Advanced Statistical Modelling III (Epiphany term)

Department of Mathematical Sciences at Durham University

09 Jan, 2025

Contents

General Information

- These are the lecture notes for the second term of the module MATH3411 – Advanced Statistical Modelling III of Durham University's degree for Mathematics and Statistics.
- **Acknowledgements**: This material is based on the lecture notes in previous modules taught by Dr Samuel Jackson, Dr Jochen Einbeck, Dr Ric Crossman, Dr Emmanuel Ogundimu, Dr Cuong Nguyen, Dr Ian Jermyn, Dr Louis Aslett, and Dr Reza Drikvandi.

Chapter 1

Introduction and Review

1.1 Introduction

In the first term, we have learned about categorical data analysis and some basics of generalised linear models (GLMs). In this term, we will continue to explore GLMs in more detail and study some of its more general variants. In particular, we will learn:

- How to estimate the parameters of a GLM from data.
- How to make a prediction and do inference once a GLM has been fitted.
- How to perform deviance analysis with GLMs.
- How to reduce overdispersion using quasi-likelihood methods.
- How to model repeated measures data using marginal models.
- How to model mixed effects using linear mixed models (LMMs) and generalised linear mixed models (GLMMs).

For the rest of this chapter, we will review some random vector and random matrix identities that will be useful later. Then we will review the basics of GLMs.

1.2 Random Vectors and Random Matrices: A Review

- A *random vector* is a vector of random variables:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

- The mean or expectation of $\boldsymbol{X}$ is defined as:

$$E[\boldsymbol{X}] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}.$$

- A *random matrix* is a matrix of random variables:

$$\boldsymbol{Z} = (Z_{ij}) = \begin{pmatrix} Z_{11} & \dots & Z_{1m} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \dots & Z_{nm} \end{pmatrix}.$$

- The expectation of $\boldsymbol{Z}$ is defined as:

$$E[\boldsymbol{Z}] = (E[Z_{ij}]) = \begin{pmatrix} E[Z_{11}] & \dots & E[Z_{1m}] \\ \vdots & \ddots & \vdots \\ E[Z_{n1}] & \dots & E[Z_{nm}] \end{pmatrix}.$$

- Some properties of random vectors and random matrices:
  - If $\boldsymbol{a}$ is a constant (i.e., non-random) vector, $E[\boldsymbol{a}] = \boldsymbol{a}$.
  - If $\boldsymbol{A}$ is a constant matrix, $E[\boldsymbol{A}] = \boldsymbol{A}$.
  - $E[\boldsymbol{X} + \boldsymbol{Y}] = E[\boldsymbol{X}] + E[\boldsymbol{Y}]$ for any random matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$.
  - $E[\boldsymbol{AX}] = \boldsymbol{A}E[\boldsymbol{X}]$ for a constant matrix $\boldsymbol{A}$ and a random matrix $\boldsymbol{X}$.
  - More generally, $E[\boldsymbol{AXB} + \boldsymbol{C}] = \boldsymbol{A}E[\boldsymbol{X}]\boldsymbol{B} + \boldsymbol{C}$ for constant matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.
- Let $\boldsymbol{X}$ be a random vector. The *covariance matrix* of $\boldsymbol{X}$ is defined as:

$$cov(\boldsymbol{X}) = (cov(X_i, X_j)) = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \dots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & var(X_n) \end{pmatrix}.$$

- We can also write:

$$\begin{aligned} cov(\boldsymbol{X}) &= E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T] \\ &= E\left[\begin{pmatrix} X_1 - E[X_1] \\ \vdots \\ X_n - E[X_n] \end{pmatrix} (X_1 - E[X_1], \dots, X_n - E[X_n])\right]. \end{aligned}$$

- Some properties of covariance matrices:
  - They are symmetric: $cov(\boldsymbol{X}) = cov(\boldsymbol{X})^T$.
  - If $\boldsymbol{a}$ is a constant vector, $cov(\boldsymbol{X} + \boldsymbol{a}) = cov(\boldsymbol{X})$.
  - If $\boldsymbol{A}$ is a constant matrix, $cov(\boldsymbol{AX}) = \boldsymbol{A}cov(\boldsymbol{X})\boldsymbol{A}^T$.
  - $cov(\boldsymbol{X}) = E[\boldsymbol{X}\boldsymbol{X}^T] - E[\boldsymbol{X}]E[\boldsymbol{X}]^T$.
- Let $\boldsymbol{X}_{n\times 1}$ and $\boldsymbol{Y}_{m\times 1}$ be random vectors where n and m could be different. The covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as:

$$cov(\boldsymbol{X}, \boldsymbol{Y}) = (cov(X_i, Y_j)) = \begin{pmatrix} cov(X_1, Y_1) & cov(X_1, Y_2) & \dots & cov(X_1, Y_m) \\ cov(X_2, Y_1) & cov(X_2, Y_2) & \dots & cov(X_2, Y_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, Y_1) & cov(X_n, Y_2) & \dots & cov(X_n, Y_m) \end{pmatrix}.$$

- We can also write:

$$\begin{aligned} cov(\boldsymbol{X}, \boldsymbol{Y}) &= E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T] \\ &= E\left[\begin{pmatrix} X_1 - E[X_1] \\ \vdots \\ X_n - E[X_n] \end{pmatrix} (Y_1 - E[Y_1], \dots, Y_m - E[Y_m]) \right]. \end{aligned}$$

- Some properties of covariance matrices between two vectors:
  - If $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices, $cov(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{A}cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T$.
  - If $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}$, then $cov(\boldsymbol{Z}) = \begin{pmatrix} cov(\boldsymbol{X}) & cov(\boldsymbol{X}, \boldsymbol{Y}) \\ cov(\boldsymbol{Y}, \boldsymbol{X}) & cov(\boldsymbol{Y}) \end{pmatrix}$.
- Let $\boldsymbol{X}$ be a random vector. The *correlation matrix* of $\boldsymbol{X}$ is defined as:

$$corr(\boldsymbol{X}) = (corr(X_i, X_j)) = \begin{pmatrix} 1 & corr(X_1, X_2) & \dots & corr(X_1, X_n) \\ corr(X_2, X_1) & 1 & \dots & corr(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ corr(X_n, X_1) & corr(X_n, X_2) & \dots & 1 \end{pmatrix},$$

where

$$corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i) var(X_j)}}.$$

1.3 Generalised Linear Models: A Review

Generalised linear models (GLMs) are developed by Nelder and Wedderburn [1972] as a way to unify other statistical models as well as the statistical methods operated on these models. In this section, we will briefly review the definition of GLMs. Please refer to the last term's lecture notes for more detailed derivations and examples.

