$$

We fit the model

$$\ln \frac{P_i}{1-P_i} = \tilde{x}_i' \beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_p x_{ip}.$$

$$E(Y_i) = P_i = \frac{\exp(\tilde{x}_i' \beta)}{1+\exp(\tilde{x}_i' \beta)}$$

This is the model instead of fitting $y_i = \tilde{x}_i' \beta + \epsilon$

finally,

$$P_i = \frac{\exp(\tilde{x}_i' \beta)}{1+\exp(\tilde{x}_i' \beta)} \quad \text{———} \quad (*)$$

when $\tilde{x}_i' \beta = \beta_1 + \beta_2 x_{i2}$, if this is a situation (*) is called Logistic function.

Estimation via Maximum Likelihood

To estimate β, we use the method of maximum likelihood, Compute likelihood function L.

$$L = \exp\left\{ \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{i=1}^{n} \ln \binom{n_i}{y_i} \right\}$$

$$\ln L = \sum_{i=1}^{n} y_i \ln \frac{p_i}{1-p_i} + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{i=1}^{n} \ln \binom{n_i}{y_i}$$

fit the model $p_i = \dfrac{\exp(\tilde{x}_i' \beta)}{1+\exp(\tilde{x}_i' \beta)}$

Write loglikelihood in terms of β

$$= \sum_{i=1}^{n} y_i \tilde{x}_i' \beta - \sum_{i=1}^{n} n_i \ln\left(1+\exp(\tilde{x}_i' \beta)\right)$$

$$+ \sum_{i=1}^{n} \ln \binom{n_i}{y_i}$$

Maximize log likelihood $\ln L$ with respect to β

differentiate $\ln L$ in terms of β

It is not easy

Numerical Search Method / iteratively reweighted least square (IRLS) could be used to compute MLEs of β.

Example Pneumoconiosis Data

Number of Year of Exposure	No. of Severe Cases	Total no of miners	Proportion of severe Cases, y
5.8	0	98	0
15.0	1	54	0.0185
21.5	3	43	0.0698
27.5	8	48	0.01667
33.5	9	51	0.1765
39.5	8	38	0.2105
46.0	10	28	0.3571
51.5	5	11	0.4545

y: proportion of miners who have Severe Symptoms.

x_i No. of years of exposure

Probability distribution for the number of Severe cases is binomial

We will fit logistic regression model to the data:

$$P_i = E(y_i) = \frac{\exp(\tilde{x}_i'\beta)}{1+\exp(\tilde{x}_i'\beta)}$$

$$\tilde{x}_i'\beta = \beta_1 + \beta_2 x.$$

finally fitted Model

$$\hat{y}_i = \frac{\exp(4.79 - 0.0935x)}{1+\exp(4.79 - 0.0935x)}$$

Poisson distribution

Data $(y_i, \tilde{x}_i')$ from poisson $P(\mu_i)$,

$$E(y_i) = \mu_i$$

$$f(y, \mu) = \exp\{y \ln\mu - \mu - \ln y!\}$$

$\ln\mu$ is the natural parameter

The variation in y_i could be explained in terms of the $\tilde{x}_i'$ values.

We fit the model

$$g(\mu_i) = x_i' \beta$$

$$\ln \mu_i = \tilde{x}_i' \beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}$$

$$\boxed{y_i = e^{\tilde{x}_i' \beta} + \epsilon}$$

Choice of Link function

Distribution	Link function	Name
Normal (continuous data)	$g(\mu) = \mu$	identity link
Binomial (Binary (or) Proportion data)	$g(p) = \ln \frac{p}{1-p}$	logistic link
Poisson (for count data)	$g(\mu) = \ln \mu$	log link
Exponential	$g(\mu) = \frac{1}{\mu}$	reciprocal link
Gamma (positive Continuous data)	$g(\mu) = \frac{1}{\mu}$	reciprocal link

Link Function

→ It is the transformation that connects predicted values of the dependent variables to the observed values