

Any model that satisfies

$$f(y_i|\theta_i, \phi) = \exp [(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)]$$

with

$$\theta_i = \theta(\eta_i) \quad \text{and} \quad \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

is called a **generalized linear model (GLM)**.

Distributions and their canonical links:

| <u>Distribution</u> | <u>Canonical link</u> |
|---------------------|------------------------------|
| Normal | $\eta = \mu$ |
| Poisson | $\eta = \log \mu$ |
| Binomial | $\eta = \log(\mu/(1 - \mu))$ |
| Gamma | $\eta = 1/\mu$ |

Before discussing Bayesian inference in the GLM, we'll discuss maximum likelihood estimation and frequentist inference.

In general, MLEs do not have a “closed” form for GLMs. Need to rely on **Newton-Raphson** or **Fisher scoring** to approximate MLEs.

We'll derive the likelihood equations for β . We have

$$\begin{aligned} \ell(\beta, \phi) &= \sum_{i=1}^n \left\{ \frac{[y_i \theta_i - b(\theta_i)]}{\phi} + c(y_i, \phi) \right\} \\ &= \sum_{i=1}^n \ell_i(\theta_i, \phi) = \sum_{i=1}^n \ell_i. \end{aligned}$$

In the previous equation we have

$$\theta_i = \theta(\mathbf{x}_i^T \boldsymbol{\beta}) = \theta(\eta_i).$$

For any GLM, the likelihood function depends on $\boldsymbol{\beta}$ **only** through η_i . Now,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial}{\partial \beta_j} \ell_i \\ &= \sum_{i=1}^n \frac{\partial \ell_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{\partial \ell_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \eta_i} \cdot x_{ij}. \end{aligned}$$

Define $\Delta_i = \partial \theta_i / \partial \eta_i$, which is sometimes called the **link adjustment**. If the canonical link is used, then $\Delta_i \equiv 1$, since in this case $\theta_i = \eta_i$.

We now have

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}.$$

The likelihood equations may thus be written

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\beta}, \phi) &= 0 \quad \iff \\ \sum_{i=1}^n (y_i - \mu_i) \Delta_i x_{ij} &= 0, \quad j = 1, \dots, p. \end{aligned}$$

These equations can be written in matrix form as

$$\mathbf{X}^T \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0},$$

where \mathbf{X} is $n \times p$ with element x_{ij} in the i th row and j th column, $\boldsymbol{\Delta}$ is an $n \times n$ diagonal matrix with diagonal elements $\Delta_1, \dots, \Delta_n$, $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$.

Letting $\mathbf{S} = (\mathbf{y} - \boldsymbol{\mu})$, the equations are

$$\mathbf{X}^T \boldsymbol{\Delta} \mathbf{S} = \mathbf{0}.$$

When using a canonical link, $\boldsymbol{\Delta} = \mathbf{I}$.

In general, the equations are nonlinear in $\boldsymbol{\beta}$. We can use Newton-Raphson for obtaining the MLE of $\boldsymbol{\beta}$, call it $\hat{\boldsymbol{\beta}}$.

Note that the likelihood equations do not depend on ϕ , and hence estimation of $\boldsymbol{\beta}$ is **not dependent on ϕ** .

The Newton-Raphson procedure may be expressed as follows:

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} - \left[\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{X}^T \boldsymbol{\Delta} \mathbf{S}) \right]^{-1} \\ &\quad \times (\mathbf{X}^T \boldsymbol{\Delta} \mathbf{S}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}. \end{aligned}$$

The book *Matrix Algebra from a Statistician's Perspective*, by David A. Harville, is a great resource. Among many other things it has various formulas for differentiating matrices.

According to p. 298 of Harville,

$$\frac{\partial}{\partial \beta_j} \mathbf{X}^T \Delta \mathbf{S} = \mathbf{X}^T \left[\Delta \frac{\partial \mathbf{S}}{\partial \beta_j} + \frac{\partial \Delta}{\partial \beta_j} \mathbf{S} \right].$$

Using the definition of \mathbf{S} ,

$$\begin{aligned} \frac{\partial \mathbf{S}}{\partial \beta_j} &= - \left(b''(\theta_1) \theta'(\eta_1) x_{1j}, \dots, b''(\theta_n) \theta'(\eta_n) x_{nj} \right)^T \\ &= -\mathbf{V} \Delta (x_{1j}, \dots, x_{nj})^T, \end{aligned}$$

where \mathbf{V} is the $n \times n$ diagonal matrix with diagonal elements $b''(\theta_i)$, $i = 1, \dots, n$.

