Notes: Generalized Additive Models – Ch3 Generalized Linear Models (GLM)

Yingbo Li

03/21/2021

Table of Contents

GLM overview

- In a GLM, a smooth monotonic link function $g(\cdot)$ connects the expectation $\mu_i = E(Y_i)$ with the linear combination of $\mathbf{X}_i$,

$$g(\mu_i) = \eta_i = \mathbf{X}_i \boldsymbol{\beta} \tag{1}$$

- In a generalized linear mixed model (GLMM), we have

$$g(\mu_i) = \eta_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}, \quad \mathbf{b} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi})$$

Exponential family of distributions

- The density function for an exponential family distribution

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \qquad (2)$$

  - a, b, c: arbitrary functions
  - ϕ: an arbitrary scale parameter
  - θ: the canonical parameter; completely depend on the model parameter β

- Properties about exponential family mean and variance

$$E(Y) = b'(\theta)$$
$$var(Y) = b''(\theta)a(\phi)$$

  - In most practical cases, $a(\phi) = \phi/\omega$ where ω is a known constant
  - We define a function

$$V(\mu) = b''(\theta)/w, \quad \text{so that } var(Y) = V(\mu)\phi$$

Exponential family examples

	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$	$\frac{\mu^y\exp(-\mu)}{y!}$	$\binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}$	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1}\exp\left(-\frac{\nu y}{\mu}\right)$	$\sqrt{\frac{\gamma}{2\pi y^3}}\exp\left\{\frac{-\gamma(y-\mu)^2}{2\mu^2 y}\right\}$
Range	$-\infty < y < \infty$	$y = 0, 1, 2, \ldots$	$y = 0, 1, \ldots, n$	$y > 0$	$y > 0$
θ	μ	$\log(\mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$
ϕ	σ^2	1	1	$\frac{1}{\nu}$	$\frac{1}{\gamma}$
$a(\phi)$	$\phi(=\sigma^2)$	$\phi(=1)$	$\phi(=1)$	$\phi\left(=\frac{1}{\nu}\right)$	$\phi\left(=\frac{1}{\gamma}\right)$
$b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$n\log\left(1+e^\theta\right)$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
$c(y,\phi)$	$-\frac{1}{2}\left\{\frac{y^2}{\phi}+\log(2\pi\phi)\right\}$	$-\log(y!)$	$\log\binom{n}{y}$	$\nu\log(\nu y)-\log\{y\Gamma(\nu)\}$	$-\frac{1}{2}\left\{\log(2\pi y^3\phi)+\frac{1}{\phi y}\right\}$
$V(\mu)$	1	μ	$\mu(1-\mu/n)$	μ^2	μ^3
$g_c(\mu)$	μ	$\log(\mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
$D(y,\hat{\mu})$	$(y-\hat{\mu})^2$	$2y\log\left(\frac{y}{\hat{\mu}}\right)-2(y-\hat{\mu})$	$2\left\{y\log\left(\frac{y}{\hat{\mu}}\right)+(n-y)\log\left(\frac{n-y}{n-\hat{\mu}}\right)\right\}$	$2\left\{\frac{y-\hat{\mu}}{\hat{\mu}}-\log\left(\frac{y}{\hat{\mu}}\right)\right\}$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}^2 y}$

Table 3.1 *Some exponential family distributions. Note that when $y = 0$, $y\log(y/\hat{\mu})$ is taken to be zero (its limit as $y \to 0$).*

Fitting GLMs

- For the GLM model (1) and (2), assuming $a_i(\phi) = \phi/\omega_i$, the log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \omega_i \left[y_i \theta_i - b_i(\theta_i) \right] / \phi + c_i(\phi, y_i)$$

- To optimize, we use the Newton's method, which is an iterative optimization approach

$$\theta^{(t+1)} = \theta^{(t)} - \left(\nabla^2 l \right)^{-1} \nabla l$$

  - Where both $\nabla^2 l$ and ∇l are evaluated at the current iteration $\theta^{(t)}$
  - Alternatively, we can use the Fisher scoring variant of the Newton's method, by replacing the Hessian matrix with its expectation

- Next, we will need to compute the gradient vector and expected Hessian matrix of l