Definition. A GLM is specified through the following components:

- A *linear predictor*: $\eta = \boldsymbol{\beta}^T \boldsymbol{x}$.
- An *injective response function* h, such that $\mu = \mathrm{E}[Y|\boldsymbol{x}, \boldsymbol{\beta}] = h(\eta) = h(\boldsymbol{\beta}^T \boldsymbol{x})$. Equivalently, one can write $g(\mu) = \boldsymbol{\beta}^T \boldsymbol{x}$, where $g = h^{-1}$ is the *link* function.
- The *distributional assumption*: $P(Y|\boldsymbol{x}, \boldsymbol{\beta})$ is an EDF, that is:

$$P(y|\boldsymbol{x}, \boldsymbol{\beta}) = P(y|\theta(\boldsymbol{x}, \boldsymbol{\beta}), \phi(\boldsymbol{x}, \boldsymbol{\beta})) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right). \tag{1.1}$$

From the properties of the EDF, the mean and variance of this distribution are:

$$\mathrm{E}[Y|\theta,\phi] = \mu = b'(\theta) \tag{1.2}$$
$$\mathrm{Var}[Y|\theta,\phi] = \phi\, b''(\theta) = \phi\, b''((b')^{-1}(\mu)) = \phi\, \mathcal{V}(\mu). \tag{1.3}$$

- We also assume *independent data*, that is:

$$P(\{y_i\} \mid \{\boldsymbol{x}_i\}, \boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) \tag{1.4}$$

where $\{y_i, i = 1, ..., n\}$ are response data given the $\{\boldsymbol{x}_i, i = 1, ..., n\}$.

The Natural/Canonical Link. Recall that we have both:

$$\mu = \mathrm{E}[Y|\theta,\phi] \ = b'(\theta) \tag{1.5}$$
$$\mu = \mathrm{E}[Y|\boldsymbol{x},\boldsymbol{\beta}] = h(\boldsymbol{\beta}^T\boldsymbol{x}) = h(\eta) \tag{1.6}$$

with Equation (1.5) holding as a result of $P(y|\theta,\phi)$ following an EDF distribution, and Equation (1.6) holding by definition for a GLM.

The *natural link* is the choice $h = b'$, or equivalently $g = (b')^{-1}$, resulting in the equation

$$\theta = \boldsymbol{\beta}^T\boldsymbol{x} = \eta. \tag{1.7}$$

1.4 Exercises

1. Prove the identities in Section 1.2.
2. The table below gives some common link functions for GLMs. In the table, $\Phi(\cdot)$ is the cdf of the standard normal distribution. Find their inverses (i.e., the response function h).

Link function	$\eta_i = g(\mu_i)$
Identity	μ_i
Log	$\log(\mu_i)$
Inverse	μ_i^{-1}
Inverse-square	μ_i^{-2}
Square-root	$\sqrt{\mu_i}$
Logit	$\log \frac{\mu_i}{1-\mu_i}$
Probit	$\Phi^{-1}(\mu_i)$
Log-log	$-\log(-\log \mu_i)$
Complementary log-log	$\log(-\log(1-\mu_i))$

3. Write down the GLM components and the corresponding natural link when the response variable Y is assumed to follow each distribution below. Note that the values of y_i have to be in the correct range for each distribution.

- Gaussian: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim \mathcal{N}(\mu_i, \sigma^2)$.

- Bernoulli: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim Bernoulli(\mu_i)$.
- Binomial: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim Bin(n_i, \mu_i)$. Note that here y_i is the proportion of successes in n_i independent trials with probability of success μ_i. Thus, the range of y_i is $\{\frac{0}{n_i}, \frac{1}{n_i}, \ldots, \frac{n_i}{n_i}\}$.
- Poisson: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim Poi(\mu_i)$.
- Gamma: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim Gamma(\nu, \alpha_i)$, where ν and α_i are the shape and rate parameters of the Gamma distribution.
- Inverse-Gaussian: $y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim IG(\mu_i, \lambda)$.

From these results, show that the linear regression model, logistic regression model, and Poisson regression model are all special cases of GLMs.

4. In `R`, find the default link function for each GLM family. Are they the natural link for the corresponding family?

Chapter 2

Estimation

Suppose we are given a dataset that we would like to model using a GLM. After checking the data and possibly performing some exploratory data analysis, we have already chosen a specific form of GLMs that is most suitable for the dataset (say e.g., the Poisson GLM). Now we wish to estimate the parameters of the model; that is, finding the value of $\boldsymbol{\beta}$ that best explains our data. For this purpose, we can use the maximum likelihood method to obtain an estimate $\hat{\boldsymbol{\beta}}$.

One advantage of the GLM formulation is that it allows us to derive a solution for the maximum likelihood method directly in the general form, thus unifying the estimation methods for various statistical models. In other words, we do not need to derive solutions for the Poisson GLM, the Binomial GLM, etc. separately. Instead, we can just derive the solution for the general form of the GLMs. The solution for each specific form of the GLM will be an instance of this general solution.