It follows that

$$\mathbf{X}^T \Delta \frac{\partial \mathbf{S}}{\partial \beta} = -\mathbf{X}^T \Delta \mathbf{V} \Delta \mathbf{X}.$$

Also,

$$\frac{\partial \Delta}{\partial \beta_j} = \text{diagonal matrix with diagonal entries}$$

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \theta'(\eta_i) &= \theta''(\eta_i) \frac{\partial \eta_i}{\partial \beta_j} \\ &= \theta''(\eta_i) x_{ij}, \quad i = 1, \dots, n. \end{aligned}$$

It follows that

$$\frac{\partial \Delta}{\partial \beta_j} \mathbf{S} = \dot{\Delta} \mathbf{H}(x_{1j}, \dots, x_{nj})^T,$$

where $\dot{\Delta}$ and \mathbf{H} are diagonal with diagonal entries $\theta''(\eta_i)$ and $y_i - \mu_i$, respectively, $i = 1, \dots, n$.

Finally then

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{X}^T \boldsymbol{\Delta} \mathbf{S} = -\mathbf{X}^T (\boldsymbol{\Delta} \mathbf{V} \boldsymbol{\Delta} - \dot{\boldsymbol{\Delta}} \mathbf{H}) \mathbf{X}.$$

\mathbf{V} is called the matrix of variance functions.

Recalling the form of $\partial \ell(\boldsymbol{\beta}, \phi) / \partial \boldsymbol{\beta}$, we see that

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \ell(\boldsymbol{\beta}, \phi) = -\frac{1}{\phi} \mathbf{X}^T (\boldsymbol{\Delta} \mathbf{V} \boldsymbol{\Delta} - \dot{\boldsymbol{\Delta}} \mathbf{H}) \mathbf{X}.$$

So, the Newton-Raphson iterations may be written

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left[\mathbf{X}^T (\boldsymbol{\Delta} \mathbf{V} \boldsymbol{\Delta} - \dot{\boldsymbol{\Delta}} \mathbf{H}) \mathbf{X} \right]^{-1} \\ &\quad \times \left(\mathbf{X}^T \boldsymbol{\Delta} \mathbf{S} \right) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}. \end{aligned}$$

When using a canonical link, $\boldsymbol{\Delta} = \mathbf{I}$ and $\dot{\boldsymbol{\Delta}} = \mathbf{0}$, in which case

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}^T \mathbf{V} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{S} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}.$$

In the canonical link case, the Newton-Raphson scheme is an **iteratively reweighted least squares (IRLS)** algorithm. Define

$$\mathbf{Z}^{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{V}^{-1}\mathbf{S} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}.$$

Since

$$\boldsymbol{\beta}^{(t)} = (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{V} \mathbf{X}) \boldsymbol{\beta}^{(t)}$$

and

$$(\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{S} = (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{V} \mathbf{V}^{-1}) \mathbf{S},$$

we have

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{V}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} \mathbf{Z}^{(t)} \\ &= (\mathbf{X}^T \mathbf{V}^{(t)} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{V}^{(t)}) \mathbf{Z}^{(t)}, \end{aligned}$$

where

$$\mathbf{V}^{(t)} = \mathbf{V} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}.$$

Therefore, $\boldsymbol{\beta}^{(t+1)}$ corresponds to a weighted least squares regression of $\mathbf{Z}^{(t)}$ on \mathbf{X} with weight matrix $\mathbf{V}^{(t)}$.

What should be used as an initial estimate for β ? Let g be the μ -link. Then

$$\begin{aligned}g(Y_i) &\approx g(\mu_i) + (Y_i - \mu_i)g'(\mu_i) \\ &= \mathbf{x}_i^T \beta + (Y_i - \mu_i)g'(\mu_i).\end{aligned}$$

So, we have an approximate linear model

$$g(Y_i) = \mathbf{x}_i^T \beta + Z_i, \quad i = 1, \dots, n,$$

where $E(Z_i) = 0$, $i = 1, \dots, n$. This suggests that we take

$$\beta^{(0)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T g(\mathbf{Y}),$$

where $g(\mathbf{Y}) = (g(Y_1), \dots, g(Y_n))^T$.

Fisher Scoring

In Fisher scoring, the **expected Hessian** is used instead of the Hessian as in Newton-Raphson.

