Generalized linear models

Exponential dispersion family

- Assume the response Y has pdf or pmf

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - h(y,\phi)\right\},$$

for some functions $a(\cdot), b(\cdot), h(\cdot,\cdot)$. The parameter θ is called "canonical parameter" and ϕ is a dispersion parameter.

- The above family is called exponential dispersion family.
- This family includes many distributions as special cases: normal, Poisson, Binomial, Gamma,$\cdots$

Link function

- The heart of a generalized linear model is to model some function of μ as a linear function of predictors X. That is for some link function $g(\cdot)$ and some p-dim coefficient vector β, we assume that

$$g(\mu) = X^T \beta.$$

- A canonical link is to set $\theta = b'^{-1}(\mu) = X^T \beta$.
- Therefore, a canonical link function is $g(\cdot) = b'^{-1}(\cdot)$.

Maximum likelihood

- Assume that $Y_1, \cdots, Y_n$ are independent sample from $f(y; \theta, \phi)$ and $X_1, \cdots, X_n$ are the corresponding covariates.

  The log-likelihood for β and ϕ is

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} - h(Y_i, \phi) \right\}$$

  where $\theta_i = b'^{-1}(\mu_i) = b'^{-1}\{g^{-1}(X_i^T \beta)\}$.

- The maximum likelihood estimator of β is

  $\hat{\beta} = \arg\max_\beta \sum_{i=1}^{n} \{Y_i \theta_i - b(\theta_i)\}$.

Score function in a matrix form

Let $X = (X_1, \cdots, X_n)^T$ be an $n \times p$ matrix, $Y = (Y_1, \cdots, Y_n)^T$ and $\mu = (\mu_1, \cdots, \mu_n)^T$.

Define the following $n \times n$ matrices,

$$W^{-1} = \text{diag}\{V(\mu_1)\{g'(\mu_1)\}^2, \cdots, V(\mu_n)\{g'(\mu_n)\}^2\}$$

and

$$\Delta = \text{diag}\{g'(\mu_1), \cdots, g'(\mu_n)\}.$$

The score function can be written as

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \{a(\phi)\}^{-1} X^T W \Delta (Y - \mu).$$

Fisher's information matrix

- The Fisher's information matrix for β is

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta \partial \beta^T} = -\{a(\phi)\}^{-1} X^T W \Delta \frac{\partial \mu}{\partial \beta^T} + \{a(\phi)\}^{-1} X^T \frac{\partial W \Delta}{\partial \beta^T} (Y - \mu).$$

- The Fisher's information matrix is

$$I_n = \{a(\phi)\}^{-1} X^T W X.$$

Computation issue

- The maximum likelihood estimator of β is a solution of the score function $\partial \ell(\beta, \phi)/\partial \beta = 0$.
- The estimating function typically does not have a closed form solution. We need to use an numerical algorithm to find the solutions of the estimating equations.
- In the score function, the quantities W, Δ and μ all contain the unknown parameter β.

Fisher's scoring method

Let $S(\beta)$ denote the score function $\partial \ell(\beta, \phi)/\partial \beta$. A Newton algorithm updates the parameter by considering the following Taylor expansion at some initial values β_0,

$$S(\beta) \approx S(\beta_0) + \frac{\partial \ell(\beta, \phi)}{\partial \beta \partial \beta^T}(\beta - \beta_0).$$

We then update the parameter β_0 by

$$\beta = \beta_0 - \left[\frac{\partial \ell(\beta, \phi)}{\partial \beta \partial \beta^T}\right]^{-1} S(\beta_0).$$

Fisher's scoring method replaces $-\partial \ell(\beta, \phi)/\{\partial \beta \partial \beta^T\}$ by its corresponding expectation, which is the Fisher's information matrix.

Fisher's scoring method

Step 1: Choose some initial value $\hat{\beta}^{(0)}$.

Step 2: At the k-th step, we update the parameter $\hat{\beta}^{(k)}$ by $\hat{\beta}^{(k+1)}$ through solving the following equations

$$I_n(\hat{\beta}^{(k)})\hat{\beta}^{(k+1)} = I_n(\hat{\beta}^{(k)})\hat{\beta}^{(k)} + S(\hat{\beta}^{(k)}).$$

Step 3: Repeat Step 1 and 2 until convergence.

