MSc Applied Statistics Programmes: Statistical Analysis (Spring Term)

1 The Generalized Linear Model

1.1 Introduction

The *Generalized Linear Model* was introduced by Nelder and Wedderburn (1972) in JRSS (A) and extends the assumptions underpinning the *General Linear Model*, which was treated extensively in the Autumn Term.

Generalized Linear Models (which we shall sometimes call GLMs for short) take into account a much wider class of relationships between the mean of the dependent variable and the explanatory variables, as well as a much wider class of error structure.

In the first part of this chapter, we briefly review the General Linear Model and its assumptions. We then go on to describe the extensions to this model that lead to the GLM.

1.2 Review of the General Linear Model

Each member i of a sample of size n is associated with an observation on a response (dependent) random variable Y, which in turn is associated with values on p explanatory variables $X_1, X_2, \ldots, X_p$. Note that the explanatory variables may be continuous or categorical (factors and their interactions). Their values may be corresponding observations on associated random variables (measured without error) or controlled by the design.

Let the observed value for Y be y_i and the observed (or controlled) values for the explanatory variables $X_1, X_2, \ldots, X_p$ be $x_{i1}, x_{i2}, \ldots, x_{ip}$ for the i-th individual in the sample.

We relate the observed value of Y with the corresponding values for $X_1, X_2, \ldots, X_p$ via the relation:

$$y_i \;=\; \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \;=\; \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad i=1,\ldots,n, \tag{1}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are a set of fixed (but usually unknown) parameters, and the ϵ_i, $i=1,\ldots,n$, are assumed NID$(0, \sigma^2)$.

From this construction, we can see that the mean of the i-th dependent variable is:

$$E[Y_i|X_1 = x_{i1}, \ldots, X_p = x_{ip}] \;=\; E[Y_i] \;=\; \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \;=\; \mathbf{x}_i'\boldsymbol{\beta}. \tag{2}$$

So y_i is described by its mean, a linear function in the parameters $\beta_1, \ldots, \beta_p$, plus a random perturbation ϵ_i taken from the Normal distribution N$(0, \sigma^2)$.

As usual these relationships can be conveniently expressed in vector-matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of parameters, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is an $n \times 1$ vector of random perturbations, and

$\mathbf{X}$ is an $n \times p$ *design matrix*, which takes the form

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

A constant term is incorporated into this model by setting $x_{11} = x_{21} = \ldots = x_{n1} = 1$

The above formulation, known as the *General Linear Model*, can be characterized by the following 3 conditions:

(I′) The response r.v.'s $Y_i, i = 1, \ldots, n$ have the NID(μ_i, σ^2) distribution.

(II′) The *linear predictor* for the i-th observation is defined as

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

for $i = 1, \ldots, n$.

(III′) The mean of the i-th observation equals the linear predictor, i.e.

$$E[Y_i] = \mu_i = \eta_i$$

for $i = 1, \ldots, n$.

Remarks 1.1

(i) The least squares estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ say, is the vector that minimizes

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \left(= \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \right)$$

with respect to $\boldsymbol{\beta}$, and for this model the least squares estimate $\hat{\boldsymbol{\beta}}$ is also the *maximum likelihood estimate*. This is easy to see if we recall that the observations y_i are assumed to be independently sampled from distributions N$(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$. So the corresponding *Likelihood function*

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \end{aligned}$$

The maximum likelihood estimate of $\boldsymbol{\beta}$ is the vector that maximizes $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ with respect to $\boldsymbol{\beta}$, and clearly this is $\hat{\boldsymbol{\beta}}$, the vector that minimizes

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

(ii) The residual sum-of-squares is used to estimate σ^2 and as the basis of goodness-of-fit measures. We shall see later that it is also proportional to a Likelihood Ratio goodness-of-fit statistic.

(iii) Changes in the residual sum-of-squares, as terms are added to or subtracted from the model, are used to test the importance of these terms. Where there is an hierarchical sequence of models a sequential sum-of-squares (or ANOVA) table can be built up and a series of tests, starting with the largest model, can be carried out to select the most appropriate model. These tests use ratios of mean squares tested as F distributions; it will be seen that they are Likelihood Ratio tests.

Recall also that:

(iv) $\hat{\boldsymbol{\beta}}$ satisfies the *Normal equations*, which, in matrix form, are given by

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}, \tag{4}$$

where these are derived by taking partial derivatives of the sum-of-squares (or log-likelihood) criterion and equating to zero.

(v) If $\mathbf{X}$ is of full rank (equal to p if $n \geq p$), then $\widehat{\boldsymbol{\beta}}$ is unique, and is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5}$$

Moreover $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which can be approximated using our estimate of σ^2.