The Hessian is

$$-X^T(\Delta V \Delta - \dot{\Delta} H)X,$$

which has expectation $-X^T \Delta V \Delta X$ since

$$E(H) = 0.$$

The **Fisher Information matrix** for β is

$$\begin{aligned} I(\beta) &= -(-X^T \Delta V \Delta X) \\ &= X^T \Delta V \Delta X. \end{aligned}$$

The Fisher scoring algorithm is

$$\beta^{(t+1)} = \beta^{(t)} + (X^T \Delta V \Delta X)^{-1} X^T \Delta S \Big|_{\beta = \beta^{(t)}}.$$

When using a canonical link, the Fisher scoring algorithm is identical to Newton-Raphson.

Example: Logistic Regression

In this case ϕ is known and equal to 1. The link is canonical and the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) \right].$$

For Newton-Raphson, we need \mathbf{V} and \mathbf{S} .

$$\mathbf{S} = \left(y_1 - \frac{e^{\mathbf{x}_1^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_1^T \boldsymbol{\beta}}}, \dots, y_n - \frac{e^{\mathbf{x}_n^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_n^T \boldsymbol{\beta}}} \right)^T$$

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

$$\begin{aligned} b''(\theta) &= -\frac{e^{2\theta}}{(1 + e^\theta)^2} + \frac{e^\theta}{1 + e^\theta} \\ &= \frac{e^\theta}{(1 + e^\theta)^2} \end{aligned}$$

V is the $n \times n$ diagonal matrix with diagonal elements $b''(\theta_i)$, $i = 1, \dots, n$, where $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, \dots, n$.

The inverse of b' is the μ -link.

$$\log \left(\frac{\mu}{1 - \mu} \right) = \theta$$

To obtain an initial estimate of $\boldsymbol{\beta}$, we can't use the method mentioned before, at least without modification.

$$\log \left(\frac{y_i}{1 - y_i} \right) = \pm \infty$$

To take care of this problem, define

$$\tilde{y}_i = \begin{cases} 0.99, & \text{if } y_i = 1 \\ 0.01, & \text{if } y_i = 0. \end{cases}$$

As an initial estimate of β , use

$$\beta^{(0)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T g(\tilde{\mathbf{y}}),$$

where

$$g(\tilde{\mathbf{y}}) = \left(\log \left(\frac{\tilde{y}_1}{1 - \tilde{y}_1} \right), \dots, \log \left(\frac{\tilde{y}_n}{1 - \tilde{y}_n} \right) \right)^T.$$

We'll illustrate Newton-Raphson with a simulated example where there is only one covariate, but \mathbf{X} is $n \times 3$ since θ is modeled as a quadratic function of the covariate.

In the simulated example, we have observations Y_1, \dots, Y_{25} that are independent and such that

$$Y_i \sim \text{Bin}(1, \mu(d_i)), \quad i = 1, \dots, 25,$$

where $d_i = (i - 1/2)/25$, $i = 1, \dots, 25$,

$$\log \left(\frac{\mu(d)}{1 - \mu(d)} \right) = \frac{\beta_0}{5} + \beta_1 p_1(d) + \beta_2 p_2(d),$$
$$0 \leq d \leq 1,$$

and p_1 and p_2 are known linear and quadratic functions that are orthogonal on $[0, 1]$.

Observations were generated from this model with $\beta_0 = 6.66$, $\beta_1 = 17.307$ and $\beta_2 = 3.712$.

The \mathbf{X} matrix has columns $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(3)}$, where

