# Generalised Linear Models (GLMs) I

# Learning goals

- Be able to define Generalised Linear Model (GLM).
- Be able to obtain the cannonical 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

①

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

②
$$\mu := \mathbb{E}Y = g^{-1}(x^T\beta)$$

$= b'(\theta)$ $\qquad x^T\beta = g(\mu)$. relationship between predictors and $\mu$.

where

- $g$ is a monotonic differentiable function called the *link function*.
- $x$ is a vector of independent (predictor) variables, and
- $\beta$ is a vector of parameters

$\Rightarrow x^T\beta = g(\hat{\mu})$

$\hat{\mu}_i = y_i$

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

✲ Interpretation

① log vs logit.

② estimation of parameter

$\mu \Leftrightarrow p$.

for binomial $\mu = np$

# Canonical link

Recall $Y$ is from an exponential family if

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

Def:

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

$b'(\theta) = \mu$.

$\theta = (b')^{-1}(\mu)$.

$g(\mu) = x^T\beta = \theta$.

$\mu = b'(\theta)$.

$\theta = (b')^{-1}(\mu)$.

$\underbrace{g(\mu) = x^T\beta}_{\text{def.}} = \theta$

requirement for cannonical link.

## Examples: canonical links

① normal

$b(\theta)=\frac{\theta^2}{2}$ $\quad b'(\theta)=\theta$

[1] $b'(\theta)$ identity function

inverse of identity is still identity

$g(\mu)=\mu$.

[2] $g(\mu)=\theta$ (by def)

$\mu=\theta \quad g(\theta)=\theta$

$\Rightarrow g(\mu)=\mu$

② Poisson

$b(\theta)=e^\theta$

$b'(\theta)=e^\theta=\mu$.

$\theta=\log\mu$.

$g(\mu)=\log\mu$.

$b'(\theta)=e^\theta=\mu$.

$\theta=\log\mu$.

$g(\mu)=\theta=\log\mu$

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

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

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

③ binomial

from lab

$\theta=\log\frac{p}{1-p}$

$b(\theta)=m\log(1+e^\theta)$

$\mu=b'(\theta)=\frac{me^\theta}{1+e^\theta}$

$\Rightarrow \frac{\mu}{m}=\frac{e^\theta+1-1}{1+e^\theta}=1-\frac{1}{1+e^\theta}$

$\Rightarrow e^\theta=\frac{1}{1-\frac{\mu}{m}}-1=\frac{m-(m-\mu)}{m-\mu}=\frac{\mu}{m-\mu} \Rightarrow \boxed{\theta=\log\frac{\mu}{m-\mu}}$

logit link

or $\theta=\log\frac{\frac{\mu}{m}}{1-\frac{\mu}{m}}$

$\therefore g(\mu)=\theta=\log\frac{\mu}{m-\mu}$

# 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 $\theta_i$ and dispersion parameter $\phi$, for $i = 1, \ldots, n$.

Furthermore suppose that $y_i$ has mean

sample specific

$$\mu_i = b'(\theta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

If $g$ is the canonical link then $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

The log-likelihood is then

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

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

# Aside: Newton-Raphson method

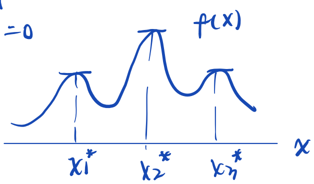
If $f(x)$ has a continuous derivative $f'(x)$, then the problem of finding the value that maximise $f(x)$ is equivalent to 1) finding $x_1^*, \ldots, x_k^*$ that are the roots of $f'(x)$, and then 2) among them, finding the one that maximise $f(x)$.