1.3 Some Drawbacks with the General Linear Model

Under the general linear model assumptions, the $\{Y_i\}$ are continuous, Normally distributed (range $\mathbb{R}$), and have the same variance. Moreover the $E[Y_i] = \mu_i$ are *linearly* related to the explanatory variables and random errors are additive. These are substantial limitations. We may wish to model count data (eg Poisson, range $\mathbb{N}$), binomial proportions (constrained to lie between 0 and 1), or lifetime data (nonnegative and often substantially skewed). The variance may be a function of the mean rather than a constant, and the effect of explanatory variables on the mean may not be linear. Consider a very common example, a dose-response model with binomial data.

Example 1.2 (A dose-response model, *Bliss (1935)*)
The table below shows the numbers of insects that have died after several hours exposure to a certain pesticide.

Dose x_i $\log_{10}\mathrm{CS}_2$ mgl^{-1}	Number of insects, n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

These data are binomial. Now consider the problem of determining a relationship, or regression equation, between the proportion killed and the dose. Clearly, the observed proportion killed, at dose x_i, is equal to $p_i = y_i/n_i$. We wish to model $E[p_i] = \pi_i$ in terms of the dose. But $0 \leq p_i, \pi_i \leq 1$ and so a linear regression is not really appropriate.

Instead we consider a very simple logistic regression model (not the only possibility):

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

Here $0 \leq \pi_i \leq 1$ and, provided $\beta_2 > 0$, $\pi_i \to 1$ as dose $\to \infty$ and $\pi_i \to 0$ as dose $\to -\infty$. We can reorganize this equation to obtain

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i, \tag{6}$$

where we see that a *nonlinear* function of the mean π_i is modelled as a linear function of the explanatory variable x_i. To fit a model of this type to the data, we need to develop an extension to our previous model assumptions, as well as to procedures for the estimation and testing of the parameters.

Remarks 1.3
(i) Equation (6) is an example of a *logistic* regression and is readily interpretable.

(ii) The ratio $\pi/(1-\pi)$ is the ratio of the probability of success to the probability of failure, known as the *odds* of success. The term on the left hand side is known as the *log odds*.

(iii) As well as calculating the odds o from the probability π as $o = \pi/(1-\pi)$, we can calculate the probability from the odds as $\pi = o/(1+o)$.

(iv) How should we interpret β_1 and β_2, the parameter that determines the effect of changes in dose? We can write the model in (6) as

$$\text{odds of success} \;=\; \pi/(1-\pi) \;=\; \exp(\beta_1)\,\exp(\beta_2 x).$$

Clearly $\exp(\beta_1)$ represents the odds of success at $x = 0$, and $\exp(\beta_2)$ is the odds *multiplier*, the amount by which the odds of success are multiplied for any unit increase in explanatory variable x (the dose).

(v) When comparing probabilities, a useful statistic is the *odds ratio*. The odds multiplier $\exp(\beta_2)$ is actually an odds ratio since, for any dose x, it can be expressed as

$$\exp(\beta_2) \;=\; \frac{\text{odds of success for dose } (x+1)}{\text{odds of success for dose } x}.$$

In order to fit such (*generalized linear*) models, we clearly need to consider a wider class of error structure (of which the Normal distribution is a member) and a wider class of relationships between the mean and the explanatory variables.

1.4 The Generalized Exponential Family of Distributions

Definition 1.4 (Generalized Exponential Family)
Y is said to have a distribution from within the *generalized exponential family* if its p.d.f/p.m.f., $f(y)$ say, can be expressed as

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \tag{7}$$

for functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and parameters θ and ϕ. The range space (or support) of Y, that is the set $\{y : f(y;\theta,\phi) > 0\} \subseteq \mathbb{R}$, is also required to be independent of the parameters.

Remarks 1.5
(i) In all our examples $a(\phi) = \phi$. The only extension we allow is ϕ/w, where w is a given weighting constant that is allowed to vary from observation to observation.

(ii) The parameter θ is a location parameter and is termed the *natural* or *canonical* parameter. With $a(\phi) = \phi$ (or ϕ/w) the parameter ϕ is known as the *scale* or *dispersion* parameter.

Proposition 1.6 (Mean and Variance of Gen. Exp. family)
Suppose Y is a r.v. from the generalized exponential family. Then

$$\mathrm{E}[Y] = b'(\theta) \tag{8}$$
$$\mathrm{Var}(Y) = a(\phi)b''(\theta). \tag{9}$$

Proof See the Appendix to Chapter 1.

Note that we can also write:

$$\mathrm{Var}(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu),$$

where $V(\mu)$ is called the *variance function*.

1.5 The Generalized Linear Model

Here, we introduce the conditions which characterize the Generalized Linear Model (GLM). Essentially, we extend conditions (I′) and (III′) introduced earlier.

(I) The $\{Y_i\}$ are independent random variables sharing the same form of distribution from the generalized exponential family.

(II) The *linear predictor* for the i-th observation is

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad i = 1, \ldots, n.$$

(III)

$$g(\mu_i) = \eta_i,$$

where $\mu_i = E[Y_i]$. The function $g(\cdot)$ is known as the *link function* and is required to be invertible.

