Statistics for Applications

Chapter 10: Generalized Linear Models (GLMs)

Linear model

A linear model assumes

$$Y|X \sim \mathcal{N}(\mu(X), \sigma^2 I),$$

And

$$\mathbb{E}(Y|X) = \mu(X) = X^\top \beta,$$

Components of a linear model

The two components (that we are going to relax) are

1. Random component: the response variable $Y|X$ is continuous and normally distributed with mean $\mu = \mu(X) = \mathbb{E}(Y|X)$.

2. Link: between the random and covariates $X = (X^{(1)}, X^{(2)}, \cdots, X^{(p)})^\top$: $\mu(X) = X^\top \beta$.

Generalization

A generalized linear model (GLM) generalizes normal linear regression models in the following directions.

1. Random component:

$$Y \sim \text{some exponential family distribution}$$

2. Link: between the random and covariates:

$$g\big(\mu(X)\big) = X^\top \beta$$

   where g called link function and $\mu = \mathbb{E}(Y|X)$.

Example 1: Disease Occuring Rate

In the early stages of a disease epidemic, the rate at which new cases occur can often increase exponentially through time. Hence, if μ_i is the expected number of new cases on day t_i, a model of the form

$$\mu_i = \gamma \exp(\delta t_i)$$

seems appropriate.

- Such a model can be turned into GLM form, by using a log link so that

$$\log(\mu_i) = \log(\gamma) + \delta t_i = \beta_0 + \beta_1 t_i.$$

- Since this is a count, the Poisson distribution (with expected value μ_i) is probably a reasonable distribution to try.

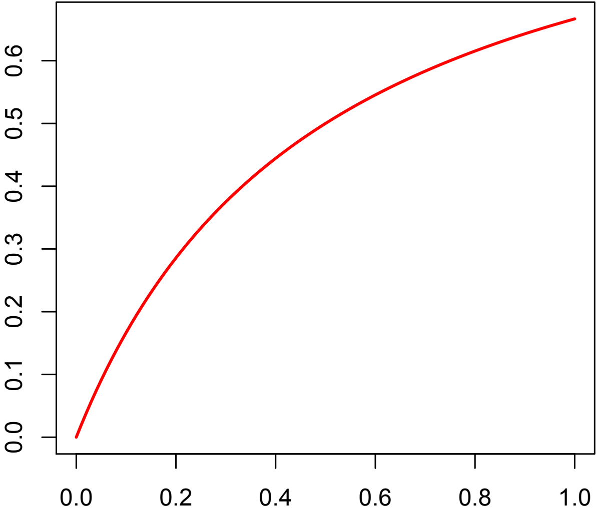
Example 2: Prey Capture Rate(1)

The rate of capture of preys, y_i, by a hunting animal, tends to increase with increasing density of prey, x_i, but to eventually level off, when the predator is catching as much as it can cope with. A suitable model for this situation might be

$$\mu_i = \frac{\alpha x_i}{h + x_i},$$

where α represents the maximum capture rate, and h represents the prey density at which the capture rate is half the maximum rate.

Example 2: Prey Capture Rate (2)



Example 2: Prey Capture Rate (3)

- Obviously this model is non-linear in its parameters, but, by using a reciprocal link, the right-hand side can be made linear in the parameters,

$$g(\mu_i) = \frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{h}{\alpha}\frac{1}{x_i} = \beta_0 + \beta_1 \frac{1}{x_i}.$$

- The standard deviation of capture rate might be approximately proportional to the mean rate, suggesting the use of a Gamma distribution for the response.

Example 3: Kyphosis Data

The Kyphosis data consist of measurements on 81 children following corrective spinal surgery. The binary response variable, Kyphosis, indicates the presence or absence of a postoperative deforming. The three covariates are, Age of the child in month, Number of the vertebrae involved in the operation, and the Start of the range of the vertebrae involved.