For the rest of this chapter,

2.1 Likelihood Function

Consider the grouped data setup where we have predictors and data with possible replicates $\{(\boldsymbol{x}_i, y_{ir_i})\}_{i\in[1..n], r_i\in[1..m_i]}$. Recall that under a GLM, given predictors $\{\boldsymbol{x}_i\}_{i\in[1..n]}$, each response y_{ir_i} is independent of the other y_{jr_j}, and of the values of all predictors $\boldsymbol{x}_j$ with $j \neq i$, so that the joint probability of the data — that is, the likelihood — is given by

$$L(\boldsymbol{\beta}) = P(\{y_{ir_i}\} \,|\, \{\boldsymbol{x}_i\}, \boldsymbol{\beta}) = P(\{y_{ir_i}\} \,|\, \{\theta_i\}, \phi) = \prod_{i=1}^{n} \prod_{r_i=1}^{m_i} P(y_{ir_i}|\theta_i, \phi) \tag{2.1}$$

where

$$P(y_{ir_i}|\theta_i, \phi) = \exp\left(\frac{y_{ir_i}\theta_i - b(\theta_i)}{\phi} + c(y_{ir_i}, \phi)\right) \tag{2.2}$$

with

$$\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(h(\eta_i)) = (b')^{-1}(h(\boldsymbol{\beta}^T \boldsymbol{x}_i)). \tag{2.3}$$

2.2 Log-Likelihood Function

The log probability of the data — or log-likelihood — is thus given by

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log L(\boldsymbol{\beta}) = \log P(\{y_{ir_i}\} \mid \{\theta_i\}, \phi) && (2.4)\\ &= \sum_i \sum_{r_i} \left(\frac{y_{ir_i}\theta_i - b(\theta_i)}{\phi} + c(y_{ir_i}, \phi) \right) && (2.5)\\ &= \sum_i \left(m_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + \sum_{r_i} c(y_{ir_i}, \phi) \right) && (2.6)\\ &= \sum_i l_i && (2.7) \end{aligned}$$

where we have defined

$$\begin{aligned} y_i &= \frac{1}{m_i} \sum_{r_i} y_{ir_i} && (2.8)\\ l_i &= \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{r_i} c(y_{ir_i}, \phi) && (2.9)\\ \phi_i &= \phi / m_i. && (2.10) \end{aligned}$$

2.3 Score Function and Score Equation

The *score function* is given by

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}^T} = \sum_i \frac{\partial l_i}{\partial \boldsymbol{\beta}^T} = \sum_i \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T} \tag{2.11}$$

where, recalling that $\mu_i = b'(\theta_i)$, $\mathcal{V}(\mu_i) = b''(\theta_i)$, $\mu_i = h(\eta_i)$ and $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$, we have:[1]

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\phi_i} = \frac{y_i - \mu_i}{\phi_i} && (2.12)\\ \frac{\partial \theta_i}{\partial \mu_i} &= 1 / \left(\frac{\partial \mu_i}{\partial \theta_i} \right) = \frac{1}{b''(\theta_i)} = \frac{1}{\mathcal{V}(\mu_i)} && (2.13)\\ \frac{\partial \mu_i}{\partial \eta_i} &= h'(\eta_i) && (2.14)\\ \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T} &= \boldsymbol{x}_i. && (2.15) \end{aligned}$$

[1] Note that here and in this module, we assume that ϕ does not depend on $\boldsymbol{\beta}$.

The score function is thus given by

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i \left(\frac{y_i - \mu_i}{\phi_i}\right) \left(\frac{1}{\mathcal{V}(\mu_i)}\right) h'(\eta_i)\, \boldsymbol{x}_i \tag{2.16}$$

$$= \frac{1}{\phi} \sum_i m_i (y_i - \mu_i) \frac{1}{\mathcal{V}(\mu_i)} h'(\eta_i)\, \boldsymbol{x}_i. \tag{2.17}$$

The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ must then satisfy the *score equation*:

$$\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0. \tag{2.18}$$

Note that the dispersion parameter ϕ cancels from the score equation, which implies that $\hat{\boldsymbol{\beta}}$ does not depend on ϕ. This is another important property of EDFs.

2.3.1 Special Case: Natural Link

For the natural link, $\theta_i = \eta_i$, so Equations (2.13) and (2.14) combine to give

$$\frac{h'(\eta_i)}{\mathcal{V}(\mu_i)} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \theta_i}{\partial \eta_i} = 1. \tag{2.19}$$

The score function thus simplifies to

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_i m_i (y_i - \mu_i)\, \boldsymbol{x}_i. \tag{2.20}$$

2.4 Fisher Information

To solve the score equation, we will also need the second derivative of the log-likelihood. Its negative is called the *Observed Fisher Information*, defined as

$$\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = -\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\beta}}. \tag{2.21}$$

Note that, at the MLE, $\boldsymbol{F}_{\text{obs}}(\hat{\boldsymbol{\beta}})$ is positive by definition. Because it is a function of the data $\{y_i\}$, $\boldsymbol{F}_{\text{obs}}$ has a probability distribution. In practice, the Observed Fisher Information is often approximated by the *Expected Fisher Information*, otherwise known simply as the *Fisher Information*:[2]

$$\boldsymbol{F}(\boldsymbol{\beta}) = E\left[-\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\beta}}\right] \tag{2.22}$$

[2]Some texts refer to $\boldsymbol{F}_{\text{obs}}(\hat{\boldsymbol{\beta}})$ as the Observed Fisher Information, and to $\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})$ simply as the Fisher Information. Some don't refer to either of these at all. Just to be clear, we will refer to $\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})$ as the Observed Fisher Information and the Expected Fisher Information $\boldsymbol{F}(\boldsymbol{\beta})$ simply as the Fisher Information.

where the expectation is taken over the joint probability distribution of the data $P(\{y_{ir_i}\}|\boldsymbol{\beta}, \{\boldsymbol{x}_i\})$.