Remarks 1.7
(i) More precisely, statement (I) says that the $\{Y_i\}$ are from a particular distribution from the generalized exponential family (e.g. all Normal, all Binomial, etc.), but with θ varying with i. [$a(\phi) = \phi$ is assumed constant, but we allow known weights w to vary with i.]

(ii) Instead of μ_i being linearly related to η_i, we now have the more general (possibly non-linear) relationship $\mu_i = g^{-1}(\eta_i)$, for invertible g.

Example 1.8
Show that the Binomial distribution is a member of the generalized exponential family.

Solution: The following probability mass function is for the Binomial distribution with parameters $n \in \mathbb{N}$ and $\pi \in (0, 1)$, i.e. for $Y \sim Bin(n, \pi)$.

$$f(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \qquad y \in \{0, \ldots, n\}$$

The natural way to express this as a member of the generalized exponential family is to consider the random variable $P = Y/n$, (the *proportion* of 'successes'), with p.m.f.

$$\begin{aligned} f(p) &= \binom{n}{np}\pi^{np}(1-\pi)^{n-np} = \exp\left\{\log\left[\binom{n}{np}\pi^{np}(1-\pi)^{n-np}\right]\right\} \\ &= \exp\left\{np\log(\pi) + (n-np)\log(1-\pi) + \log\binom{n}{np}\right\} \\ &= \exp\left\{n\left[p\log(\pi) + \log(1-\pi) - p\log(1-\pi)\right] + \log\binom{n}{np}\right\} \\ &= \exp\left\{\frac{1}{n^{-1}}\left[p\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right] + \log\binom{n}{np}\right\} \end{aligned}$$

Now, set $\theta = \log\left(\frac{\pi}{1-\pi}\right)$. Then $\pi = \frac{e^\theta}{1+e^\theta}$ which implies that $1-\pi = \frac{1}{1+e^\theta}$.

Set $b(\theta) = -\log(1-\pi) = -\log(1+e^\theta)^{-1} = \log(1+e^\theta)$.

Also set $a(\phi) = \phi/n = 1/n$ and $c(p, \phi) = \log\binom{n}{np}$.

This is now the standard form for a distribution of the generalized exponential family. Note that $a(\phi) = \phi/n = 1/n$, so that n may be considered as the *weight* discussed earlier. Furthermore,

$$E[P] = b'(\theta) = \frac{d}{d\theta}\log(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi$$

and

$$E[Y] = n\pi.$$

Also

$$\text{var}(P) = a(\phi)b''(\theta) = \frac{1}{n}\frac{d}{d\theta}b'(\theta) = \frac{1}{n}\frac{e^\theta}{(1+e^\theta)^2} = \frac{1}{n}\frac{e^\theta}{1+e^\theta} \times \frac{1}{1+e^\theta} = \frac{1}{n}\pi(1-\pi)$$

so that

$$\text{var}(Y) = \text{var}(nP) = n\pi(1-\pi).$$

Example 1.9
Show that the Poisson distribution is a member of the generalized exponential family.

Solution: We suppose that $Y \sim Po(\lambda)$.

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp\left\{\log\left(\frac{\lambda^y}{y!}\right)\right\}\exp\{-\lambda\}$$

$$= \exp\{y\log\lambda - \lambda - \log y!\}$$

with

$$\theta = \log\lambda, \quad a(\phi) = \phi = 1, \quad b(\theta) = e^\theta, \quad c(y,\phi) = -\log y!.$$

Thus $f(\cdot)$ is of the correct form. Furthermore,

$$E[Y] = b'(\theta) = e^\theta = \lambda$$
$$\text{var}(Y) = a(\phi)b''(\theta) = 1.e^\theta = e^\theta = \lambda$$

Example 1.10 Show that the $N(\mu, \sigma^2)$ distribution is a member of the generalized exponential family.

Solution: The p.d.f. - evaluated at y - is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

which can be re-arranged to give

$$\exp\left\{\frac{\mu y - \mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}.$$

So, setting $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$, $a(\phi) = \phi = \sigma^2$, and $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$, we have the correct form. Furthermore,

$$E[Y] = b'(\theta) = \mu$$
$$\text{var}[Y] = a(\phi)b''(\theta) = \sigma^2.1 = \sigma^2$$

1.6 Link Functions

For any particular problem, it is possible that there may be several plausible candidates for the link function. An interesting challenge is to propose the most suitable candidate for the problem at hand. See Exercises 1, question 3 for some examples of relationships that can be modelled with the right choice of link function and linear predictor.