Compute the gradient vector and expected Hessian of l

- By the chain rule,

$$\begin{aligned}\frac{\partial \theta_i}{\partial \beta_j} &= \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{1}{b''(\theta_i)} \cdot \frac{1}{g'(\mu_i)} \cdot X_{ij}\end{aligned}$$

- Therefore, the gradient vector of l is

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \left[y_i - b_i'(\theta_i) \right] \frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i) V(\mu_i)} X_{ij}$$

- The expected Hessian (expectation taken wrt Y) is

$$E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) = -\frac{1}{\phi} \sum_{i=1}^{n} \frac{X_{ij} X_{ik}}{g'(\mu_i)^2 V(\mu_i)}$$

The Fisher scoring update

- Define the matrices

$$\mathbf{W} = \text{diag}\{w_i\}, \quad w_i = \frac{1}{g'(\mu_i)^2 V(\mu_i)} \tag{3}$$

$$\mathbf{G} = \text{diag}\left\{g'(\mu_i)\right\} \tag{4}$$

- The Fisher scoring update for β is

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\underbrace{\left[\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) + \mathbf{X}\boldsymbol{\beta}^{(t)}\right]}_{\mathbf{z}}\end{aligned}$$

Iteratively re-weighted least square (IRLS) algorithm

1. Initialization:

$$\hat{\mu}_i = y_i + \delta_i, \quad \hat{\eta}_i = g\left(\hat{\mu}_i\right)$$

   - δ_i is usually zero, but may be a small constant ensuring $\hat{\eta}_i$ is finite

2. Compute pseudo data z_i and iterative weights w_i:

$$z_i = g'\left(\hat{\mu}_i\right)\left(y_i - \hat{\mu}_i\right) + \hat{\eta}_i$$
$$w_i = \frac{1}{g'\left(\hat{\mu}_i\right)^2 V\left(\hat{\mu}_i\right)}$$

3. Find $\hat{\beta}$ by minimizing the weighted least squares objective

$$\sum_{i=1}^{n} w_i \left(z_i - \mathbf{X}_i \boldsymbol{\beta}\right)^2$$

   then update

$$\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \quad \hat{\mu}_i = g^{-1}\left(\hat{\eta}_i\right)$$

- Repeat Step 2-3 until the change in deviance is near zero

IRLS example 1: logistic regression

- For logistic regression,

$$
\begin{aligned}
g(\mu) &= \log\left(\frac{\mu}{1-\mu}\right), \quad & g'(\mu) &= \frac{1}{\mu(1-\mu)} \\
V(\mu) &= \mu(1-\mu), & \phi &= 1
\end{aligned}
$$

- Therefore, in Step 2 of IRLS,

$$
\begin{aligned}
z_i &= \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i\left(1-\hat{\mu}_i\right)} + \hat{\eta}_i \\
w_i &= \hat{\mu}_i\left(1-\hat{\mu}_i\right)
\end{aligned}
$$

IRLS example 2: GLM with independent normal priors

- Assume that the vector β has independent normal priors

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0}, \frac{\phi}{\lambda}\mathbf{I}_p\right)$$

- Log posterior density (we still call it l, with some abuse of notation)

$$l(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}\omega_i\left[y_i\theta_i - b_i(\theta_i)\right] - \frac{\lambda}{2\phi}\boldsymbol{\beta}^T\boldsymbol{\beta} + \text{const}$$

- Gradient vector and expected Hessian matrix (wrt β)

$$\nabla l = \frac{1}{\phi}\left[\mathbf{X}^T\mathbf{W}\mathbf{G}(\mathbf{y} - \boldsymbol{\mu}) - \lambda\boldsymbol{\beta}\right]$$

$$E\left(\nabla^2 l\right) = -\frac{1}{\phi}\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I}_p\right)$$

  - Here, $\mathbf{W}$ and $\mathbf{G}$ are the same as in Equation (3) and (4)

- IRLS for GLM with independent normal priors

$$
\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\left[\mathbf{X}^T\mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) - \lambda\boldsymbol{\beta}^{(t)}\right] \\
&= \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{W}\underbrace{\left[\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) + \mathbf{X}\boldsymbol{\beta}^{(t)}\right]}_{\mathbf{z}} \qquad (5)
\end{aligned}
$$