2.5 Example: Poisson Regression

We look at two example calculations of the score function and Fisher Information for Poisson Regression, that is we have

- $y|\boldsymbol{x}, \boldsymbol{\beta} \sim \text{Poi}(\lambda(\boldsymbol{x}, \boldsymbol{\beta}))$
- $\phi = 1$.

In this example, let us assume $m_i = 1$, i.e., the data are ungrouped.

2.5.1 With Natural Link

We have that: $\lambda(\boldsymbol{x}, \boldsymbol{\beta}) = \mu(\boldsymbol{x}, \boldsymbol{\beta}) = h(\eta(\boldsymbol{x}, \boldsymbol{\beta})) = e^{\eta(\boldsymbol{x}, \boldsymbol{\beta})} = e^{\boldsymbol{\beta}^T \boldsymbol{x}}$.

Equation (2.20) then gives

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i (y_i - e^{\boldsymbol{\beta}^T \boldsymbol{x}_i})\, \boldsymbol{x}_i \tag{2.23}$$

while Equation (2.21) gives

$$\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = \sum_i e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\, \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{2.24}$$

Note that this does not depend on the data, so that Equation (2.22) gives

$$\boldsymbol{F}(\boldsymbol{\beta}) = \text{E}[\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})] = \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}). \tag{2.25}$$

2.5.2 With Identity Link

The identity link is defined such that $h(\eta) = \eta$.

In this case, we have that:

- $\lambda(\boldsymbol{x}, \boldsymbol{\beta}) = \mu(\boldsymbol{x}, \boldsymbol{\beta}) = h(\eta(\boldsymbol{x}, \boldsymbol{\beta})) = \eta(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{x}$.
- $\mathcal{V}(\mu) = \mu$ (see Poisson example in EDF chapter).
- $h'(\eta) = 1$.

Equation (2.17) thus gives

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i (y_i - \mu_i)\, \frac{1}{\mu_i}\, 1\, \boldsymbol{x}_i \tag{2.26}$$

$$= \sum_i (y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)\, \frac{1}{\boldsymbol{\beta}^T \boldsymbol{x}_i}\, \boldsymbol{x}_i \tag{2.27}$$

$$= \sum_i \left(\frac{y_i}{\boldsymbol{\beta}^T \boldsymbol{x}_i} - 1 \right) \boldsymbol{x}_i. \tag{2.28}$$

Using Equation (2.21) and the chain rule, we can obtain the Observed Fisher Information:

$$\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = \sum_i \frac{y_i}{(\boldsymbol{\beta}^T \boldsymbol{x}_i)^2}\, \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{2.29}$$

Hence, the Fisher Information is:

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}[\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})] \tag{2.30}$$

$$= \mathrm{E}\left[\sum_i \frac{Y_i}{(\boldsymbol{\beta}^T \boldsymbol{x}_i)^2}\, \boldsymbol{x}_i \boldsymbol{x}_i^T \right] \tag{2.31}$$

$$= \sum_i \frac{\mathrm{E}[Y_i | \boldsymbol{\beta}, \boldsymbol{x}_i]}{(\boldsymbol{\beta}^T \boldsymbol{x}_i)^2}\, \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{2.32}$$

$$= \sum_i \frac{\boldsymbol{\beta}^T \boldsymbol{x}_i}{(\boldsymbol{\beta}^T \boldsymbol{x}_i)^2}\, \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{2.33}$$

$$= \sum_i \frac{1}{\boldsymbol{\beta}^T \boldsymbol{x}_i}\, \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{2.34}$$

Note that $\boldsymbol{F}(\boldsymbol{\beta}) \neq \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})$ in this case.

2.6 Properties of $\boldsymbol{S}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$

Having defined the score function $\boldsymbol{S}(\boldsymbol{\beta})$ and the Fisher information $\boldsymbol{F}(\boldsymbol{\beta})$, in this section we will investigate some of their properties. First, let us define $S_i(\boldsymbol{\beta}) = \dfrac{\partial l_i}{\partial \boldsymbol{\beta}}$. Then we have $\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i \boldsymbol{S}_i(\boldsymbol{\beta})$.

2.6.1 Expectation of $\boldsymbol{S}(\boldsymbol{\beta})$

The expectation of $\boldsymbol{S}(\boldsymbol{\beta})$ can be computed from Equations (2.11) and (2.17) as follows:

$$\mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})] = \sum_i \mathrm{E}[\boldsymbol{S}_i(\boldsymbol{\beta})] \tag{2.35}$$

$$= \sum_i \frac{\mathrm{E}[Y_i | \boldsymbol{\beta}, \boldsymbol{x}_i] - \mu_i}{\phi_i} \frac{1}{\mathcal{V}(\mu_i)} h'(\eta_i)\, \boldsymbol{x}_i \tag{2.36}$$

$$= 0 \tag{2.37}$$

because $\mathrm{E}[Y_i|\boldsymbol{\beta}, \boldsymbol{x}_i] = \mu_i$.

2.6.2 Variance of $\boldsymbol{S}(\boldsymbol{\beta})$

Using Equation (2.17) and the properties of covariance matrices in Section 1.2, we can calculate the variance of $\boldsymbol{S}(\boldsymbol{\beta})$ as follows:

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})] &= \sum_i \mathrm{Var}[\boldsymbol{S}_i(\boldsymbol{\beta})] && (2.38)\\
&= \sum_i \mathrm{Var}\left[\frac{h'(\eta_i)}{\phi_i \mathcal{V}(\mu_i)}\boldsymbol{x}_i(Y_i - \mu_i)\right] && (2.39)\\
&= \sum_i \left(\frac{h'(\eta_i)}{\phi_i \mathcal{V}(\mu_i)}\boldsymbol{x}_i\right) \mathrm{Var}[Y_i - \mu_i] \left(\frac{h'(\eta_i)}{\phi_i \mathcal{V}(\mu_i)}\boldsymbol{x}_i^T\right) && (2.40)\\
&= \sum_i \left(\frac{h'(\eta_i)^2}{\phi_i^2 \mathcal{V}(\mu_i)^2}\boldsymbol{x}_i\boldsymbol{x}_i^T\right) \mathrm{Var}[Y_i]. && (2.41)
\end{aligned}$$

Note that in the first equality above, we can break the variance into sum of smaller components due to the independent data assumption.