However, for particular error structures satisfying (I), we can associate with each of them a default, or *canonical* link function. The canonical link is defined to be that function which maps the mean μ to the canonical parameter, θ. That is, that function g for which

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

Example 1.11 *(Canonical link for $Bin(n, \pi)$)*
In Example 1.8, it was shown that the natural way to express $Y \sim Bin(n, \pi)$ as a member of the exponential family is to consider the random variable $P = Y/n$. Then $a(\phi) = \phi/n = 1/n$ and

$$\pi = b'(\theta) = \frac{e^\theta}{1 + e^\theta}. \tag{10}$$

The canonical link is found by making θ the subject of the formula. Now (10) implies that

$$\begin{aligned} \pi(1 + e^\theta) &= e^\theta \\ \Rightarrow \quad \pi &= (1 - \pi)e^\theta \\ \Rightarrow \quad e^\theta &= \tfrac{\pi}{1-\pi} \\ \Rightarrow \quad \theta &= \log\left(\tfrac{\pi}{1-\pi}\right). \end{aligned}$$

Hence

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right).$$

This is known as the *logit link* function.

Example 1.12 *(Canonical link for $Po(\lambda)$)*

$$\mu = e^\theta \Rightarrow \theta = \log \mu$$

This means that the canonical link for the Poisson distribution is $g(\mu) = \log(\mu)$, which is known as the the *log link* function.

What about the Normal distribution?

Example 1.13 *(Canonical link for $N(\mu, \sigma^2)$)*

Since $\theta = \mu$ (and $b(\theta) = \frac{\theta^2}{2}$), the canonical link is $g(\mu) = \mu$, known as the *identity* link.

It turns out, therefore, that with the Normal error structure, along with its canonical link, we recover the General Linear Model (which was characterized by (I′) – (III′)).

Appendix to Section 1

Proof of Proposition 1.6
Since $f(\cdot;\theta,\phi)$ is a p.d.f. (proof for discrete case similar), then

$$\int_{\Omega_Y} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy = 1,$$

where Ω_Y is the support or range space of Y. Differentiating w.r.t. θ yields

$$\int_{\Omega_Y} \frac{\partial}{\partial\theta} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy = 0.$$

This leads to

$$\int_{\Omega_Y} \frac{y - b'(\theta)}{a(\phi)} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy = 0 \tag{11}$$

which, upon rearrangement, yields

$$\int_{\Omega_Y} y \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy = b'(\theta) \int_{\Omega_Y} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy$$

i.e.

$$\int_{\Omega_Y} y f(y;\theta,\phi) dy = b'(\theta) \int_{\Omega_Y} f(y;\theta,\phi) dy.$$

Thus we have the result that

$$E[Y] = b'(\theta) \times 1 = b'(\theta)$$

since $f(\cdot;\theta,\phi)$ is a p.d.f.

To find var(Y), differentiate (11) w.r.t. θ to obtain

$$-\frac{b''(\theta)}{a(\phi)} \int_{\Omega_Y} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy$$

$$+ \int_{\Omega_Y} \frac{(y - b'(\theta))^2}{a^2(\phi)} \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\} dy = 0.$$

This implies that

$$\frac{1}{a^2(\phi)} \int_{\Omega_Y} (y - b'(\theta)))^2 f(y;\theta,\phi) dy = \frac{b''(\theta)}{a(\phi)} \int_{\Omega_Y} f(y;\theta,\phi) dy.$$

$$\frac{1}{a^2(\phi)} \text{var}(Y) = \frac{b''(\theta)}{a(\phi)} \times 1$$
$$\Rightarrow \text{var}(Y) = a(\phi) b''(\theta).$$

2 Parameter Estimation for GLMs

2.1 Introduction

We now need to consider the problem of estimating the parameter vector $\boldsymbol{\beta}$ in the linear predictor of a generalized linear model (GLM). A set of observations are collected on a dependent variable Y and explanatory variables $X_1, X_2, \ldots, X_p$, from a random sample of Y of size n. The aim is to model $E(Y) = \mu$ using a GLM.

Recall the characterization of the GLM:

(I) $\{Y_i\}$ are independent random variables sharing the same form of distribution from the generalized exponential family.

(II) The *linear predictor* for the i-th observation is

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = \mathbf{x}_i' \boldsymbol{\beta} \quad i = 1, \ldots, n.$$

(III)

$$g(\mu_i) = \eta_i,$$

where $\mu_i = E[Y_i]$ and $g(\cdot)$ is the (invertible) link function.

In vector/matrix form,

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

We shall estimate the unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$ by *maximum likelihood*.

Recall that,

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

where ϕ is assumed fixed, but $a_i(\cdot) = \phi / w_i$ and θ_i may vary with i.

Also recall that

$$E[Y_i] = \mu_i = b'(\theta_i)$$

and

$$\mathrm{Var}[Y_i] = a_i(\phi) b''(\theta_i) = a_i(\phi) V(\mu_i),$$

where $V(\mu)$ is the *variance function*, and the *canonical* link is given by

$$g(\mu_i) = \theta_i = \eta_i.$$

The *likelihood function* for the n observations is given by:

$$L(\theta_1, \ldots, \theta_n; \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

where the product-form follows from the independence of the Y_i's.

