Generalized Linear Models

Zhe Gao, Junhao Zhu

School of Mathematics
Sun Yat-sen University

September 9, 2020

Table of Contents

Table of Contents

Gaussian linear model

Given a response Y and predictors X, the Gaussian linear model is

$$Y = \beta^T X + \varepsilon.$$

We assume ε has a normal distribution,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$$

then we have

$$Y|X, \theta \sim \mathcal{N}(\beta^T X, \sigma^2 \boldsymbol{I})$$

Gaussian linear model has property that

$$\begin{gathered} E(Y|X) = \beta^T X \\ var(Y|X) = \sigma^2 \boldsymbol{I}. \end{gathered}$$

Motivation of nonlinear models

The key properties of a linear model are that

$$E(Y|X) = \beta^T X$$
$$var(Y|X) \propto \boldsymbol{I}.$$

In some cases where these conditions are not met, we can transform Y so that the properties are satisfied.

However it is often difficult to find a transformation that simultaneously linearizes the mean and gives constant variance.

Generalized linear models(GLMs) are a class of nonlinear regression models that can be used in certain cases where linear models do not fit well.

Logistic regression

Logistic regression is a specific type of GLM. The response Y is binary and has Bernoulli distribution instead of Gaussian. We define

$$P(Y=1|X=x) = \frac{1}{1+\exp(-\beta^T x)}$$
$$P(Y=0|X=x) = \frac{1}{1+\exp(\beta^T x)}.$$

The logit function

$$\mathrm{logit}(x) = \log(\frac{x}{1-x})$$

maps the unit interval onto the real line.
Thus, the transformation of Logistic regression is logit function.

$$\mathrm{logit}(E(Y|X)) = \log(\frac{E(Y|X)}{1-E(Y|X)}) = \log(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta^T X.$$

Estimation for logistic regression

Assuming independent cases, the log-likelihood for logistic regression is

$$L(\beta|y,X)=\sum_{i:y_i=1}\beta^T x_i-\sum_i \log(1+\exp(\beta^T x_i)).$$

The gradient of the log-likelihood function (the score function) is

$$G(\beta|y,X)=\sum_i (y_i-\frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)})x_i.$$

The Hessian of the log-likelihood is

$$H(\beta|y,X)=-\sum_i \frac{\exp(\beta^T x_i)}{(1+\exp(\beta^T x_i))^2}x_i x_i^T.$$

The Hessian is strictly negative definite as long as the design matrix has independent columns. Then MLE is unique.

Poisson regression

Poisson regression is a specific type of GLM. The response Y has Poisson distribution. We define

The CDF of Poisson distribution is

$$P(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda), k = 0, 1, 2, \dots$$

A key property of the Poisson distribution is that the mean is equal to the variance.

To create a regression methodology based on the Poisson distribution, we can formulate a regression model in which $Y|X$ is Poisson, with mean and variance equal to $\lambda(x) = \exp(\beta^T X)$.

Thus transformation of Poisson regression is log function

$$\log(E(Y|X)) = \beta^T X.$$

Estimation for Poisson regression

Assuming independent cases, the log-likelihood for logistic regression is

$$L(\beta|y, X) = \sum_i y_i\beta^T x_i - \log(y_i!) - \exp(\beta^T x_i).$$

The gradient of the log-likelihood function (the score function) is

$$G(\beta|y, X) = \sum_i y_i x_i - x_i \exp(\beta^T x_i) = \sum_i (y_i - \exp(\beta^T x_i))x_i.$$

The Hessian of the log-likelihood is

$$H(\beta|y, X) = -\sum_i (y_i - \exp(\beta^T x_i))x_i x_i^T.$$

Compare three regression

We see that for all three types of regression models, the following equation is satisfied.

$$\sum_i (y_i - \beta^T x_i)xi = 0 \text{ in Gaussian regression}$$

$$\sum_i (y_i - \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)})xi = 0 \text{ in logistic regression}$$

$$\sum_i (y_i - \exp(\beta^T x_i))x_i = 0 \text{ in Poisson regression}$$

Let $\mu_i = E[y_i|x_i]$, the equation can be written as

$$\sum_i (y_i - \mu_i)x_i = 0.$$

This shows that the residuals are orthogonal to each covariate in all of these models, and that achieving this orthogonality characterizes the MLE.