- The response variable is binary so there is no choice: $Y|X$ is Bernoulli with expected value $\mu(X) \in (0,1)$.
- We cannot write
$$\mu(X) = X^\top \beta$$
because the right-hand side ranges through $\mathbb{R}$.
- We need an invertible function f such that $f(X^\top \beta) \in (0,1)$

GLM: motivation

- clearly, normal LM is not appropriate for these examples;
- need a more general regression framework to account for various types of response data
  - Exponential family distributions
- develop methods for model fitting and inferences in this framework
  - Maximum Likelihood estimation.

Exponential Family

A family of distribution $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be a k-parameter exponential family on $\mathbb{R}^q$, if there exist real valued functions:

- $\eta_1, \eta_2, \cdots, \eta_k$ and B of θ,
- T_1, $T_2, \cdots, T_k$, and h of $x \in \mathbb{R}^q$ such that the density function (pmf or pdf) of P_θ can be written as

$$p_\theta(x) = \exp[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta)] h(x)$$

Normal distribution example

- Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. The density is

$$p_\theta(x) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\right)\frac{1}{\sigma\sqrt{2\pi}},$$

which forms a two-parameter exponential family with

$$\eta_1 = \frac{\mu}{\sigma^2},\ \eta_2 = -\frac{1}{2\sigma^2},\ T_1(x) = x,\ T_2(x) = x^2,$$

$$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}),\ h(x) = 1.$$

- When σ^2 is known, it becomes a one-parameter exponential family on $\mathbb{R}$:

$$\eta = \frac{\mu}{\sigma^2},\ T(x) = x,\ B(\theta) = \frac{\mu^2}{2\sigma^2},\ h(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}.$$

Examples of discrete distributions

The following distributions form discrete exponential families of distributions with pmf

- Bernoulli(p): $\quad p^x(1-p)^{1-x}$, $x \in \{0,1\}$
- Poisson(λ): $\dfrac{\lambda^x}{x!}e^{-\lambda}$, $x = 0, 1, \ldots$.

Examples of Continuous distributions

The following distributions form continuous exponential families of distributions with pdf:

- Gamma(a, b): $\frac{1}{\Gamma(a)b^a}x^{a-1}e^{-\frac{x}{b}}$;
  - above: a: shape parameter, b: scale parameter
  - reparametrize: $\mu = ab$: mean parameter

$$\frac{1}{\Gamma(a)}\left(\frac{a}{\mu}\right)^a x^{a-1}e^{-\frac{ax}{\mu}}.$$

- Inverse Gamma(α, β): $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$.
- Inverse Gaussian(μ, σ^2): $\sqrt{\frac{\sigma^2}{2\pi x^3}}e^{\frac{-\sigma^2(x-\mu)^2}{2\mu^2 x}}$.

Others: Chi-square, Beta, Binomial, Negative binomial distributions.

Components of GLM

1. Random component:

$$Y \sim \text{some exponential family distribution}$$

2. Link: between the random and covariates:

$$g\big(\mu(X)\big) = X^\top \beta$$

where g called link function and $\mu(X) = \mathbb{E}(Y|X)$.

One-parameter canonical exponential family

- Canonical exponential family for $k = 1$, $y \in \mathbb{R}$

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

  for some *known* functions $b(\cdot)$ and $c(\cdot, \cdot)$.

- If ϕ is known, this is a one-parameter exponential family with θ being the canonical parameter .
- If ϕ is unknown, this may/may not be a two-parameter exponential family. ϕ is called dispersion parameter.
- In this class, we always assume that ϕ is *known*.

Normal distribution example

- Consider the following Normal density function with known variance σ^2,

$$
\begin{aligned}
f_\theta(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\
&= \exp\left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\},
\end{aligned}
$$

- Therefore $\theta = \mu, \phi = \sigma^2, \ , b(\theta) = \frac{\theta^2}{2}$, and

$$
c(y, \phi) = -\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi)).
$$

Other distributions

Table 1: Exponential Family

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of y	$(-\infty, \infty)$	$[0, -\infty)$	$\{0, 1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1 + e^\theta)$
$c(y, \phi)$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	$-\log y!$	1

Likelihood

Let $\ell(\theta) = \log f_\theta(Y)$ denote the log-likelihood function.
The mean $\mathbb{E}(Y)$ and the variance $\mathrm{var}(Y)$ can be derived from the following identities

- First identity
$$\mathbb{E}(\frac{\partial \ell}{\partial \theta}) = 0,$$
- Second identity
$$\mathbb{E}(\frac{\partial^2 \ell}{\partial \theta^2}) + \mathbb{E}(\frac{\partial \ell}{\partial \theta})^2 = 0.$$

Obtained from $\displaystyle\int f_\theta(y)dy \equiv 1$.