$$\mathbf{x}^{(1)} = (1/5, \dots, 1/5)^T,$$

$$\mathbf{x}^{(2)} = (p_1(d_1), \dots, p_1(d_{25}))^T$$

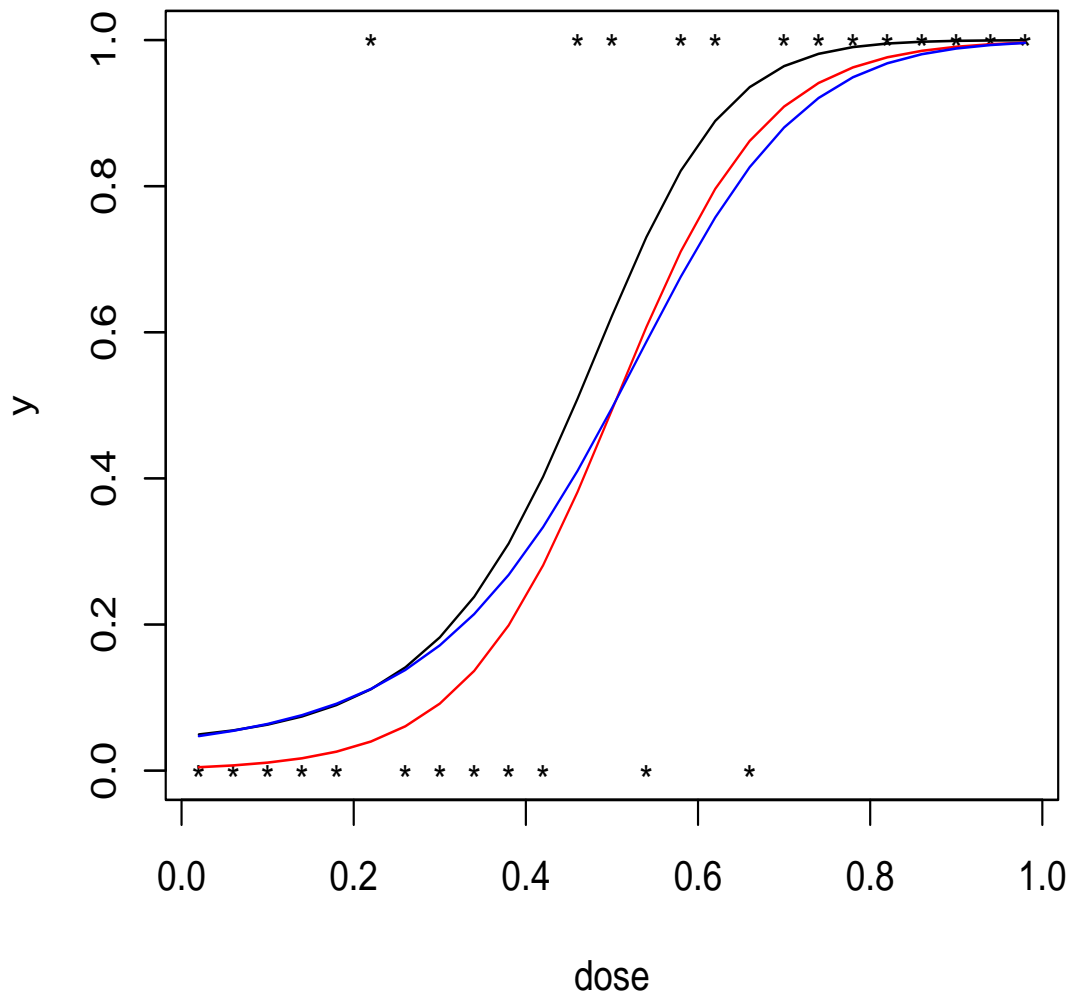
and

$$\mathbf{x}^{(3)} = (p_2(d_1), \dots, p_2(d_{25}))^T.$$

A plot of the generated data along with the true mean curve and two estimates is shown on the next page.

The initial estimate of β was gotten using the method described on pp. 210-211N.

Results for simulated logistic regression data



True mean curve: — MLE: —
Initial estimate: —

Newton-Raphson iterations

| Iteration | $\beta_0^{(t)}$ | $\beta_1^{(t)}$ | $\beta_2^{(t)}$ |
|-----------|-----------------|-----------------|-----------------|
| 0 | 0.184 | 16.568 | 0.277 |
| 1 | 5.581 | 10.404 | 5.490 |
| 2 | 3.596 | 12.162 | 3.588 |
| 3 | 2.594 | 12.690 | 2.473 |
| 4 | 2.293 | 12.861 | 2.118 |
| 5 | 2.269 | 12.877 | 2.090 |
| 6 | 2.269 | 12.877 | 2.090 |
| 7 | 2.269 | 12.877 | 2.090 |
| Truth | 6.660 | 17.307 | 3.712 |

Frequentist Inference

Sampling distributions of statistics in GLM are generally not known in closed form. We can use asymptotic theory to **approximate** sampling distributions.

Under general conditions, when n is sufficiently large $\hat{\beta}$ is approximately **normally distributed** with mean vector β and covariance matrix

$$\phi(\mathbf{X}^T \Delta \mathbf{V} \Delta \mathbf{X})^{-1}.$$

This fact can be used to test hypotheses about β or construct confidence regions for β .