Compare three regression

Let the mean function be $\mu(\beta)$, the variance function be $V(\mu)$ or $V(\beta)$

Table: Three regression model

Family	Mean	Variance	$d\mu/d\beta$
Gaussian	$\beta^T X$	$\sigma^2\mathcal{I}$	X
Binomial	$\frac{1}{1+\exp(-\beta^T X)})$	$\frac{1}{2+\exp(-\beta^T X)+\exp(\beta^T X))}$	$\frac{-X}{2+\exp(-\beta^T X)+\exp(\beta^T X))}$
Poisson	$exp(\beta^T X)$	$exp(\beta^T X)$	$exp(\beta^T X)X$

We can see $d\mu/d\beta$ is proportional to $V(\beta)X$.

The estimating equations can be written as:

$$\sum_i \partial\mu_i/\partial\beta_i(yi - \mu_i(\beta))/V_i(\beta) = 0$$

Generalized linear model

A GLM is based on the following conditions:

- The y_i are conditionally independent given X.
- Assume Y follow a distribution from an exponential family, then pdf is

$$p(y_i|\theta_i, \sigma^2) = \exp(\frac{y_i\theta_i - A(\theta_i)}{\sigma^2} + c(y_i, \sigma^2)).$$

  where σ^2 is the dispersion parameter (often set to 1), $\theta_i = g(\beta^T x_i)$ is the natural parameter with an unknown vector of regression slopes β, A is the partition function, and c is a normalization constant.
- The homogeneity of variance does NOT need to be satisfied.
- Errors need to be independent but NOT normally distributed.

Generalized linear model

The log-likelihood function is

$$L(\beta, \sigma^2 | Y, X) = \sum_i \frac{y_i \theta_i - A(\theta_i)}{\sigma^2} + c(y_i, \sigma^2)$$

The score function with respect to θ_i is

$$\frac{y_i - A^{'}(\theta_i)}{\sigma^2}$$

To get expected value of the score function, we rewrite the score function

$$\frac{\partial}{\partial\theta} \log f_\theta(y) = f_\theta(y)^{-1} \frac{\partial}{\partial\theta} f_\theta(y).$$

where $f_\theta(y)$ is a density in y with parameter θ. The expected value is

$$\begin{aligned} E\frac{\partial}{\partial\theta} \log f_\theta(y) &= \int f_\theta(y)^{-1} (\frac{\partial}{\partial\theta} f_\theta(y)) f_\theta(y) dy \\ &= \frac{\partial}{\partial\theta} \int f_\theta(y) dy = 0. \end{aligned}$$

Generalized linear model

For the score function we have

$$E(\frac{y_i - A^{'}(\theta_i)}{\sigma^2}|X) = 0$$

$$E(y_i|x_i) = A^{'}(g(\beta^T x_i)).$$

Let $T^{-1}(\cdot) = A^{'}(g(\cdot))$, function $T(\cdot)$ is called link function, it maps $\mu = E[Y|X]$ to $\beta^T X$.

Generalized linear model

The Hessian is

$$\frac{\partial}{\partial\theta\theta^T}\log f_\theta(y) = f_\theta(y)^{-2}(f_\theta(y)\frac{\partial}{\partial\theta\theta^T}f_\theta(y) - \frac{\partial f_\theta(y)}{\partial\theta}\frac{\partial f_\theta(y)}{\partial\theta^T})$$

The expected value of the Hessian is

$$\begin{aligned} E\frac{\partial}{\partial\theta\theta^T}\log f_\theta(y) &= \frac{\partial}{\partial\theta\theta^T}\int f_\theta(y)dy - \int(\frac{1}{f_\theta(y)}\frac{\partial f_\theta(y)}{\partial\theta}\frac{\partial f_\theta(y)}{\partial\theta^T})f_\theta(y)dy \\ &= -\mathrm{cov}(\frac{\partial}{\partial\theta}\log f_\theta(y)|X). \end{aligned}$$

Thus,

$$var(\frac{y_i - A^{'}(\theta_i)}{\sigma^2}|X) = -E\frac{\partial^2}{\partial\theta_i^2}\log f_\theta(y) = A^{''}(\theta_i)\sigma^2.$$