Expected value

Note that

$$\ell(\theta) = \frac{Y\theta - b(\theta}{\phi} + c(Y;\phi),$$

Therefore

$$\frac{\partial \ell}{\partial \theta} = \frac{Y - b'(\theta)}{\phi}$$

It yields

$$0 = \mathbb{E}(\frac{\partial \ell}{\partial \theta}) = \frac{\mathbb{E}(Y) - b'(\theta))}{\phi},$$

which leads to

$$\mathbb{E}(Y) = \mu = b'(\theta).$$

Variance

On the other hand we have we have

$$\frac{\partial^2 \ell}{\partial \theta^2} + (\frac{\partial \ell}{\partial \theta})^2 = -\frac{b''(\theta)}{\phi} + \Big(\frac{Y - b'(\theta)}{\phi}\Big)^2$$

and from the previous result,

$$\frac{Y - b'(\theta)}{\phi} = \frac{Y - \mathbb{E}(Y)}{\phi}$$

Together, with the second identity, this yields

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\mathrm{var}(Y)}{\phi^2},$$

which leads to

$$\mathrm{var}(Y) = V(Y) = b''(\theta)\phi.$$

Example: Poisson distribution

Example: Consider a Poisson likelihood,

$$f(y) = \frac{\mu^y}{y!} e^{-\mu} = e^{y \log \mu - \mu - \log(y!)},$$

Thus,

$$\theta = \log \mu, \quad b(\theta) = \mu, \quad c(y, \phi) = -\log(y!),$$

$$\phi = 1,$$

$$\mu = e^{\theta},$$

$$b(\theta) = e^{\theta},$$

$$b''(\theta) = e^{\theta} = \mu,$$

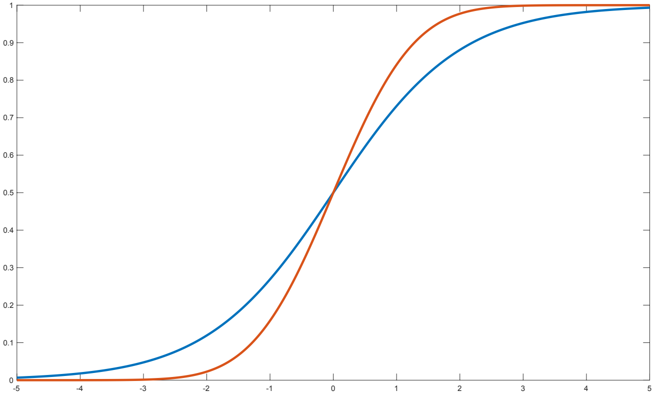
Link function

- β is the parameter of interest, and needs to appear somehow in the likelihood function to use maximum likelihood.
- A link function g relates the linear predictor $X^\top \beta$ to the mean parameter μ,
$$X^\top \beta = g(\mu).$$
- g is required to be monotone increasing and differentiable
$$\mu = g^{-1}(X^\top \beta).$$