Now using $\mathrm{Var}[Y_i] = \phi_i \mathcal{V}(\mu_i)$, we can obtain the expression for $\mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})]$:

$$\mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})] = \sum_i \frac{h'(\eta_i)^2}{\phi_i \mathcal{V}(\mu_i)}\boldsymbol{x}_i\boldsymbol{x}_i^T. \tag{2.42}$$

2.6.3 Properties of $\boldsymbol{F}(\boldsymbol{\beta})$

Recall from Equation (2.22) that $\boldsymbol{F}(\boldsymbol{\beta}) = E\left[-\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\beta}}\right] = -E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right]$.

We will first show that $E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right] = E\left[-\frac{\partial l}{\partial \boldsymbol{\beta}^T}\frac{\partial l}{\partial \boldsymbol{\beta}}\right]$.

2.6.3.1 An Important Identity

Let $\rho = e^l$, where l is the log-likelihood, so that $\rho = L(\boldsymbol{\beta}) = P(\{y_{ir_i}\} \,|\, \{\boldsymbol{x}_i\}, \boldsymbol{\beta})$ is the likelihood/probability of the data. Then

$$\frac{\partial l}{\partial \boldsymbol{\beta}^T} = \frac{\partial l}{\partial \rho}\frac{\partial \rho}{\partial \boldsymbol{\beta}^T} = \frac{1}{\rho}\frac{\partial \rho}{\partial \boldsymbol{\beta}^T}. \tag{2.43}$$

Using the product rule and chain rule, we have

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\frac{1}{\rho^2}\frac{\partial \rho}{\partial \boldsymbol{\beta}^T}\frac{\partial \rho}{\partial \boldsymbol{\beta}} + \frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \tag{2.44}$$

$$= -\frac{1}{\rho^2}\left(\frac{\partial \rho}{\partial l}\frac{\partial l}{\partial \boldsymbol{\beta}^T}\right)\left(\frac{\partial \rho}{\partial l}\frac{\partial l}{\partial \boldsymbol{\beta}}\right) + \frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \tag{2.45}$$

$$= -\frac{1}{\rho^2}\left(\rho\frac{\partial l}{\partial \boldsymbol{\beta}^T}\right)\left(\rho\frac{\partial l}{\partial \boldsymbol{\beta}}\right) + \frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \tag{2.46}$$

$$= -\frac{\partial l}{\partial \boldsymbol{\beta}^T}\frac{\partial l}{\partial \boldsymbol{\beta}} + \frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}. \tag{2.47}$$

Note that the expectation (over the data) of the second term is

$$\mathrm{E}\left[\frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right] = \int \rho\, \frac{1}{\rho}\frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = \int \frac{\partial^2 \rho}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = \frac{\partial^2}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \int \rho = \frac{\partial^2}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\, 1 = 0. \tag{2.48}$$

Thus, we have:

$$E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right] = E\left[-\frac{\partial l}{\partial \boldsymbol{\beta}^T}\frac{\partial l}{\partial \boldsymbol{\beta}}\right]. \tag{2.49}$$

2.6.3.2 Relating $\boldsymbol{F}(\boldsymbol{\beta})$ and $\mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})]$

Using Equation (2.49), we have that:

$$\boldsymbol{F}(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\beta}^T}\frac{\partial l}{\partial \boldsymbol{\beta}}\right] \tag{2.50}$$

$$= \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T] \tag{2.51}$$

$$= \mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})] + \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})]\ \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})]^T \tag{2.52}$$

$$= \mathrm{Var}[\boldsymbol{S}(\boldsymbol{\beta})] \tag{2.53}$$

where the last equality is due to $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})] = 0$.

Therefore, an important property of the Fisher Information is that it is equal to the variance of the score function, whose expression is given in Equation (2.42).

2.6.3.3 Special Case: Natural Link

For the natural link, recall that $\dfrac{h'(\eta_i)}{\mathcal{V}(\mu_i)} = 1$. So we have:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i \frac{1}{\phi_i}(y_i - h(\eta_i))\boldsymbol{x}_i. \tag{2.54}$$

Let $\boldsymbol{S}_i = \dfrac{1}{\phi_i}(y_i - h(\eta_i))\boldsymbol{x}_i$. We have $\boldsymbol{S} = \sum_i \boldsymbol{S}_i$ and thus:

$$\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = -\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\beta}} = -\sum_i \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{\beta}} = -\sum_i \frac{\partial \boldsymbol{S}_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \sum_i \frac{h'(\eta_i)}{\phi_i}\boldsymbol{x}_i\boldsymbol{x}_i^T \tag{2.55}$$

$$\boldsymbol{F}(\boldsymbol{\beta}) = \text{Var}[\boldsymbol{S}(\boldsymbol{\beta})] = \sum_i \frac{h'(\eta_i)}{\phi_i}\boldsymbol{x}_i\boldsymbol{x}_i^T. \tag{2.56}$$

Thus, for the natural link, we see that $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta})$.

2.7 Matrix Notation

For the next section, it is useful to establish a condensed, matrix notation for some of the previous quantities, analogous to the matrix notation used for linear models.