Some example of GLM - logistic regression

Recall the logistic regression and Poisson regression above, we can set

$$A(\theta) = \log(1 + \exp(\theta)), \quad g(x) = x, \quad \sigma^2 = 1, \quad c(y_i, \sigma^2) = 1.$$

we obtain the logistic regression, then the link function is $logit(x)$. The mean and variance are

$$\begin{aligned} E(y|X) &= A^{'}(g(\beta^T X)) = \frac{1}{1 + \exp(\beta^T X)} \\ var(y|X) &= A^{''}(\theta)\sigma^2 = \frac{1}{2 + \exp(-\beta^T X) + \exp(\beta^T X))}. \end{aligned}$$

If we consider a new type of function Φ, then CDF of normal distribution. We can obtain another GLM - Probit distribution.

Some example of GLM - Poisson regression

If we can set

$$A(\theta) = \exp(\theta), \quad g(x) = x, \quad \sigma^2 = 1, \quad c(y_i, \sigma^2) = -\log(y_i!).$$

we obtain the Poisson regression, then the link function is $\log(x)$. The mean and variance are

$$E(y|X) = A^{'}(g(\beta^T X)) = \exp(\beta^T X)$$
$$var(y|X) = A^{''}(\theta)\sigma^2 = \exp(\beta^T X).$$

Some example of GLM - Negative binomial regression

When the response has a negative binomial distribution

$$P(y_i = y|x) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y}.$$

The mean is μ_i, and the variance is $\mu_i + \alpha\mu_i^2$. If $\alpha = 0$, we get Poisson distribution. The log-likelihood is

$$\log P(y_i = y|x_i) \propto y\log(\frac{\alpha\mu_i}{1+\alpha\mu_i}) - \alpha^{-1}\log(1+\alpha\mu_i).$$

Similar to Poisson regression, we set the mean $\mu_i = \exp(\beta^T x_i)$, $g(x) = \log(\alpha) + x - log(1 + \alpha\exp(x))$, $A(\theta) = -\frac{\log(1-\exp(\theta))}{\alpha}$. We get GLM.

Table of Contents

1. Generalized Linear Models

2. Quasi-likelihood

Motivation

In GLM, the parameters can be estimated by solving

$$\sum_i \partial\mu_i/\partial\beta_i(yi - \mu_i(\beta))/V_i(\beta) = 0$$

In GLM, we often specify $V_i(\beta)$ up to a constant, such as 1. If $V_i(\beta)$ is not a constant, we may specify a large class of regression models.

Why we use quasi-likelihood

Based on the equation above, we can give the estimate only by $\mu(\beta)$ and $V(\beta)$, the first two moments. We don't need to determine a distribution of the response.

In real data, sometimes it's difficult to give out the distribution, but determine the moment is more easy.

For example, we can get quasi-Poisson regression by specifying $\mu_i(\beta) = \exp(\beta^T x_i)$ and $V_i(\beta) = \mu_i(\beta)$. This formulation of quasi-Poisson regression never refers to the Poisson distribution directly, it only depends on moments.

Quasi-likelihood function

Suppose we have a vector of independent responses Y, we specify the first two moments of the data:

$$E(Y|\beta) = \mu(\beta);$$
$$\mathrm{cov}(Y|\beta) = \alpha V(\mu(\beta));$$

where μ is a function of regression parameters β, $V(\mu)$ is made up of known functions. Since the response is independent, $V(\mu)$ is diagonal. The quasi-likelihood function $K(Y,\mu)$ is defined as:

$$U = \frac{\partial K(Y,\mu)}{\partial \mu} = \frac{Y-\mu}{\alpha V(\mu)}$$

or equivalently

$$K(Y,\mu) = \int_{y}^{\mu} \frac{Y-t}{\alpha V(t)} dt.$$

U is the score function.