The *log-likelihood* function is given by

$$l(\theta_1, \ldots, \theta_n; \phi) = \log L(\theta_1, \ldots, \theta_n; \phi)$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} = \sum_{i=1}^{n} l_i$$

where

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

2.2 Derivation of the Likelihood Equations

The goal is to maximize the (log-)likelihood function $l(\theta_1, \ldots, \theta_n; \phi)$ w.r.t. to the parameters $\beta_1, \ldots \beta_p$. That is, we require

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0 \qquad j = 1, \ldots, p$$

Rather than express l as a function of these parameters explicitly, we will instead use the *chain rule*.

By the chain rule, $\dfrac{\partial l_i}{\partial \beta_j}$ can be written as

$$\underset{}{\frac{\partial l_i}{\partial \beta_j}} = \underset{\text{I}}{\frac{\partial l_i}{\partial \theta_i}} \times \underset{\text{II}}{\frac{\partial \theta_i}{\partial \mu_i}} \times \underset{\text{III}}{\frac{\partial \mu_i}{\partial \eta_i}} \times \underset{\text{IV}}{\frac{\partial \eta_i}{\partial \beta_j}} \tag{1}$$

We shall consider each of the terms on the R.H.S. in turn. Note first, however, that if we use the canonical link, so that $\theta_i = \eta_i$, then

$$\underset{}{\frac{\partial l_i}{\partial \beta_j}} = \underset{\text{I}}{\frac{\partial l_i}{\partial \theta_i}} \times \underset{\text{IV}}{\frac{\partial \theta_i}{\partial \beta_j}} \tag{1'}$$

Term I

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

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

Term II

Since $\mu_i = b'(\theta_i)$, then, by implicit differentiation w.r.t. μ_i, we have $1 = b''(\theta_i)\dfrac{\partial \theta_i}{\partial \mu_i}$, so that

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)},$$

where $V(\mu_i)$ is the variance function.

Term III

Recalling that $g(\mu_i) = \eta_i$, then, by implicit differentiation w.r.t. η_i, we have $g'(\mu_i)\dfrac{\partial \mu_i}{\partial \eta_i} = 1$, so that

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}.$$

Term IV

Since

$$\eta_i = \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip}$$

then

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Plugging terms (I)-(IV) into (1) yields

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)} \times \frac{1}{V(\mu_i)} \times \frac{1}{g'(\mu_i)} \times x_{ij}$$

$\Big[$or, for the canonical link, $\dfrac{\partial l_i}{\partial \beta_j} = \dfrac{y_i - \mu_i}{a_i(\phi)} \times x_{ij}\Big]$.

It now follows that

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i)} \frac{(y_i - \mu_i)}{a_i(\phi) V(\mu_i)} \quad j = 1, \ldots, p.$$

Thus, the maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ satisfy

$$S_j(\boldsymbol{\beta}) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i)} \frac{(y_i - \mu_i)}{a_i(\phi) V(\mu_i)} = 0 \quad j = 1, \ldots, p, \tag{2}$$

the *(log-)likelihood equations*, where $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$.

The function

$$S(\boldsymbol{\beta}) = (S_1(\boldsymbol{\beta}), S_2(\boldsymbol{\beta}), \ldots, S_p(\boldsymbol{\beta}))' = \frac{\partial l}{\partial \boldsymbol{\beta}}$$

is termed the *score function* for $\boldsymbol{\beta}$, and the (log-)likelihood equations are equivalent to

$$S(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

[For the canonical link

$$S_j(\boldsymbol{\beta}) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \frac{(y_i - \mu_i)}{a_i(\phi)} = 0 \quad j=1,\ldots,p, \tag{2'}$$

where $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$.

If $a_i(\phi) = \phi$ for all i, then this yields

$$S(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \Rightarrow \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

and then $\mathbf{X}'\mathbf{y}$ gives *sufficient statistics* for $\boldsymbol{\beta}$.]

In the case of the General Linear Model we have the canonical link (the identity) and $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Thus the (log-)likelihood equations (2′) reduce to the familiar Normal equations

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y},$$

which are readily solved to give

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

In general, however, equations (2) are non-linear in the $\{\beta_i\}$, and require numerical iterative techniques for their solution. We explore some of these techniques in the following sections.

2.3 Newton-Raphson Procedure

Let $\mathbf{b}^{(m-1)}$ be a guess at the solutions to the (log-)likelihood equations (2) at stage $m-1$ of the iterative scheme, and let $\mathbf{b}^{(m)}$ be the new updated guess at stage m. If $\mathbf{b}^{(m)}$ is to be correct, then

$$S_j(\mathbf{b}^{(m)}) = 0, \quad j=1,\ldots,p. \tag{3}$$

Using Taylor's theorem, we expand $S_j(\mathbf{b}^{(m)})$ about the previous guess, $\mathbf{b}^{(m-1)}$, yielding the following approximation.

$$0 = S_j\left(\mathbf{b}^{(m)}\right) \approx S_j\left(\mathbf{b}^{(m-1)}\right) + \sum_{i=1}^{p}\left(b_i^{(m)} - b_i^{(m-1)}\right)\left.\frac{\partial S_j}{\partial \beta_i}\right|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}} \quad j=1,\ldots,p$$

i.e.