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be the random vector with components Y_i, the response values. This is exactly the same quantity as in the linear model case.
- Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the *design matrix*, the matrix with components $x_{i,a}$, the value of the a^{th} component of the predictor vector for the i^{th} data point. This is exactly the same quantity as in the linear model case. This matrix is sometimes called the *model matrix*.
- Let $\boldsymbol{\mu} \in \mathbb{R}^n$ be the vector with components $\mu_i = h(\boldsymbol{\beta}^T x_i)$, so that $\boldsymbol{\mu} = \text{E}[\boldsymbol{Y}]$.
- Let $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be the diagonal matrix with components $D_{ii} = h'(\eta_i)$. For example, if $h(\eta) = e^\eta$, then

$$\boldsymbol{D} = \begin{pmatrix} e^{\boldsymbol{\beta}^T x_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{\boldsymbol{\beta}^T x_n} \end{pmatrix}. \tag{2.57}$$

- Let $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ be the covariance matrix for $\boldsymbol{Y}$, with components:

$$\Sigma_{ij} = \text{Cov}[Y_i, Y_j] = \text{Var}[Y_i]\ \delta_{ij} = \phi_i \mathcal{V}(\mu_i)\ \delta_{ij}. \tag{2.58}$$

That is,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}[Y_1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \text{Var}[Y_n] \end{pmatrix} = \begin{pmatrix} \phi_1\mathcal{V}(\mu_1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \phi_n\mathcal{V}(\mu_n) \end{pmatrix}. \tag{2.59}$$

2.7.1 Score Function and Fisher Information

Recall that

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i \left(\frac{y_i - \mu_i}{\phi_i \mathcal{V}(\mu_i)} \right) h'(\eta_i)\, \boldsymbol{x}_i \tag{2.60}$$

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_i \frac{h'(\eta_i)^2}{\phi_i \mathcal{V}(\mu_i)}\, \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{2.61}$$

In terms of the matrix notation, these become

$$\boldsymbol{S} = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \tag{2.62}$$

$$\boldsymbol{F} = \boldsymbol{X}^T \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D} \boldsymbol{X}. \tag{2.63}$$

2.7.2 Natural Link

Note that for the natural link,

$$\frac{\partial \theta_i}{\partial \eta_i} = \frac{h'(\eta_i)}{\mathcal{V}(\mu_i)} = 1. \tag{2.64}$$

Thus, with $\phi_i = \phi / m_i$, we have:

$$h'(\eta_i) = \mathcal{V}(\mu_i) = \frac{\mathrm{Var}[Y_i]}{\phi_i} = m_i \frac{\mathrm{Var}[Y_i]}{\phi}. \tag{2.65}$$

Now let $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ be the diagonal matrix with components $m_i \delta_{ij}$, known as the *grouping* matrix. Then

$$\boldsymbol{D} = \frac{1}{\phi} \boldsymbol{G} \boldsymbol{\Sigma} = \frac{1}{\phi} \boldsymbol{\Sigma} \boldsymbol{G}, \tag{2.66}$$

and therefore,

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{G} (\boldsymbol{Y} - \boldsymbol{\mu}) \tag{2.67}$$

$$\boldsymbol{F}(\boldsymbol{\beta}) = \frac{1}{\phi^2} \boldsymbol{X}^T \boldsymbol{G}^T \boldsymbol{\Sigma} \boldsymbol{G} \boldsymbol{X}. \tag{2.68}$$

2.8 Iterative Solution of $\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0$

So far we have seen how to set up the score equation for the maximum likelihood estimate. We have also proven some of its properties as well as those of the Fisher Information. We now turn to the question of how to solve the score equation. In general, this cannot be done in closed form, except in rare cases. So we need to turn to numerical methods implemented on a computer.

We have the same two options here as in the binary regression case: we can try to optimise l directly, or we can attempt to solve the score equation. There are many algorithms that can be used to perform these tasks. Here we focus on one: *Iteratively Reweighted Least Squares (IRLS)*, also known as *Iterative Weighted Least Squares (IWLS)*.[3]

We start by recalling the Newton-Raphson method for finding the zero of a function. Note that we wish to solve the equation:

$$\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0. \tag{2.69}$$

We then approximate $\boldsymbol{S}$ linearly about some point $\boldsymbol{\beta}_0$:

$$\boldsymbol{S}(\boldsymbol{\beta}_0 + \delta\boldsymbol{\beta}_0) = \boldsymbol{S}(\boldsymbol{\beta}_0) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\delta\boldsymbol{\beta}_0 + \mathcal{O}(\delta\boldsymbol{\beta}_0^2) \tag{2.70}$$

where the reason for the subscript 0 will become apparent soon. In the case when $\boldsymbol{S}(\boldsymbol{\beta}_0+\delta\boldsymbol{\beta}_0) = 0$ (such as the case we are interested in), we have approximately that:

$$\frac{\partial \boldsymbol{S}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\delta\boldsymbol{\beta}_0 = -\boldsymbol{S}(\boldsymbol{\beta}_0). \tag{2.71}$$

Now in our case,

$$-\frac{\partial \boldsymbol{S}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}_0), \tag{2.72}$$

so Equation (2.71) becomes

$$\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}_0)\delta\boldsymbol{\beta}_0 = \boldsymbol{S}(\boldsymbol{\beta}_0), \tag{2.73}$$

or equivalently,

$$\delta\boldsymbol{\beta}_0 = \left(\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}_0)\right)^{-1}\boldsymbol{S}(\boldsymbol{\beta}_0). \tag{2.74}$$

This then gives us a new value:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \delta\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 + \left(\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}_0)\right)^{-1}\boldsymbol{S}(\boldsymbol{\beta}_0). \tag{2.75}$$

Then we can iterate the steps above for $m = 1, 2, \ldots,$

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m + \delta\boldsymbol{\beta}_m \tag{2.76}$$

[3]You have cause to be particularly interested in this algorithm as a Durham student. It is on the undergraduate syllabus of nearly every maths degree in the world which includes a large statistical component and some of the important early development was researched by Dr Peter Green when he was a lecturer at Durham: Green [1984].

