Chapter 16
Generalized Linear Models

The usual linear regression model assumes a normal distribution of study variables whereas nonlinear logistic and Poison regressions are based on Bernoulli and Poisson distributions, respectively of study variables. Similar to as in logistic and Poisson regressions, the study variable can follow different probability distributions like exponential, gamma, inverse normal etc. One such family of distribution is described by the **exponential family of distributions**. The generalized linear model is based on this distribution and unifies linear and nonlinear regression models. It assumes that the distribution of the study variable is a member of the exponential family of distribution.

Exponential family of distribution

A random variable X belongs to the exponential family with a single parameter θ has a probability density function

$$f(X,\theta) = \exp\left[a(X)b(\theta) + c(\theta) + d(X)\right]$$

where $a(X), b(\theta), c(\theta)$ and $d(X)$ are all known function.

If $a(X) = X,$ the distribution is said to be in **canonical form**. The function $b(\theta)$ is called the **natural parameter** of the distribution. The parameter θ is of interest, and all other parameters which are not of interest are called **nuisance parameters**.

Example:

Normal distribution

$$\begin{aligned} f_x(x,\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]; -\infty < x < \infty; -\infty < \mu < \infty; \sigma^2 > 0 \\ &= \exp\left[x\left(\frac{\mu}{\sigma^2}\right) + \left(-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2\right) - \frac{x^2}{2\sigma^2}\right]. \end{aligned}$$

Here $a(x) = x,\ b(\theta) = \dfrac{\mu}{\sigma^2}.$

Binomial distribution

$$\begin{aligned} f(x,p) &= \binom{n}{x} p^x (1-p)^{n-x},\ 0<p<1,\ x=0,1,...,n \\ &= \exp\left[x\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\binom{n}{x}\right]. \end{aligned}$$

Here

$$a(x)=x,\ b(\theta)=\ln\left(\frac{p}{1-p}\right).$$

Expected values and variance of *a*(*X*):

The exponential family of distribution for a random variable X and parameter of interest θ is

$$f(X,\theta)=\exp\left[a(X)b(\theta)+c(\theta)+d(X)\right]$$
$$L=\ln f(X,\theta)=a(X)b(\theta)+c(\theta)+d(X).$$

Let $U=\dfrac{dL}{d\theta}$

then for any distribution

$$E(U)=0$$

$$Var(U)=E(U^2)=E(-U')$$

where $U'=\dfrac{dU}{d\theta}$. The function U is called **score** and $Var(U)$ is called **information.**

The log-likelihood function is

$$L=\ln\left[f(X,\theta)\right]=a(X)b(\theta)+c(\theta)+d(y)$$

and then

$$U=\frac{dL}{d\theta}=a(X)b'(\theta)+c'(\theta)$$

$$U'=\frac{d^2L}{d\theta^2}=a(X)b''(\theta)+c''(\theta)$$

where $b'(\theta)=\dfrac{db(\theta)}{d\theta}$, $b''(\theta)=\dfrac{d^2b(\theta)}{d\theta^2}$, $c'(\theta)=\dfrac{dc(\theta)}{d\theta}$ and $c''(\theta)=\dfrac{d^2c(\theta)}{d\theta^2}$.

Since $E(U)=0$, so

$$\begin{aligned} E(U) &= b'(\theta)E\left[a(X)\right]+c'(\theta) \\ 0 &= b'(\theta)E\left[a(X)\right]+c'(\theta) \\ \Rightarrow E\left[a(X)\right] &= -\frac{c'(\theta)}{b'(\theta)}. \end{aligned}$$

Since

$$\begin{aligned} Var(U) &= \left[b'(\theta)\right]^2 Var\left[a(X)\right], \\ E(-U') &= -b''(\theta)E\left[a(X)\right]-c''(\theta) \\ Var(U) &= E(-U') \end{aligned}$$

$$\begin{aligned} \Rightarrow Var\left[a(X)\right] &= \frac{-b''(\theta)E\left[a(X)\right]-c''(\theta)}{\left[b'(\theta)\right]^2} \\ &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\left[b'(\theta)\right]^3}. \end{aligned}$$

Now we consider two examples which illustrate how other distribution and their properties can be obtained as particular cases:

Example: Binomial distribution

Consider X follows a Binomial distribution with parameters n and π, i.e. $X \sim Bin(n,\pi)$. Then in the framework of an exponential class of family

$$\begin{aligned} f(x,\pi) &= \binom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= \exp\left[x\ln\pi - x\ln(1-\pi)+n\ln(1-\pi)+\ln\binom{n}{x}\right]. \end{aligned}$$