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

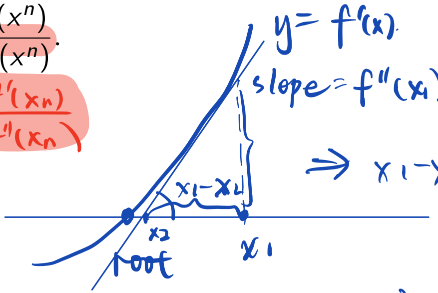
$$x^{n+1} = x^n - \frac{f'(x^n)}{f''(x^n)}.$$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

(1) find $f'(x)=0$




(2) evaluate $f(x)$.
find max




$$\text{slope} = f''(x_1) = \frac{f'(x_1)}{x_2 - x_1}$$

$$\Rightarrow x_1 - x_2 = \frac{f'(x_1)}{f''(x_1)}.$$

$$x_2 = x_1 - \frac{f'(x_1)}{f''(x_1)}$$

# Newton-Raphson method to find MLE

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

Our update step is

$\theta_{n+1} = \theta_n - \frac{f'(\theta_n)}{f''(\theta_n)}$

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

second derivative.

where

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

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

may assume no independence

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

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

What is the least squares estimator of $\beta$?

$Y_* = X_*\beta + \varepsilon'$

The estimator of $\beta$ that minimises the sum of squares

$\hat{\beta} = (X_*^T X_*)^{-1} X_*^T Y_*$

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

is $(Y_* - X_*\beta)^T(Y_* - X_*\beta)$

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

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

$( ? \ (\Sigma^{-1/2})^T = \Sigma^{-1/2})$

where the weight $W = \Sigma^{-1}$ → different weight to different sample.

# Weighted Least Squares for GLM

Z1

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

Taylor expansion

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

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

where

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

$\hat{\beta}$

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

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

where $\Sigma$ is diagonal with $\Sigma_{ii} = \mathrm{Var}\,\epsilon_i$.

# Weighted Least Squares for GLM

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

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

This suggests

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

$\hat{\beta}_0 \to \hat{\mu}_0 \to \hat{z}, \hat{\Sigma}$

$\hat{\beta}_1$

**Problem:** $z_i$ and $\Sigma_{ii}$ depend on $\boldsymbol{\beta}$ because $\mu_i = g^{-1}(\mathrm{x}_i^T \boldsymbol{\beta})$.

**Solution:** iterate!

$\Sigma$ depend on $\mu_i$. $\mu_i$ depend on $\beta$

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

# Iterated Weighted Least Squares (IWLS) algorithm

Notice:

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

$= (X^T W X)^{-1} X^T W z$

$z_i = g(\mu_i) + (y_i - \mu_i) g'(\mu_i)$

$W_{ii} = \frac{1}{\Sigma_{ii}} = \frac{1}{(g'(\mu_i))^2 v(\mu_i)}$

$\mu_i = g^{-1}(x_i^T \beta)$

**IWLS algorithm (for finding MLE of $\beta$)**

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

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

$\text{var}(\hat{\beta}) = A\,\text{var}(z)\,A^T$

where, elementwise, we have

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

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

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

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 $\mu_i$ and variance $v(\mu_i)a(\phi)$. So

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

where $p$ is the number of parameters used to estimate $\boldsymbol{\mu}$.

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

linear model

$\hat{\sigma}^2 = \frac{\sum(y_i-\hat{\mu}_i)^2}{n-p}$

$\sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)^2}{\sigma^2} \sim \chi^2_{n-p}$

$E\left(\sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)^2}{\sigma^2}\right) = n-p$

the expectation of $\chi^2_{n-p}$ is $n-p$

$E\left(\frac{\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2}{n-p}\right) = \sigma^2$

GLM $E(y_i) = \mu_i$

$var(y_i) = V(\mu) \cdot a(\phi)$.

$\sum \frac{(y_i-\mu_i)^2}{V(\mu_i)a(\phi)} \sim \chi^2_{n-p}$

$E\left(\sum \frac{(y_i-\mu_i)^2}{V(\mu_i)\, n-p}\right) = a(\phi)$.

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

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

where

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