where

$$\delta\boldsymbol{\beta}_m = \left(\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}_m)\right)^{-1} \boldsymbol{S}(\boldsymbol{\beta}_m). \tag{2.77}$$

Because $\boldsymbol{F}_{\text{obs}}$ is hard to find and hard to invert in general, we approximate it with the expected Fisher Information. This is known as the *Fisher scoring* method, where we compute $\delta\boldsymbol{\beta}_m$ by:

$$\delta\boldsymbol{\beta}_m = \left(\boldsymbol{F}(\boldsymbol{\beta}_m)\right)^{-1} \boldsymbol{S}(\boldsymbol{\beta}_m). \tag{2.78}$$

2.8.1 Iteratively Reweighted Least Squares (IRLS)

Now we will use Equation (2.78) to derive the IRLS method in matrix notation. From Equation (2.78), we have that

$$\boldsymbol{F}(\boldsymbol{\beta}_m)\delta\boldsymbol{\beta}_m = \boldsymbol{S}(\boldsymbol{\beta}_m) \tag{2.79}$$

or equivalently that

$$\boldsymbol{F}(\boldsymbol{\beta}_m)\boldsymbol{\beta}_{m+1} = \boldsymbol{F}(\boldsymbol{\beta}_m)\boldsymbol{\beta}_m + \boldsymbol{S}(\boldsymbol{\beta}_m). \tag{2.80}$$

Using the matrix notation in Section 2.7.1 and defining $\boldsymbol{W} = \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}$, we can write

$$\boldsymbol{F} = \boldsymbol{X}^T\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \tag{2.81}$$

and

$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{X}^T\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}), \tag{2.82}$$

since $\boldsymbol{D}$ is a diagonal matrix.

Thus, using the subscript m to denote the value of a quantity evaluated using $\boldsymbol{\beta}_m$ or derived quantities, we can calculate the right hand side of Equation (2.80) from:

$$\boldsymbol{F}_m\boldsymbol{\beta}_m + \boldsymbol{S}_m = \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X}\boldsymbol{\beta}_m + \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{D}_m^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_m) = \boldsymbol{X}^T\boldsymbol{W}_m\tilde{\boldsymbol{Y}}_m \tag{2.83}$$

where

$$\tilde{\boldsymbol{Y}}_m = \boldsymbol{X}\boldsymbol{\beta}_m + \boldsymbol{D}_m^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_m) \tag{2.84}$$

are the so-called *working observations*.

Now replacing $\boldsymbol{F}(\boldsymbol{\beta}_m) = \boldsymbol{F}_m = \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X}$ in the left hand side of Equation (2.80), we have that

$$(\boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X})\boldsymbol{\beta}_{m+1} = \boldsymbol{X}^T\boldsymbol{W}_m\tilde{\boldsymbol{Y}}_m \tag{2.85}$$

or

$$\boldsymbol{\beta}_{m+1} = (\boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_m\tilde{\boldsymbol{Y}}_m. \tag{2.86}$$

Thus, to find a solution for $\boldsymbol{S}(\boldsymbol{\beta}) = 0$, we can start from an arbitrary point $\boldsymbol{\beta}_0$ and iteratively apply Equation (2.86) until a convergence criterion is met.

This sequence of iterated operations is called *iteratively reweighted least squares* or *iterative weighted least squares* since each iteration is the solution to the following least squares problem: minimize the quantity $l_m(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where

$$l_m(\boldsymbol{\beta}) = (\tilde{\boldsymbol{Y}}_m - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}_m(\tilde{\boldsymbol{Y}}_m - \boldsymbol{X}\boldsymbol{\beta}) \tag{2.87}$$

and $\boldsymbol{W}_m$ is known as the *weight matrix*.

2.8.2 IRLS Pseudo-Code

We now give a pseudo-code below for running IRLS. Note that the code will not run without computing $\boldsymbol{\mu}$, $\boldsymbol{D}$ and $\boldsymbol{W}$ using a specific example.

```
IRLS <- function(Y, X, phi, epsilon) {
    # Pick an initial value for hatBeta
    hatbeta = initializeBeta()

    # Set up convergence
    converged = false

    # Loop as long as convergence condition is not satisfied
    while not converged loop
    {
        # Compute mu, D, and Sigma (use h, h', V as subroutines)
        mu = computeMu(hatBeta, X)
        D = computeD(hatBeta, X)
        Sigma = computeSigma(hatBeta, phi)

        # Compute the weight matrix W
        W = t(D) %*% solve(Sigma) %*% D

        # Compute the working observations tildeY
        tildeY = X %*% hatBeta + solve(D) %*% (Y - mu)

        # Compute the new value of hatBeta
        newHatBeta = solve(t(X) %*% W %*% X) %*% (t(X) %*% W %*% tildeY)
```

```
        # Check whether we have converged
        converged = ((norm(newHatBeta - hatBeta) / norm(hatBeta)) <= epsilon)

        # Store new value of hatBeta ready for next iteration or return
        hatBeta = newHatBeta
    }

    return hatBeta
}
```

2.9 Practical Example: US Polio Data

In this example, we will use the US Polio data discussed last term to fit the GLMs. Recall that this dataset is a matrix of count data, giving the monthly number of polio cases in the United States from 1970 to 1983. We will convert this dataset into a matrix with two columns:

- covariate `time` in the first column ranging from 1 to 168, starting with January 1970.
- response `cases` in the second column, indicating the monthly number of polio cases.