Here $a(x)=x,\ \theta=\pi,\ b(\theta)=\ln\dfrac{\pi}{1-\pi},\ c(\theta)=n\ln(1-\pi),\quad d(x)=\ln\dbinom{n}{x},$

$$L = \ln\ f(x,\pi) = x\ln\pi - x\ln(1-\pi)+n\ln(1-\pi)+\ln\binom{n}{x}.$$

It is the canonical form of $f(x,\pi)$ with natural parameter $\ln\pi$.

$$\begin{aligned} U = \frac{dL}{d\pi} &= \frac{x}{\pi} + \frac{x}{1-\pi} - \frac{n}{1-\pi} \\ &= \frac{x}{\pi(1-\pi)} - \frac{n}{1-\pi} \\ &= \frac{x - n\pi}{\pi(1-\pi)} \end{aligned}$$

$$E(U) = \frac{E(x) - n\pi}{\pi(1-\pi)}$$

$$\begin{aligned} &= \frac{n\pi - n\pi}{\pi(1-\pi)} \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(U) &= \frac{Var(x)}{\pi^2(1-\pi)^2} \\ &= \frac{n\pi(1-\pi)}{\pi^2(1-\pi)^2} \\ &= \frac{n}{\pi(1-\pi)^2} \\ E(-U') &= E\left[-\frac{(-n)}{\pi(1-\pi)^2}\right] \\ &= \frac{n}{\pi(1-\pi)} \\ b'(\theta) &= \frac{1}{\pi(1-\pi)} = b'(\pi) \\ b''(\theta) &= \frac{2\pi - 1}{\left[\pi(1-\pi)\right]^2} = b''(\pi) \\ c'(\theta) &= -\frac{n}{1-\pi} = c'(\pi). \\ c''(\theta) &= -\frac{n}{(1-\pi)^2} = c''(\pi). \end{aligned}$$

Thus

$$\begin{aligned} E\left[a(X)\right] &= E(X) = -\frac{c'(\pi)}{b'(\pi)} = \pi \\ Var\left[a(X)\right] &= Var(X) \\ &= \frac{b''(\pi)c'(\pi) - c''(\pi)b'(\pi)}{\left[b'(\pi)\right]^3} = n\pi(1-\pi). \end{aligned}$$

Example: Poisson distribution

Consider that the random variable X follows a Poisson distribution with parameter λ, i.e., $X \sim P(\lambda)$. Then

$$\begin{aligned} f(x,\lambda) &= \frac{\exp(-\lambda)\lambda^x}{x!} \\ &= \exp\left[x \ln \lambda - \lambda - \ln x!\right] \\ L = \ln f(x,\lambda) &= x \ln \lambda - \lambda - \ln x! \end{aligned}$$

It is the canonical form of $f(x,\lambda)$ and $\ln \lambda$ is the natural parameter. Here

$$a(X) = X,\ b(\theta) = \ln \lambda,\ c(\theta) = -\lambda,\ d(X) = -\ln X!$$

$$U = \frac{dL}{d\lambda} = \frac{x}{\lambda} - 1$$

$$\begin{aligned} E(U) &= \frac{E(x)}{\lambda} - 1 \\ &= \frac{\lambda}{\lambda} - 1 \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(U) &= \frac{Var(x)}{\lambda^2} \\ &= \frac{\lambda}{\lambda^2} \\ &= \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned} E(-U') &= E\left[-\left\{\frac{d}{d\lambda}\left(\frac{x}{\lambda} - 1\right)\right\}\right] \\ &= E\left(\frac{x}{\lambda^2}\right) \\ &= \frac{\lambda}{\lambda^2} \\ &= \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned} b'(\theta) &= \frac{1}{\lambda} = b'(\lambda) \\ b''(\theta) &= -\frac{1}{\lambda^2} = b''(\lambda) \\ c'(\theta) &= -1 = c'(\lambda) \\ c''(\theta) &= 0 = c''(\lambda) \end{aligned}$$

$$E\left[a(X)\right] = E(X) = -\frac{c'(\lambda)}{b'(\lambda)} = \lambda$$

$$\begin{aligned} Var\left[a(X)\right] &= \frac{b''(\lambda)c'(\lambda) - c''(\lambda)b'(\lambda)}{\left[b'(\lambda)\right]^3} \\ &= \frac{\frac{1}{\lambda^2} - 0}{\frac{1}{\lambda^3}} \\ &= \lambda. \end{aligned}$$

Linear predictors and link functions

The role of the generalized model is basically to unify various distributions of the study variable. This is accomplished by developing a linear model having an appropriate function of the expected value of the study variable.

Denoting η_i to be the **linear predictor** which relates to the expected value of the study variable, it is expressed as

$$\begin{aligned} \eta_i &= g\left[E(y_i)\right] \\ &= g(\mu_i) \\ &= x_i'\beta \end{aligned}$$

where $x_i'\beta = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}$.