$$0 = S_j\left(\mathbf{b}^{(m-1)}\right) + \left[\left.\frac{\partial S_j}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}}\right]'(\mathbf{b}^{(m)} - \mathbf{b}^{(m-1)})$$

We write this as

$$0 = U_j^{(m-1)} + \mathbf{H}_j^{(m-1)} \left(\mathbf{b}^{(m)} - \mathbf{b}^{(m-1)} \right)$$

where

$$U_j^{(m-1)} = S_j(\mathbf{b}^{(m-1)})$$

is a scalar quantity, and

$$\mathbf{H}_j^{(m-1)} = \left(\frac{\partial S_j}{\partial \beta_1}, \ldots, \frac{\partial S_j}{\partial \beta_p} \right) \Bigg|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}$$

is a $1 \times p$ row vector.

When $\mathbf{b}^{(m)}$ and $\mathbf{b}^{(m-1)}$ are very close, then the following equations hold approximately

$$0 = U_j^{(m-1)} + \mathbf{H}_j^{(m-1)} \left(\mathbf{b}^{(m)} - \mathbf{b}^{(m-1)} \right), \quad j = 1, \ldots, p.$$

Writing the p equations in matrix form, we have

$$\mathbf{0} = \mathbf{U}^{(m-1)} + \mathbf{H}^{(m-1)} \left(\mathbf{b}^{(m)} - \mathbf{b}^{(m-1)} \right) \tag{4}$$

where

$$\mathbf{U}^{(m-1)} = (U_1^{(m-1)}, \ldots, U_p^{(m-1)})' \quad \left[= S\left(\mathbf{b}^{(m-1)} \right) \right]$$

and $\mathbf{H}^{(m-1)} = \left(h_{jk}^{(m-1)} \right)$ is the $p \times p$ matrix where $h_{jk}^{(m-1)} = \left. \frac{\partial S_j}{\partial \beta_k} \right|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}$.

Hence, if $\mathbf{H}^{(m-1)}$ is invertible, then (4) leads to

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - [\mathbf{H}^{(m-1)}]^{-1} \mathbf{U}^{(m-1)}. \tag{$*$}$$

So given a suitable starting point, $\mathbf{b}^{(0)}$, we can produce a series of iterates $\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots$ until a suitable convergence criterion is satisfied.

The following manipulations put this process into a convenient and interpretable computational form

First, recall that $S_j = \dfrac{\partial l}{\partial \beta_j}$, so that

$$\mathbf{U}^{(m-1)} = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_p} \right)' \Bigg|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}} = \left. \frac{\partial l}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}$$

and

$$\mathbf{H}^{(m-1)} = \left(\frac{\partial S_j}{\partial \beta_k} \right) = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) \Bigg|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}} = \left. \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right|_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}$$

is the *Hessian* matrix for l.

That is, the equations are

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - \left[\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right) \Big|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}} \right]^{-1} \frac{\partial l}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}}$$

2.4 The Method of Scoring

Computation of the elements of $\mathbf{H}(\cdot)$ can sometimes be difficult (if the canonical link is not used). We shall use

$$E\left[\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) \right] \;=\; E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] \;=\; E_{\boldsymbol{\beta}}(\mathbf{H})$$

instead, which is the negative of the Fisher Information matrix, (the inverse of which is the asymptotic covariance matrix for the estimates $\hat{\boldsymbol{\beta}}$).

Now, by the chain rule

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \eta_i} x_{ij}$$

and, by another application of the chain rule, it follows that

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = \frac{\partial^2 l_i}{\partial \eta_i^2} \frac{\partial \eta_i}{\partial \beta_k} x_{ij} = \frac{\partial^2 l_i}{\partial \eta_i^2} x_{ij} x_{ik} \tag{5}$$

The term $\dfrac{\partial^2 l_i}{\partial \eta_i^2}$ in (5) can be computed as follows.

Using the chain rule, we get

$$\frac{\partial l_i}{\partial \eta_i} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i}$$

Applying both the chain rule and the *differentiation of a product rule* to the above equation yields

$$\frac{\partial^2 l_i}{\partial \eta_i^2} = \frac{\partial}{\partial \eta_i} \left(\frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \right) = \left(\frac{\partial^2 l_i}{\partial \theta_i^2} \frac{\partial \theta_i}{\partial \eta_i} \right) \frac{\partial \theta_i}{\partial \eta_i} + \frac{\partial l_i}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \eta_i^2}$$

$$= \frac{\partial^2 l_i}{\partial \theta_i^2} \left(\frac{\partial \theta_i}{\partial \eta_i} \right)^2 + \frac{\partial l_i}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \eta_i^2} = \frac{\partial^2 l_i}{\partial \theta_i^2} \left(\frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \right)^2 + \frac{\partial l_i}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \eta_i^2}$$

$$= \frac{\partial^2 l_i}{\partial \theta_i^2} \left(\frac{\partial \theta_i}{\partial \mu_i} \right)^2 \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + \frac{\partial l_i}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \eta_i^2}.$$