Quasi-likelihood function

The function U has several properties in common with the log-likelihood derivative. In particular,

$$\begin{aligned} E(U) &= 0 \\ Var(U) &= \frac{1}{\alpha V(\mu)} \\ -E(\frac{\partial U}{\partial \mu}) &= \frac{1}{\alpha V(\mu)}. \end{aligned}$$

Some example of quasi-likelihood

If we set the variance function to be 1, then $U = \frac{Y-\mu}{\alpha}$, quasi-likelihood is

$$K(y,\mu) = \int_y^\mu \frac{Y-t}{\alpha} dt = -\frac{(y-\mu)^2}{\alpha}.$$

$K(y,\mu)$ is the same as log-likelihood of normal distribution.
If we set the variance function to be μ, then $U = \frac{Y-\mu}{\alpha\mu}$, quasi-likelihood is

$$K(y,\mu) = \int_y^\mu \frac{Y-t}{\alpha\mu} dt = \frac{1}{\alpha}[y\log(\mu) - \mu + c].$$

$K(y,\mu)$ is the same as log-likelihood of Poisson distribution.

The word "quasi" refers to the fact that the score may or not correspond to a probability function. For example, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution.

Quasi-likelihood and exponential family

The following theorem shows that the log likelihood function is identical to the quasi-likelihood if and only if this family is an exponential family.

Theorem

For one observation of Y, the log likelihood function L has the property

$$\frac{\partial L}{\partial \mu} = \frac{Y - \mu}{V(\mu)}$$

where $\mu = E(Y)$, $V(\mu) = var(Y)$, if and only if the density of Y with respect to some measure can be written in the form $\exp(y\theta - g(\theta))$, where θ is some function of μ.

Estimation of quasi-likelihood

The quasi-likelihood estimating equations for the parameters β are obtained by differentiating the function U,

$$D^T \frac{Y-t}{V(t)} = 0.$$

where D is a matrix with element $\frac{\partial \mu_i}{\partial \beta_r}$.

The covariance matrix is

$$\frac{D^T V^{-1} D}{\alpha}.$$

This matrix plays the same role as the Fisher information for likelihood functions. Then

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \frac{D^T V^{-1} D}{n\alpha}).$$

Estimation of quasi-likelihood

The scale parameter is usually estimated in a separate step, after the regression parameters are estimated. A common approach is to use

$$\hat{\alpha} = \frac{\sum_i (y_i - \hat{\mu}_i)/\hat{V}_i}{n - p}.$$

Overdispersion

In statistics, overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model.

One mechanism that may give rise to overdispersion is heterogeneity. Suppose there exists a binary covariate, Z_i, and that

$$\begin{aligned} Y_i|Zi = 0 &\sim Poisson(\lambda_0) \\ Y_i|Zi = 1 &\sim Poisson(\lambda_1) \\ P(Z_i = 1) &= \pi \\ E(Y_i) &= \pi\lambda_1 + (1-\pi)\lambda_0 = \mu \\ var(Y_i) &= \mu + (\lambda_1 - \lambda_0)^2\pi(1-\pi). \end{aligned}$$

Therefore, if we do not observe Z_i then this omitted factor leads to increased variation.

Overdispersion-Poisson

Poisson regression analysis is commonly used to model count data. If overdispersion is a feature, an alternative model with additional free parameters may provide a better fit.

We can use a Poisson mixture model, in which the mean of the Poisson distribution can itself be thought of as a random variable drawn – in this case – from the gamma distribution thereby introducing an additional free parameter.

$$Y|X \sim Poisson(\lambda)$$
$$\lambda \sim \Gamma(a, b).$$

Overdispersion-Binomial

It has been observed that the number of boys born to families does not conform faithfully to a binomial distribution as might be expected. Instead, the sex ratios of families seem to skew toward either boys or girls. Therefore, empirical variance is larger than specified by a binomial model.

In this case, we can use the beta-binomial model, in which the probability parameter as a beta distribution.

$$\begin{aligned} Y|X &\sim Binomial(p) \\ p &\sim beta(a, b). \end{aligned}$$

Adjusting for Overdispersion

The most popular method for adjusting for overdispersion comes from the theory of quasilikelihood.

In quasilikelihood, after specify the mean function, we need to determines the relationship between the variance of the response variable and its mean. For a Poisson model, the variance function is μ.

To account for overdispersion, we will include another factor α called the "scale parameter," so that

$$V = \alpha\mu.$$

If $\alpha \neq 1$ then the model is not binomial; $\alpha > 1$ is called "overdispersion" and $\alpha < 1$ is called "underdispersion."