Thus

$$\begin{aligned} E(y_i) &= g^{-1}(\eta_i) \\ &= g^{-1}(x_i'\beta) \end{aligned}$$

where g is the function called a **link function.**

Several choices of link functions are available. If

$$\eta_i = \theta_i$$

then η_i is the **canonical link**. The choice of θ_i and canonical link is related to the distribution of study variable, which in turn governs the appropriate, usually nonlinear regression models. The canonical link provides mathematical convenience in deriving the statistical properties of the model and compatibility with sensible conclusions on scientific grounds.

For example, is the case y has a

- normal distribution, the canonical link function is the **identity link** defined as $\eta_i = \mu_i$.
- Binomial distribution, then logistic regression is used, and the **logistic link** is used as a canonical link which is defined as $\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$.
- Poisson distribution, then **log link** is used as a canonical link which is given as $\eta_i = \ln \lambda$.
- Exponential and gamma distribution, then the canonical link function used is a **reciprocal link** given by $\eta_i = \frac{1}{\lambda_1}$.

Other types of link functions are

- **probit link** given as $\eta_i = \Phi^{-1}\left[E(y_i)\right]$ where Φ is the cumulative distribution function of $N(0,1)$ distribution.
- **Complementary log-log link given by**

$$\eta_i = \ln\left[\ln\left\{1-E(y_i)\right\}\right]$$

- **power family link**

$$\eta_i = \begin{cases} \left[E(y_i)\right]^{\lambda}, & \lambda \neq 0 \\ \ln\left[E(y_i)\right], & \lambda = 0 \end{cases}$$

which is based on power transformation, similar to Box-Cox transformation.

A link is preferable if it maps the range of μ_i onto the whole real line and provides a good empirical approximation. It should also carry a meaningful interpretation in case of real applications.

There are two components in any generalized linear model:

i) distribution of study variable and

ii) link function.

The choice of link function is like choosing an appropriate transformation on the study variable. The link function takes advantage of the natural distribution of the study variable. The incorrect choice of link function can give rise to incorrect statistical inferences.

Maximum likelihood estimation of GLM:

The least-squares method can not be directly applied when the study variable is not continuous. So we use the maximum likelihood estimation method in GLM, which has a close connection with iteratively weighted least squares method.

Given the data $(x_i, y_i),\ i = 1, 2, ..., n$ and y following exponential family of distribution, the joint p.d.f. is

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

where θ is the parameter of interest and ϕ is nuisance parameters. The θ and/or ϕ can be a vector also like $(\theta_1, \theta_2, ..., \theta_n)$ and/or $(\phi_1, \phi_2, ..., \phi_n)$ respectively.

Consider a smaller set of the parameter $\beta = (\beta_1, \beta_2, ..., \beta_k)'$ which relates some function $g(\mu_i)$ to μ_i. In case μ_i is $E(y_i)$ then $g(\mu_i)$ relates μ_i to a linear combinations of $\beta's$ via $g(\mu_i) = x_i'\beta.$

For example, if data on y_i and n_i such that $y_i \sim Bin(n_i, \pi_i),$ then $y_i = \frac{r_i}{n_i}$ is the number of successes is n_i trials where π_i is the probability of success. Then joint p.d.f. of all n data set is

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

Assuming that the variation in π_i is explained by x_i values, choose suitable link function $g(\pi_i) = x_i'\beta$. A sensible link function is log-odds as

$$g(\pi_i) = \ln\frac{\pi_i}{1-\pi_i}.$$

Now the objective is to fit a model

$$\ln\frac{\pi_i}{1-\pi_i} = x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

or equivalently

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}.$$

The general log-likelihood function is

$$L(\beta) = \ln f(y; \theta, \phi) = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i).$$

The log-likelihood function is numerically maximized for a given data set. Generally, the iteratively reweighted least squares method is used.

Suppose $\hat{\beta}$ is the final value obtained after optimization and is the maximum likelihood estimator of β, then asymptotically

$$E(\hat{\beta}) = \beta$$
$$V(\hat{\beta}) = a(\phi)(X'V^{-1}X)^{-1}$$

where V is a diagonal matrix formed by the variances of estimated parameters in the linear predictor, apart from $a(\phi)$. The covariance matrix can be estimated by replacing V by its estimate $\hat{V}$.

In GLM, the variance of y_i is not constant and so generalized least squares estimation is used to get more efficient estimators.

To conduct the test of hypothesis in GLM, the model deviance is used for testing the goodness of model fit. The difference in deviance of full and reduced models is used to decide for the subset model.