Substituting for l_i and using $\mu_i = b'(\theta_i)$ yields

$$\frac{\partial^2 l_i}{\partial \eta_i^2} = -\frac{b''(\theta_i)}{a_i(\phi)} \left(\frac{1}{b''(\theta_i)} \right)^2 \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + \frac{(y_i - \mu_i)}{a_i(\phi)} \frac{\partial^2 \theta_i}{\partial \eta_i^2}.$$

Plugging this back into (5) we arrive at the expression

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = \left[-\frac{1}{b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + (y_i - \mu_i) \frac{\partial^2 \theta_i}{\partial \eta_i^2} \right] \frac{x_{ij} x_{ik}}{a_i(\phi)}.$$

Hence

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \left[-\frac{1}{b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + (y_i - \mu_i) \frac{\partial^2 \theta_i}{\partial \eta_i^2} \right] \frac{x_{ij} x_{ik}}{a_i(\phi)} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}$$

which is the (j, k)-th element of the Hessian matrix.

In the *method of scoring*, this element is replaced in our iteration scheme by

$$E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right].$$

We have

$$E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = \sum_{i=1}^{n} -\frac{x_{ij} x_{ik}}{a_i(\phi) b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + 0 = -\sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

We can exploit this simplification to produce a (more) concise iteration scheme.

[Note that if the canonical link is used then

$$E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{V(\mu_i)}{a_i(\phi)} x_{ij} x_{ik}.$$

Verify this as an exercise. Note also that in the case of the General Linear Model

$$\frac{V(\mu_i)}{a_i(\phi)} = \frac{1}{\sigma^2}$$

and

$$E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik}.$$

so that

$$E \left[\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) \right] = E \left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} \right] = -\frac{1}{\sigma^2} \mathbf{X}'\mathbf{X}.$$]

Now, observe that

$$\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

is the (j,k)-th element of the matrix $\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix, whose i-th diagonal element is

$$w_{ii} \;=\; \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{g'(\mu_i)^2 \,\text{Var}(Y_i)}.$$

i.e. The w_{ii} may be thought of as *intrinsic weights* associated with the observations y_i arising via the variance-mean relationship and the link function.

$\Big[$ In the case of the canonical link $w_{ii} \;=\; \dfrac{V(\mu_i)}{a_i(\phi)}.\Big]$

So our iteration scheme (from (*)) reduces to

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left[\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\right]^{-1} \mathbf{U}^{(m-1)} \tag{6}$$

where $\mathbf{W}^{(m-1)}$ is equal to $\mathbf{W}|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}}$ and

$$\left[\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\right]^{-1} \;=\; \left[-\text{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right)\right]^{-1}\Bigg|_{\boldsymbol{\beta}=\mathbf{b}^{(m-1)}}$$

We shall manipulate this further into a familiar form. Multiplying through by $\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}$ yields

$$\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}. \tag{7}$$

The R.H.S. of (7) is a $p \times 1$ vector whose j-th component is:

$$\sum_{k=1}^{p}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

$$= \sum_{i=1}^{n} \frac{x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \left[\sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)\right]$$

$$= \sum_{i=1}^{n} \frac{x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 z_i$$

where

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

and μ_i and $\dfrac{\partial \mu_i}{\partial \eta_i}$ are evaluated at $\boldsymbol{\beta} = \mathbf{b}^{(m-1)}$.

Note that z_i can be written

$$\begin{aligned} z_i &= \eta_i^{(m-1)} + (y_i - \mu_i^{(m-1)})g'(\mu_i^{(m-1)}) \\ &= g(\mu_i^{(m-1)}) + (y_i - \mu_i^{(m-1)})g'(\mu_i^{(m-1)}) \end{aligned}$$

i.e. $g(y_i)$ expanded about $\mu_i^{(m-1)}$, and that the associated weight, w_{ii}, is the current estimate of

$$\frac{1}{\mathrm{Var}(Z_i)} \;=\; \frac{1}{g'(\mu_i)^2\,\mathrm{Var}(Y_i)}.$$

So, finally, (7) can be written as

$$\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)} \tag{8}$$

where the i-th element of $\mathbf{z}$ is z_i.

Compare this with the *normal equations* for the general linear model

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

given in Section 1.2 equation (4). The iterative scheme suggested by (8) for obtaining the maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ is known as an *iterative weighted least squares procedure*. R, and a number of other packages that fit GLMs, use modified versions of this algorithm.