Fisher's scoring method for generalized linear models

At the k-th step, we update the parameter $\hat{\beta}^{(k)}$ by $\hat{\beta}^{(k+1)}$ through solving the following equations

$$\begin{aligned} X^T W X \hat{\beta}^{(k+1)} &= X^T W X \hat{\beta}^{(k)} + X^T W \Delta(Y - \mu) \\ &= X^T W \{X \hat{\beta}^{(k)} + \Delta(Y - \mu)\}. \end{aligned}$$

The solution $\hat{\beta}^{(k+1)}$ is equivalent to a weighted least squares estimator with weights matrix W and dependent response vector $X \hat{\beta}^{(k)} + \Delta(Y - \mu)$.

Iteratively reweighted least squares (IRWLS) algorithm

Step 1: set initial estimates $X_i^T \hat{\beta}^{(0)}$ and $\hat{\mu}^{(0)}$.

Step 2: At k-th step, compute "adjusted dependent response" as

$$Z_i^{(k)} = X_i^T \hat{\beta}^{(k)} + (Y_i - \hat{\mu}^{(k)}) g'(\hat{\mu}^{(k)}).$$

Step 3: Compute weights $\{W_i^{(k)}\}^{-1} = V(\hat{\mu}^{(k)})\{g'(\hat{\mu}^{(k)})\}^2$.

Step 4: Update $\hat{\beta}^{(k)}$ by $\hat{\beta}^{(k+1)}$ through a weighted least squares regression with response $Z_i^{(k)}$ and weights $W_i^{(k)}$.

Step 5: Repeat Steps 2-4 until convergence.

Example: Logistic regression model

Recall that in a logistic regression model with Bernoulli distributed response,

$$\mu = p,\ g(\mu) = \log\{\mu/(1-\mu)\} \text{ and } V(\mu) = \mu(1-\mu).$$

Therefore,

$$g'(\mu) = \frac{1}{\mu} + \frac{1}{1-\mu} = \frac{1}{\mu(1-\mu)}.$$

and

$$W_i = [\{g'(\mu)\}^2 V(\mu)]^{-1} = \mu(1-\mu).$$

Example: Logistic regression model

Step 1: set initial estimates $X_i^T \hat{\beta}^{(0)}$ and

$\hat{\mu}^{(0)} = \exp(X_i^T \hat{\beta}^{(0)})/\{1 + \exp(X_i^T \hat{\beta}^{(0)})\}$.

Step 2: At k-th step, compute "adjusted dependent response"

as

$$Z_i^{(k)} = X_i^T \hat{\beta}^{(k)} + (Y_i - \hat{\mu}^{(k)})/\{\mu^{(k)}(1 - \mu^{(k)})\}.$$

Step 3: Compute weights $W_i^{(k)} = \mu^{(k)}(1 - \mu^{(k)})$.

Step 4: Update $\hat{\beta}^{(k)}$ by $\hat{\beta}^{(k+1)}$ through a weighted least squares

regression with response $Z_i^{(k)}$ and weights $W_i^{(k)}$.

Step 5: Repeat Steps 2-4 until convergence.

Example: Poisson regression model

Recall that in a logistic regression model with Bernoulli distributed response,

$$\mu = exp(X^T\beta),\ g(\mu) = \log(\mu) \text{ and } V(\mu) = \mu.$$

Therefore,

$$g'(\mu) = \frac{1}{\mu}.$$

and

$$W_i = [\{g'(\mu)\}^2 V(\mu)]^{-1} = \mu.$$

Example: Poisson regression model

Step 1: set initial estimates $X_i^T \hat{\beta}^{(0)}$ and $\hat{\mu}^{(0)} = \exp(X_i^T \hat{\beta}^{(0)})$.

Step 2: At k-th step, compute "adjusted dependent response" as

$$Z_i^{(k)} = X_i^T \hat{\beta}^{(k)} + (Y_i - \hat{\mu}^{(k)})/\mu^{(k)}.$$

Step 3: Compute weights $W_i^{(k)} = \mu^{(k)}$.

Step 4: Update $\hat{\beta}^{(k)}$ by $\hat{\beta}^{(k+1)}$ through a weighted least squares regression with response $Z_i^{(k)}$ and weights $W_i^{(k)}$.

Step 5: Repeat Steps 2-4 until convergence.