Large sample distribution of $\hat{\beta}$

- Hessian of the negative log likelihood (also called observed information)

$$\hat{\mathcal{I}} = \frac{\mathbf{X}^T\mathbf{W}\mathbf{X}}{\phi}$$

- Fisher information, also called expected information

$$\mathcal{I} = E\left(\hat{\mathcal{I}}\right)$$

- Asymptotic normality the MLE $\hat{\beta}$

$$\hat{\beta} \sim \mathsf{N}\left(\beta, \mathcal{I}^{-1}\right) \quad \text{or} \quad \hat{\beta} \sim \mathsf{N}\left(\beta, \hat{\mathcal{I}}^{-1}\right)$$

Deviance

- Deviance is the GLM counterpart of the residual sum of squares in normal linear regression

$$
\begin{aligned}
D &= 2\phi\left[l\left(\hat{\boldsymbol{\beta}}_{\max}\right) - l\left(\hat{\boldsymbol{\beta}}\right)\right] \\
&= \sum_{i=1}^{n} 2\omega_i\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]
\end{aligned} \tag{6}
$$

  - Here, $l\left(\hat{\boldsymbol{\beta}}_{\max}\right)$ is the maximized likelihood of the saturated model: the model with one parameter per data point. For exponential family distribution, it is computed by simply setting $\hat{\boldsymbol{\mu}} = \mathbf{y}$.
  - $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ are the maximum likelihood estimates of the canonical parameters for the saturated model and the model of interest, respectively

- From the second equality, we can see that deviance is independent of ϕ

- For normal linear regression, deviance equals the residual sum of squares

Scaled deviance

- Scaled deviance does depend on ϕ

$$D^* = \frac{D}{\phi}$$

- If the model is specified correctly, then approximately

$$D^* \sim \chi^2_{n-p}$$

- To compare two nested models,
  - If ϕ is known, then under H_0, we can use

$$D^*_0 - D^*_1 \sim \chi^2_{p_1 - p_0}$$

  - If ϕ is unknown, then under H_0, we can use

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \sim F_{p_1 - p_0, n - p_1}$$

Canonical link functions

- The canonical link g_c is the link function such that

$$g_c(\mu_i) = \theta_i = \eta_i$$

  where θ_i is the canonical parameter of the distribution

- Under canonical links, the observed information $\hat{\mathcal{I}}$ and the expected information $\mathcal{I}$ matrices are the same

- Under canonical links, since $\frac{\partial \theta_i}{\partial \beta_j} = X_{ij}$, the system of equations that the MLE satisfies becomes

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \omega_i (y_i - \mu_i) X_{ij} = 0$$

  Thus, if $\omega_i = 1$, we have

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \hat{\boldsymbol{\mu}}$$

  - For any GLM with an intercept term and canonical link, the residuals sum to zero, i.e., $\sum_i y_i = \sum_i \hat{\mu}_i$

GLM residuals

- Model checking is perhaps the most important part of applied statistical modeling
- It is usual to standardize GLM residuals so that if the model assumptions are correct,
  - the standardized residuals should have approximately equal variance, and
  - behave like residuals from an ordinary linear model
- Pearson residuals

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\mu_i)}}$$

  - In practice, the distribution of the Pearson residuals can be quite asymmetric around zero. So the deviance residuals (introduced next) are often preferred.

Deviance residuals

- Denote d_i as the ith component in the deviance definition (6), so that the deviance is $D = \sum_{i=1}^{n} d_i$
- By analogy with the ordinary linear model,we define the deviance residual

$$\hat{\epsilon}_i^d = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

  - The sum of squares of the deviance residuals gives the deviance itself

Quasi-likelihood

- Consider an observation y_i, of a random variable with mean μ_i and *known* variance function $V(\mu_i)$
  - Getting the distribution of Y_i exactly right is rather unimportant, as long as the **mean-variance relationship** $V(\cdot)$ is correct
- Then the log quasi-likelihood for μ_i, given y_i, is

$$q_i(\mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - z}{\phi V(z)} \, dz$$