At convergence, i.e. when $\mathbf{b}^{(m)}$ is deemed to have converged to the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$, we have

$$\left[-\mathrm{E}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right)\right]^{-1}\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \;=\; [\mathbf{X}'\mathbf{W}\mathbf{X}]^{-1}\big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\,.$$

From likelihood theory, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\left[-\mathrm{E}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right)\right]^{-1},$$

the inverse of the Fisher Information matrix for $\boldsymbol{\beta}$. Evaluating at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ gives us an approximation to $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}})$. Note that if the scale parameter ϕ is unknown, we need to find a way of estimating it.

Compare the corresponding exact result for the General Linear Model given in Section 1.2, equation (5):

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) \;=\; \sigma^2(\mathbf{X}'\mathbf{X})^{-1},\ \text{which is just}\ \left[-\mathrm{E}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right)\right]^{-1} \;=\; \left[-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]^{-1} \text{ (see p 16).}$$

2.5 Issues

A serious issue to contend with is the sensible selection of the starting point. If the seed is too far away from the solution to the equation(s) and/or the function is not sufficiently 'nice', then it is quite possible that the iterative scheme fails to converge. The ad-hoc choice suggested is to invoke a General Linear Model-type least squares estimate, by regressing $g(y_i)$ against the explanatory variables, $i = 1, \ldots, n$. Thus $\mathbf{b}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widetilde{\mathbf{y}}$, where the i-th component of $\widetilde{\mathbf{y}}$ is equal to $g(y_i)$.

Also notice that we solve (8) by multiplying through by the inverse of

$$\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}$$

This may lead to problems if this matrix is *ill-conditioned*, i.e. it has a determinant that is almost zero. In such cases, alternative methods for solving (8) such as *Gaussian elimination* could be explored.

2.6 Exercise using R for Parameter Estimation in a GLM

Consider again the dose-response example 1.2, given in Section 1.3. These data are binomial, and the aim is to use a generalized linear model to model the proportion of insects killed as a function of the dose of insecticide using the canonical logit link function.

Run the following R code. [Note that the link function has not been specified since, by default, the canonical link is always used.]

```
dose <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.861, 1.8839)
number <- c(59, 60, 62, 56, 63, 59, 62, 60)
killed <- c(6, 13, 18, 28, 52, 53, 61, 60)
prop <- killed/number
dose <- dose - 1.78
pesticide <- data.frame(dose, number, killed, prop)
pesticide.glm <- glm(prop ~ dose, family = binomial,
    weights = number, data = pesticide)
summary(pesticide.glm)

x <- seq(-0.105,0.12,0.005)
plot(dose,prop,xlim=c(-0.11,0.12),ylim=c(0,1))
lines(x,predict(pesticide.glm,data.frame(dose=x),type="response"))
```

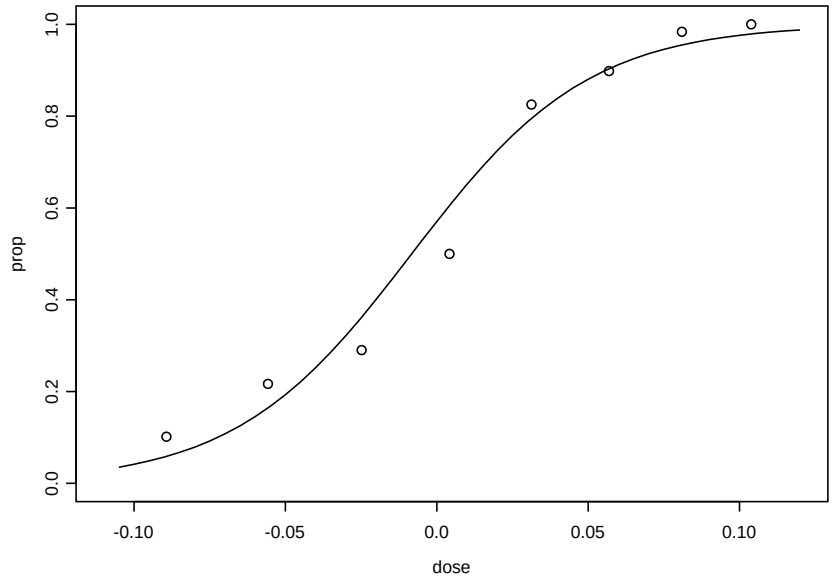
This should produce estimates of the parameters and their standard errors, plus a plot of the observed proportions and the fitted model. It will also produce other statistics some of which you are not yet able to interpret. Ignore these for the present.

Note that a constant 1.78 has been subtracted from the values for dose. The value 1.78 is about the middle of the dose range. The effect is to make the interpretation of the intercept of more interest (how?). You should obtain estimates:

```
Call:
glm(formula = prop ~ dose, family = binomial, data = pesticide,
    weights = number)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2837     0.1308   2.168   0.0301 *
dose         34.2703     2.9121  11.768   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

How can you interpret these parameter estimates? [For example, consider the effect on the odds of being killed of increases in the dose. For example what happens if `dose` is increased by 0.01, or 0.1?]

You should also obtain the following plot.hh



How well do you think the model fits? We shall see later that an improved fit can be obtained using a different link function.