Examples of link functions

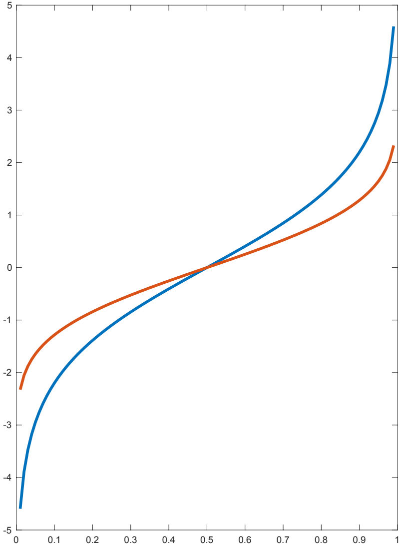
- For LM, $g(\cdot)$ = identity.
- Poisson data. Suppose $Y|X \sim \text{Poisson}(\mu(X))$.
  - $\mu(X) > 0$;
  - $\log(\mu(X)) = X^\top \beta$;
  - In general, a link function for the count data should map $(0, +\infty)$ to $\mathbb{R}$.
  - The log link is a natural one.
- Bernoulli/Binomial data.
  - $0 < \mu < 1$;
  - g should map $(0, 1)$ to $\mathbb{R}$:
  - 3 choices:
    1. logit: $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^\top \beta$;
    2. probit: $\Phi^{-1}(\mu(X)) = X^\top \beta$ where $\Phi(\cdot)$ is the normal cdf;
    3. complementary log-log: $\log(-\log(1 - \mu(X))) = X^\top \beta$
  - The logit link is the natural choice.

Examples of link functions for Bernoulli response (1)



- in blue: $f_1(x) = \dfrac{e^x}{1+e^x}$
- in red: $f_2(x) = \Phi(x)$ (Gaussian CDF)

Examples of link functions for Bernoulli response (2)



- in blue: $g_1(x) = f_1^{-1}(x) = \log \dfrac{x}{1-x}$ (logit link)
- in red: $g_2(x) = f_2^{-1}(x) = \Phi^{-1}(x)$ (probit link)

Canonical Link

- The function g that links the mean μ to the canonical parameter θ is called Canonical Link:

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

- Since $\mu = b'(\theta)$, the canonical link is given by

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

- If $\phi > 0$, the canonical link function is strictly increasing. `Why?`

Example: the Bernoulli distribution

- We can check that

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

- Hence we solve

$$b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \mu \qquad \Leftrightarrow \qquad \theta = \log\left(\frac{\mu}{1 - \mu}\right)$$

- The canonical link for the Bernoulli distribution is the logit link.

Other examples

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	μ
Poisson	$\exp(\theta)$	$\log \mu$
Bernoulli	$\log(1 + e^{\theta})$	$\log \frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

Model and notation

- Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ be independent random pairs such that the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family:

$$f_{\theta_i}(y_i) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\Big\}.$$

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbb{X} = (X_1^\top, \ldots, X_n^\top)^\top$
- Here the mean μ_i is related to the canonical parameter θ_i via

$$\mu_i = b'(\theta_i)$$

- and μ_i depends linearly on the covariates through a link function g:

$$g(\mu_i) = X_i^\top \beta\,.$$

Back to β

- Given a link function g, note the following relationship between β and θ:

$$\begin{aligned} \theta_i &= (b')^{-1}(\mu_i) \\ &= (b')^{-1}(g^{-1}(X_i^\top \beta)) \equiv h(X_i^\top \beta), \end{aligned}$$

  where h is defined as

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}.$$

- Remark: if g is the canonical link function, h is identity.

Log-likelihood

- The log-likelihood is given by

$$\begin{aligned}\ell_n(\beta; \mathbf{Y}, \mathbb{X}) &= \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} \\ &= \sum_i \frac{Y_i h(X_i^\top \beta) - b(h(X_i^\top \beta))}{\phi}\end{aligned}$$

  up to a constant term.

- Note that when we use the canonical link function, we obtain the simpler expression

$$\ell_n(\beta, \phi; \mathbf{Y}, \mathbb{X}) = \sum_i \frac{Y_i X_i^\top \beta - b(X_i^\top \beta)}{\phi}$$

Strict concavity

- The log-likelihood $\ell(\theta)$ is strictly concave using the canonical function when $\phi > 0$. `Why?`
- As a consequence the maximum likelihood estimator is unique.
- On the other hand, if another parameterization is used, the likelihood function may not be strictly concave leading to several local maxima.

Optimization Methods

Given a function $f(x)$ defined on $\mathcal{X} \subset \mathbb{R}^m$, find x^* such that $f(x^*) \geq f(x)$ for all $x \in \mathcal{X}$.

We will describe the following three methods,

- Newton-Raphson Method
- Fisher-scoring Method
- Iteratively Re-weighted Least Squares.

Gradient and Hessian

- Suppose $f : \mathbb{R}^m \to \mathbb{R}$ has two continuous derivatives.
- Define the Gradient of f at point x_0, $\nabla_f = \nabla_f(x_0)$, as

$$(\nabla_f) = (\partial f/\partial x_1, \ldots, \partial f/\partial x_m)^\top.$$

- Define the Hessian (matrix) of f at point x_0, $H_f = H_f(x_0)$, as

$$(H_f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

- For smooth functions, the Hessian is symmetric. If f is strictly concave, then $H_f(x)$ is negative definite.
- The continuous function:

$$x \mapsto H_f(x)$$

is called Hessian map.

Quadratic approximation

- Suppose f has a continuous Hessian map at x_0. Then we can approximate f quadratically in a neighborhood of x_0 using

$$f(x) \approx f(x_0) + \nabla_f^\top(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^\top H_f(x_0)(x - x_0).$$

- This leads to the following approximation to the gradient:

$$\nabla_f(x) \approx \nabla_f(x_0) + H_f(x_0)(x - x_0).$$

- If x^* is maximum, we have

$$\nabla_f(x^*) = 0$$

- We can solve for it by plugging in x^*, which gives us

$$x^* = x_0 - H_f(x_0)^{-1}\nabla_f(x_0).$$

Newton-Raphson method

- The Newton-Raphson method for multidimensional optimization uses such approximations sequentially
- We can define a sequence of iterations starting at an arbitrary value x_0, and update using the rule,

$$x^{(k+1)} = x^{(k)} - H_f(x^{(k)})^{-1} \nabla_f(x^{(k)}).$$

- The Newton-Raphson algorithm is globally convergent at quadratic rate whenever f is concave and has two continuous derivatives.

Fisher-scoring method (1)

- Newton-Raphson works for a deterministic case, which does not have to involve random data.
- Sometimes, calculation of the Hessian matrix is quite complicated (we will see an example)
- Goal: use directly the fact that we are minimizing the KL divergence

$$\text{KL} \text{``}=\text{''} -\mathbb{E}\big[\text{log-likelihood}\big]$$

- Idea: replace the Hessian with its expected value. Recall that

$$\mathbb{E}_\theta\left(H_{\ell_n}(\theta)\right) = -I(\theta)$$

is the Fisher Information

Fisher-scoring method (2)

- The Fisher Information matrix is positive definite, and can serve as a stand-in for the Hessian in the Newton-Raphson algorithm, giving the update:

$$\theta^{(k+1)} = \theta^{(k)} + I(\theta^{(k)})^{-1}\nabla_{\ell_n}(\theta^{(k)}).$$

  This is the Fisher-scoring algorithm.
- It has essentially the same convergence properties as Newton-Raphson, but it is often easier to compute I than H_{ℓ_n}.

Example: Logistic Regression (1)

- Suppose $Y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n$, are independent 0/1 indicator responses, and X_i is a $p \times 1$ vector of predictors for individual i.
- The log-likelihood is as follows:

$$\ell_n(\theta|\mathbf{Y}, \mathbb{X}) = \sum_{i=1}^{n} \left(Y_i \theta_i - \log\left(1 + e^{\theta_i}\right)\right).$$

- Under the canonical link,

$$\theta_i = \log\left(\frac{p_i}{1 - p_i}\right) = X_i^\top \beta.$$

Example: Logistic Regression (2)

- Thus, we have

$$\ell_n(\beta|\mathbf{Y}, \mathbb{X}) = \sum_{i=1}^{n} \left(Y_i X_i^\top \beta - \log\left(1 + e^{X_i^\top \beta}\right)\right).$$

- The gradient is

$$\nabla_{\ell_n}(\beta) = \sum_{i=1}^{n} \left(Y_i X_i - \frac{e^{X_i^\top \beta}}{1 + e^{X_i^\top \beta}} X_i \right).$$

- The Hessian is

$$H_{\ell_n}(\beta) = -\sum_{i=1}^{n} \frac{e^{X_i^\top \beta}}{\left(1 + e^{X_i^\top \beta}\right)^2} X_i X_i^\top.$$

- As a result, the updating rule is

$$\beta^{(k+1)} = \beta^{(k)} - H_{\ell_n}(\beta^{(k)})^{-1} \nabla_{\ell_n}(\beta^{(k)}).$$

Example: Logistic Regression (3)

- The score function is a linear combination of the X_i, and the Hessian or Information matrix is a linear combination of $X_i X_i^\top$. This is typical in exponential family regression models (i.e. GLM).
- The Hessian is negative definite, so there is a unique local maximizer, which is also the global maximizer.
- Finally, note that that the Y_i does not appear in $H_{\ell_n}(\beta)$, which yields

$$H_{\ell_n}(\beta) = \mathbb{E}\big[H_{\ell_n}(\beta)\big] = -I(\beta)$$

.

Iteratively Re-weighted Least Squares

- IRLS is an algorithm for fitting GLM obtained by Newton-Raphson/Fisher-scoring.
- Suppose $Y_i|X_i$ has a distribution from an exponential family with the following log-likelihood function,

$$\ell = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi).$$

- Observe that

$$\mu_i = b'(\theta_i),\ \ X_i^\top \beta = g(\mu_i), \frac{d\mu_i}{d\theta_i} = b^{''}(\theta_i) \equiv V_i.$$

$$\theta_i = (b')^{-1} \circ g^{-1}(X_i^\top \beta) := h(X_i^\top \beta)$$

Chain rule

- According to the chain rule, we have

$$
\begin{aligned}
\frac{\partial \ell_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} \\
&= \sum_i \frac{Y_i - \mu_i}{\phi} h'(X_i^\top \beta) X_i^j \\
&= \sum_i (\tilde{\mathbf{Y}}_i - \tilde{\mu}_i) W_i X_i^j \qquad \left(W_i \equiv \left(\frac{h'(X_i^\top \beta)}{g'(\mu_i)\phi} \right) \right).
\end{aligned}
$$

- Where $\tilde{\mathbf{Y}} = (g'(\mu_1)Y_1, \ldots g'(\mu_n)Y_n)^\top$ and $\tilde{\mu} = (g'(\mu_1)\mu_1, \ldots g'(\mu_n)\mu_n)^\top$

Gradient

- Define
$$W = \mathrm{diag}\{W_1, \ldots, W_n\},$$
- Then, the gradient is
$$\nabla_{\ell_n}(\beta) = \mathbb{X}^\top W(\tilde{\mathbf{Y}} - \tilde{\mu})$$

Hessian

- For the Hessian, we have

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_i \frac{Y_i - \mu_i}{\phi} h''(X_i^\top \beta) X_i^j X_i^j$$
$$-\frac{1}{\phi} \sum_i \left(\frac{\partial \mu_i}{\partial \beta_k} \right) h'(X_i^\top \beta) X_i^j$$

- Note that

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial b'(\theta_i)}{\partial \beta_k} = \frac{\partial b'(h(X_i^\top \beta))}{\partial \beta_k} = b''(\theta_i) h'(X_i^\top \beta) X_i^k$$

It yields

$$\mathbb{E}(H_{\ell_n}(\beta)) = -\frac{1}{\phi} \sum_i b''(\theta_i) \left[h'(X_i^\top \beta) \right]^2 X_i X_i^\top$$

Fisher information

- Note that $g^{-1}(\cdot) = b' \circ h(\cdot)$ yields

$$b'' \circ h(\cdot) \cdot h'(\cdot) = \frac{1}{g' \circ g^{-1}(\cdot)}$$

Recall that $\theta_i = h(X_i^\top \beta)$ and $\mu_i = g^{-1}(X_i^\top \beta)$, we obtain

$$b''(\theta_i) h'(X_i^\top \beta) = \frac{1}{g'(\mu_i)}$$

- As a result

$$\mathbb{E}(H_{\ell_n}(\beta)) = -\sum_i \frac{h'(X_i^\top \beta)}{g'(\mu_i)\phi} X_i X_i^\top$$