  - The log quasi-likelihood for the mean vector $\boldsymbol{\mu}$ of all the response data is $q(\boldsymbol{\mu}) = \sum_{i=1}^{n} q_i(\mu_i)$
- To obtain the maximum quasi-likelihood estimation of $\boldsymbol{\beta}$, we can differentiate q wrt β_j, for $\forall j$

$$0 = \frac{\partial q}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \Longrightarrow \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} X_{ij} = 0$$

  this is exactly the GLM maximum likelihood solution, which can be obtained through IRLS

Generalized linear mixed models (GLMM)

- A GLMM model for an exponential family random variable Y_i

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}, \quad \mathbf{b} \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\psi}_{\boldsymbol{\theta}}\right)$$

- Difficulty in moving from linear mixed models to GLMM: it is no longer possible to evaluate the marginal likelihood analytically
- One effective solution is **Taylor expansion** around $\hat{\mathrm{b}}$, the posterior mode of $f(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\beta})$

$$f(\mathbf{y} \mid \boldsymbol{\beta}) \approx \int \exp\left\{\log f(\mathbf{y}, \hat{\mathbf{b}} \mid \boldsymbol{\beta}) + \frac{1}{2}\left(\mathbf{b} - \hat{\mathbf{b}}\right)^T \frac{\partial^2 \log f(\mathbf{y}, \mathbf{b} \mid \boldsymbol{\beta})}{\partial \mathbf{b} \partial \mathbf{b}^T} \left(\mathbf{b} - \hat{\mathbf{b}}\right)\right\} d\mathbf{b}$$

Laplace approximation of GLMM marginal likelihood

- For GLM, note that the expected Hessian is

$$-\frac{\mathbf{Z}^T\mathbf{W}\mathbf{Z}}{\phi} - \boldsymbol{\psi}^{-1}$$

  – $\mathbf{W}$ is the IRLS weight vector (3) based on the $\boldsymbol{\mu}$ implied by $\hat{\mathrm{b}}$ and β

- Therefore, the approximate marginal likelihood is

$$f(\mathbf{y} \mid \boldsymbol{\beta}) \approx f(\mathbf{y}, \hat{\mathbf{b}} \mid \boldsymbol{\beta}) \frac{(2\pi)^{p/2}}{\left|\frac{\mathbf{Z}^T\mathbf{W}\mathbf{Z}}{\phi} + \boldsymbol{\psi}_{\boldsymbol{\theta}}^{-1}\right|^{1/2}}$$

Penalized likelihood and penalized IRLS

- The point estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are obtained by optimizing the penalized likelihood

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}} &= \arg\max_{\boldsymbol{\beta},\mathbf{b}} \log f(\mathbf{y}, \mathbf{b} \mid \boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta},\mathbf{b}} \left\{ \log f(\mathbf{y} \mid \mathbf{b}, \boldsymbol{\beta}) - \mathbf{b}^T \psi_\theta^{-1} \mathbf{b}/2 \right\}
\end{aligned}
$$

- To simplify notation, we denote

$$
\begin{aligned}
\mathcal{B}^T &= (\mathbf{b}, \boldsymbol{\beta})^T \\
\mathcal{X} &= (\mathbf{Z}, \mathbf{X}), \quad \mathbf{S} = \begin{bmatrix} \psi_\theta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}
\end{aligned}
$$

- A penalized version of the IRLS algorithm (PIRLS) : by (5), a single Newton update step is

$$
\mathcal{B}^{(t+1)} = \left(\mathcal{X}^T \mathbf{W} \mathcal{X} + \phi \mathbf{S} \right)^{-1} \mathcal{X}^T \mathbf{W} \left[\mathbf{G} \left(\mathbf{y} - \hat{\boldsymbol{\mu}} \right) + \mathcal{X} \mathcal{B}^{(t)} \right]
$$

Penalized quasi-likelihood method

- Since optimizing the Laplace approximate marginal likelihood can be computationally costly, it is therefore tempting to instead perform a PIRLS iteration, estimating $\boldsymbol{\theta}, \phi$ at each step based on the working mixed model

$$\mathbf{z} \mid \mathbf{b}, \boldsymbol{\beta} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{W}^{-1}\phi\right), \quad \mathbf{b} \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\psi}_{\boldsymbol{\theta}}\right)$$

References

- Wood, Simon N. (2017), *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC