

Generalised Linear Models (GLMs) I

Learning goals

- Be able to define Generalised Linear Model (GLM).
- Be able to obtain the canonical link function for GLM.
- (Challenging) Be able to explain ideas for the iterated weighted least squares (IWLS) algorithm.
- Be able to compute the variance of parameter estimates in GLM using IWLS algorithm.
- Understand (scaled) deviance
 - Be able to define (scaled) deviance.
 - Be able to compute (scaled) deviance.
 - Be able to use it to test model adequacy, perform model selection for nested models and non-nested models.
- Be able to perform diagnostics for GLM using R.
 - Understand what different 'diagnostics measurements' aim to measure.
 - Be able to obtain those diagnostic measurements from the *glm* output.
 - Be able to perform diagnostics for GLM using R.

Generalised Linear Model

Definition: Y is a GLM if it is from an exponential family, and

$$\mu := \mathbb{E}Y = g^{-1}(\mathbf{x}^T \boldsymbol{\beta})$$

where

g is a monotonic differentiable function called the *link function*.

\mathbf{x} is a vector of independent (predictor) variables, and

$\boldsymbol{\beta}$ is a vector of parameters

Remark: We model *location* using $\eta = \mathbf{x}^T \boldsymbol{\beta}$, and let the *scale* sort itself out. That is, we do not model the scale explicitly.

Canonical link

Recall Y is from an exponential family if

$$f(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right].$$

If $g(\mu) = g(\mathbb{E} Y) = \mathbf{x}^T \boldsymbol{\beta} = \theta$ then g is called the *canonical* link. Since $\mu = b'(\theta)$, it follows that the canonical link must be $(b')^{-1}$.

Examples: canonical links

normal $\theta = \mu, g(\mu) = \mu$

Poisson $\theta = \log \lambda = \log \mu, g(\mu) = \log \mu$

binomial $\theta = \log \frac{p}{1-p} = \log \frac{\mu}{m-\mu}, g(\mu) = \log \frac{\mu}{m-\mu}$

Estimation of parameters in GLM

- Iterated Weighted Least Squares (IWLS) algorithm
- Implemented in the *glm* function in R

Estimation of parameters in GLM using maximum likelihood methods

Suppose we have independent observations y_i from an exponential family, with canonical parameter θ_i and dispersion parameter ϕ , for $i = 1, \dots, n$.

Furthermore suppose that y_i has mean

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

If g is the canonical link then $\theta_i = x_i^T \beta$.

The log-likelihood is then

$$l(\beta, \phi; y) = \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

Maximum likelihood methods aim to find the values of parameters that maximise their log likelihood function.

Aside: Newton-Raphson method

If $f(x)$ has a continuous derivative $f'(x)$, then the problem of finding the value that maximise $f(x)$ is equivalent to 1) finding x_1^*, \dots, x_k^* that are the roots of $f'(x)$, and then 2) among them, finding the one that maximise $f(x)$.

We apply the Newton-Raphson method for root-finding to $f'(x)$:

$$x^{n+1} = x^n - \frac{f'(x^n)}{f''(x^n)}.$$

Newton-Raphson method to find MLE

Suppose we wish to find the value of θ that maximise a log likelihood $l(\theta)$ using Newton-Raphson method.

Our update step is

$$\theta^{n+1} = \theta^n - H(\theta^n)^{-1} U(\theta^n)$$

where

$$U(\theta) = \frac{\partial l(\theta)}{\partial \theta} \text{ and } H(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} = -\mathcal{J}(\theta).$$

If we replace \mathcal{J} by \mathcal{I} , the Fisher information, then the algorithm is called [Fisher scoring](#).

The Fisher information is guaranteed to be positive definite (unlike the observed information).

Connection with Iterated Weighted Least Squares (IWLS) algorithm?

It turns out that the Fisher scoring applied to GLM can be interpreted as a least squares problem (if you want to know details, read “McCullagh & Nelder on IWLS” posted on LMS).

The *glm* function in R computes MLEs of parameters in GLM using iterated weighted “Least Squares” algorithm.

Aside: Weighted Least Squares method

Suppose $Y = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \Sigma)$, and X and Σ are full rank.

Multiplying by $\Sigma^{-1/2}$ we get $\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \varepsilon'$ where $\varepsilon' \sim N(0, I)$.

What is the least squares estimator of β ?

The estimator of β that minimises the sum of squares

$$(\Sigma^{-1/2}y - \Sigma^{-1/2}X\beta)^T(\Sigma^{-1/2}y - \Sigma^{-1/2}X\beta) = (y - X\beta)^T\Sigma^{-1}(y - X\beta)$$

is

$$\begin{aligned}\hat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \\ &= (X^TWX)^{-1}X^TWy,\end{aligned}$$

where the weight $W = \Sigma^{-1}$.

Weighted Least Squares for GLM

$$g(Y_i) \approx g(\mu_i) + (Y_i - \mu_i)g'(\mu_i)$$

Let $Z_i = g(\mu_i) + (Y_i - \mu_i)g'(\mu_i)$ and $\epsilon_i = (Y_i - \mu_i)g'(\mu_i)$. Then,

$$Z_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where

$$\text{Var } \epsilon_i = (g'(\mu_i))^2 \text{Var } Y_i.$$

If we knew $\text{Var } \epsilon_i$ then the estimator of $\boldsymbol{\beta}$ would be the solution to the weighted least squares problem:

$$\min_{\boldsymbol{\beta}} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$$

where $\boldsymbol{\Sigma}$ is diagonal with $\Sigma_{ii} = \text{Var } \epsilon_i$.

Weighted Least Squares for GLM

We have $g(Y) \approx Z = X\beta + \epsilon$ where $\epsilon_i = (Y_i - \mu_i)g'(\mu_i)$ and $\Sigma = \text{Var } \epsilon$ is diagonal with entries

$$\Sigma_{ii} = (g'(\mu_i))^2 \text{Var } Y_i = (g'(\mu_i))^2 v(\mu_i) a(\phi).$$

This suggests

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} z.$$

Problem: z_i and Σ_{ii} depend on β because $\mu_i = g^{-1}(x_i^T \beta)$.

Solution: iterate!

Note that the $a(\phi)$ factor in the expression for $\hat{\beta}$ cancels out.

Iterated Weighted Least Squares (IWLS) algorithm

IWLS algorithm (for finding MLE of β)

- 1 Start with $\hat{\mu}^0 = y$.
- 2 Given $\hat{\mu}^n$ calculate, for each i ,
 $z_i^n = g(\hat{\mu}_i^n) + (y_i - \hat{\mu}_i^n)g'(\hat{\mu}_i^n)$ and
 $W_{ii}^n = \frac{1}{\Sigma_{ii}} = \frac{1}{g'(\hat{\mu}_i^n)^2 v(\hat{\mu}_i^n)}$.
- 3 Put $\hat{\beta}^{n+1} = (X^T W^n X)^{-1} X^T W^n z^n$ and
for each i , $\hat{\mu}_i^{n+1} = g^{-1}(x_i^T \hat{\beta}^{n+1})$.
- 4 If $\hat{\beta}^{n+1}$ is sufficiently close to $\hat{\beta}^n$ then stop, otherwise
return to (2).

For GLM, IWLS algorithm is equivalent to Fisher scoring.

Example: see the section “IWLS” in Bliss.pdf.

Variance of $\hat{\beta}$ from IWLS algorithm

Suppose that the IWLS algorithm converges to the estimate $\hat{\beta}$, then

$$\hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} z$$

where, elementwise, we have

$$\begin{aligned}\hat{\mu}_i &= g^{-1}(x_i^T \hat{\beta}) \\ z_i &= g(\hat{\mu}_i) + (y_i - \hat{\mu}_i)g'(\hat{\mu}_i) \\ \hat{\Sigma}_{ii} &= (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i).\end{aligned}$$

Since $\text{Var } z = \hat{\Sigma}$ we have

$$\begin{aligned}\text{Var } \hat{\beta} &= [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}] \hat{\Sigma} [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}]^T \\ &= (X^T \hat{\Sigma}^{-1} X)^{-1}\end{aligned}$$

Note that the $a(\phi)$ term in $\hat{\Sigma}$ does not cancel here, as it did in the IWLS algorithm, so we need to estimate it.

Estimator for $a(\phi)$

Y_i has mean μ_i and variance $v(\mu_i)a(\phi)$. So

$$\sum_i \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)a(\phi)} \approx \chi_{n-p}^2$$

where p is the number of parameters used to estimate μ .

Let

$$X^2 := \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

Then, $X^2/(n-p)$ will be an estimator for $a(\phi)$.

X^2 is called Pearson's χ^2 statistic, and it can be shown that $X^2/(n-p)$ is a consistent estimator for $a(\phi)$.

Variance of $\hat{\beta}$ from IWLS algorithm

$$\text{Var } \hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1},$$

where

$$\begin{aligned}\hat{\Sigma}_{ii} &= (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i) a(\hat{\phi}) \\ &= (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i) X^2 / (n - p).\end{aligned}$$