- Therefore,

$$I(\beta) = -\mathbb{E}(H_{\ell_n}(\beta)) = \mathbb{X}^\top W \mathbb{X} \qquad \text{where} \quad W = \operatorname{diag}\Big(\frac{h'(X_i^\top \beta)}{g'(\mu_i)}\Big)$$

Fisher-scoring updates

- According to Fisher-scoring, we can update an initial estimate $\beta^{(k)}$ to $\beta^{(k+1)}$ using

$$\beta^{(k+1)} = \beta^{(k)} + I(\beta^{(k)})^{-1} \nabla_{\ell_n}(\beta^{(k)}) \,,$$

- which is equivalent to

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} + (\mathbb{X}^\top W \mathbb{X})^{-1} \mathbb{X}^\top W (\tilde{\mathbf{Y}} - \tilde{\mu}) \\ &= (\mathbb{X}^\top W \mathbb{X})^{-1} \mathbb{X}^\top W (\tilde{\mathbf{Y}} - \tilde{\mu} + \mathbb{X}\beta^{(k)}) \end{aligned}$$

Weighted least squares (1)

Let us open a parenthesis to talk about Weighted Least Squares.

- Assume the linear model $\mathbf{Y} = \mathbb{X}\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}_n(0, W^{-1})$, where W^{-1} is a $n \times n$ diagonal matrix. When variances are different, the regression is said to be heteroskedastic.
- The maximum likelihood estimator is given by the solution to

$$\min_{\beta} (\mathbf{Y} - \mathbb{X}\beta)^\top W (\mathbf{Y} - \mathbb{X}\beta)$$

  This is a Weighted Least Squares problem
- The solution is given by

$$(\mathbb{X}^\top W \mathbb{X})^{-1} \mathbb{X}^\top W (\mathbb{X}^\top W \mathbb{X}) \mathbf{Y}$$

- Routinely implemented in statistical software.

Weighted least squares (2)

Back to our problem.
Recall that

$$\beta^{(k+1)} = (\mathbb{X}^\top W \mathbb{X})^{-1} \mathbb{X}^\top W (\tilde{\mathbf{Y}} - \tilde{\mu} + \mathbb{X}\beta^{(k)})$$

- This reminds us of Weighted Least Squares with
  1. $W = W(\beta^{(k)})$ being the weight matrix,
  2. $\tilde{\mathbf{Y}} - \tilde{\mu} + \mathbb{X}\beta^{(k)}$ being the response.

  So we can obtain $\beta^{(k+1)}$ using any system for WLS.

IRLS procedure (1)

Iteratively Reweighed Least Squares is an iterative procedure to compute the MLE in GLMs using weighted least squares.
We show how to go from $\beta^{(k)}$ to $\beta^{(k+1)}$

1. Fix $\beta^{(k)}$ and $\mu_i^{(k)} = g^{-1}(X_i^\top \beta^{(k)})$;
2. Calculate the adjusted dependent responses

$$Z_i^{(k)} = X_i^\top \beta^{(k)} + g'(\mu_i^{(k)})(Y_i - \mu_i^{(k)});$$

3. Compute the weights $W^{(k)} = W(\beta^{(k)})$

$$W^{(k)} = \mathrm{diag} \quad \frac{h'(X_i^\top \beta^{(k)})}{g'(\mu_i^{(k)})\phi}$$

4. Regress $\mathbf{Z}^{(k)}$ on the design matrix $\mathbb{X}$ with weight $W^{(k)}$ to derive a new estimate $\beta^{(k+1)}$;

We can repeat this procedure until convergence.

IRLS procedure (2)

- For this procedure, we only need to know $\mathbb{X}$, $\mathbf{Y}$, the link function $g(\cdot)$ and the variance function $V(\mu) = b^{''}(\theta)$.
- A possible starting value is to let $\mu^{(0)} = \mathbf{Y}$.
- If the canonical link is used, then Fisher scoring is the same as Newton-Raphson.

  $$\mathbb{E}\left(H_{\ell_n}\right) = H_{\ell_n}.$$

  There is no random component $(\mathbf{Y})$ in the Hessian matrix.