The Wald inference can be applied for testing hypothesis and confidence interval estimation about individual model parameters. The Wald statistic for testing the null hypothesis $H_0 : R\beta = r$ where R is $q \times (k+1)$ with $rank(R) = q$ is

$$W = (R\hat{\beta} - r)'\left[R(X'\hat{V}X)^{-1}R'\right]^{-1}(R\hat{\beta} - r).$$

The distribution of W under H_0 is χ^2 distribution with q degrees of freedom.

In particular, for $H_0 : \beta_j = \beta_0$, the test statistic is

$$Z = \sqrt{W} = \frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)}$$

which has $N(0,1)$ distribution under H_0 and $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. The confidence intervals can be constructed using the Wald test. For example, $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm Z_{\frac{\alpha}{2}}\, se(\hat{\beta}_j).$$

The likelihood ratio test comprise the maximized log-likelihood function between the full and reduced models. The reduced model is the full model under the null hypothesis.

The likelihood ratio test statistic is

$$-2(\hat{L}_{reduced} - \hat{L}_{full})$$

where $\hat{L}_{full}$ and $\hat{L}_{reduced}$ are the maximized likelihood functions under full and reduced models. The likelihood ratio test statistic has a χ^2-distribution with degrees of freedom equal to the difference in the degrees of freedom of full and reduced model.

Prediction and confidence interval with GLM

Suppose we want to estimate the mean response function at $x = x_0$. The estimate is given by

$$\hat{y}_0 = \hat{\mu}_0 = g^{-1}(x_o^{'}\beta)$$

where g is the associated link function.

It is understood that x_0 is expandable to model form if more terms, e.g., interaction forms, are to be accommodated in the linear predictor.

To find the confidence interval, the asymptotic covariance matrix of $\hat{\beta}$ given by $\Omega = a(\phi)(X'V'X)^{-1}$, is estimated as $\hat{\Omega}$. The asymptotic variance of linear predictor estimated at $x = x_0$ is

$$Var(\hat{\eta}_0) = Var(x_0^{'}\hat{\beta}) = x_0^{'}V(\hat{\beta})x_0 = x_0^{'}\Omega x_0$$

and its estimate is $x_0^{'}\hat{\Omega}x_0$. Then $100(1-\alpha)\%$ confidence interval on the true mean response at $x = x_0$ is

$$g^{-1}\left[x_0^{'}\hat{\beta} - Z_{\frac{\alpha}{2}}\, x_0^{'}\hat{\Omega}x_0\right] \le \mu(x_0) \le g^{-1}\left[x_0^{'}\hat{\beta} + Z_{\frac{\alpha}{2}}\, x_0^{'}\hat{\Omega}x_0\right].$$

This approach usually works in practice because $\hat{\beta}$ is the maximum likelihood estimate of β. So any function of $\hat{\beta}$ is also a maximum likelihood estimate. This method constructs the confidence interval in the space of linear predictor and transform back the interval into the original metric. The Wald method can also be used to derive the approximate confidence interval for the mean response.

Residual analysis is GLM

The usual approach of finding the residuals is adopted in case of GLM.

The i^{th} ordinary residual from GLM is

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\mu}_i. \end{aligned}$$

The residual analysis is generally performed in GLM using **deviance residuals** defined as

$$r_i = \sqrt{d_i}\ sign(y_i - \hat{y}_i)$$

where d_i is the contribution of i^{th} observation to the deviance.

We explain it in the context of logistic and Poisson regression. In the case of logistic regression

$$d_i = y_i \ln\left(\frac{y_i}{n_i \hat{\pi}_i}\right) + (n_i - y_i)\left[\frac{1-\frac{y_i}{n_i}}{1-\hat{\pi}_i}\right],\ i = 1, 2, ..., n$$

where $\hat{\pi}_i = \dfrac{1}{1+\exp(x_i'\beta)}$.

As fitting of data to the model becomes better, then $\hat{\pi}_i \equiv \dfrac{y_i}{n_i}$ and deviance residuals become smaller and close to zero.

In the case of Poisson regression,

$$d_i = y_i \ln\left(\frac{y_i}{\exp(x_i'\beta)}\right) - \left[y_i - \exp(x_i'\beta)\right],\ \ i = 1, 2, ..., n.$$

Here y_i and $\hat{y}_i = \exp(x_i'\hat{\beta})$ become close to each other as deviance residuals approach zero.

The behaviour of deviance residuals is like the behaviour of ordinary residuals as in standard normal linear regression model. The normal probability plot is obtained by plotting the deviance residuals on a normal probability scale versus fitted values. Usually, the fitted values are transformed to constant information scale before plotting, so

- $\hat{y}_i$ is used in case of usual regression with normal distribution,
- $2\sin^{-1}\sqrt{\hat{\pi}_i}$ is used in case of logistic regression.
- $2\sqrt{\hat{y}_i}$ is used in Poisson regression.
- $2\ln\hat{y}_i$ is used when the study variable has gamma distribution.