We now load the data from the library `gamlss.data` and do the conversion.

```
library("gamlss.data")
```

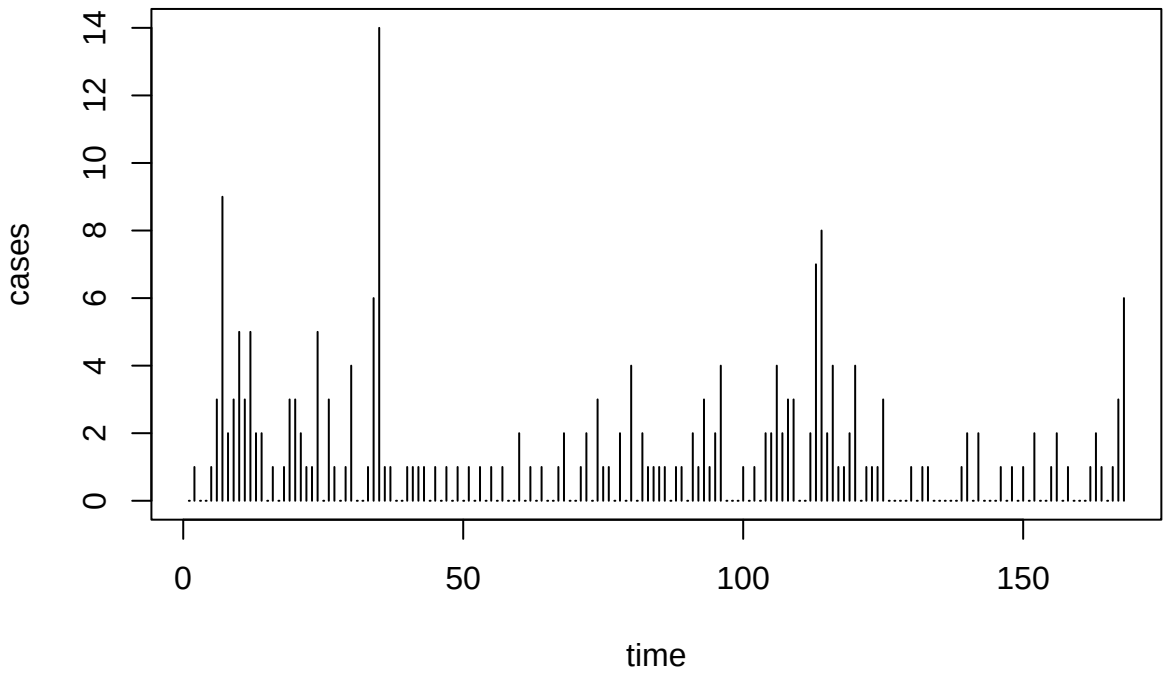
```
##
## Attaching package: 'gamlss.data'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
data("polio")
uspolio <- as.data.frame(matrix(c(1:168, t(polio)), ncol=2))
colnames(uspolio) <- c("time", "cases")
```

First, let us plot the data.

```
plot(uspolio, type="h")
```



Note that the main question we wish to consider is: *How has Polio incidence changed over time?*

Since this is count data, we begin by fitting a Poisson model with a linear time trend.

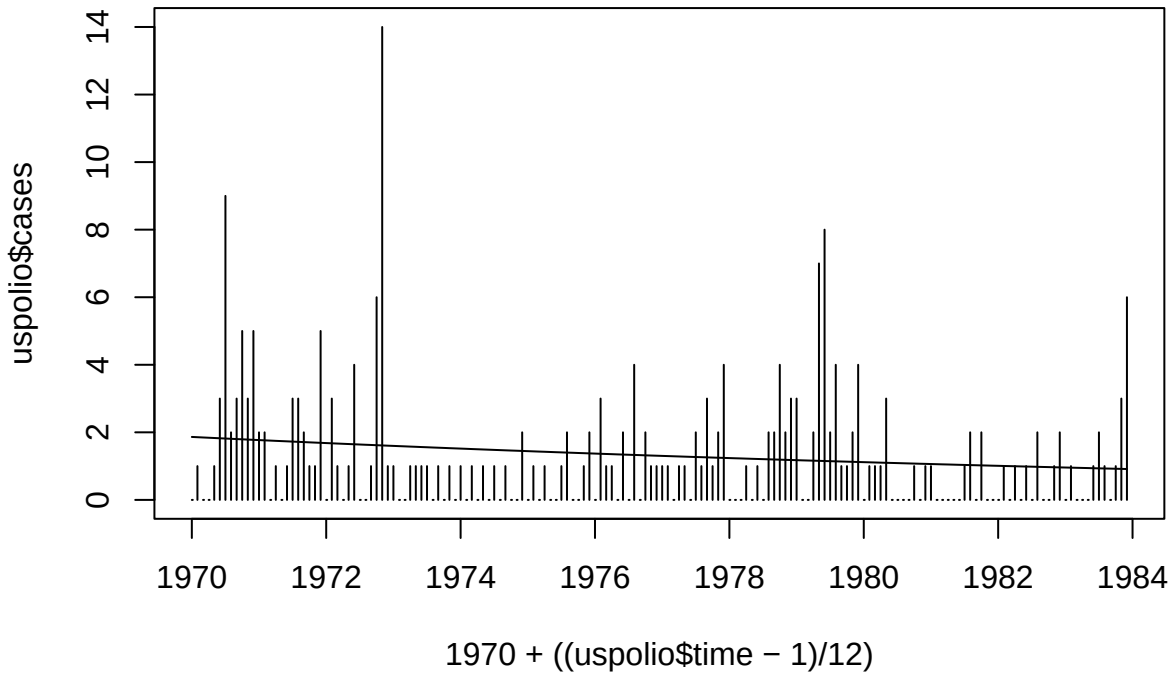
```
# Poisson model with linear time trend
polio.glm <- glm(cases ~ time, family=poisson(link=log), data=uspolio)

# Look at the model summary
summary(polio.glm)
```

```
##
## Call:
## glm(formula = cases ~ time, family = poisson(link = log), data = uspolio)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.626639   0.123641   5.068 4.02e-07 ***
## time        -0.004263   0.001395  -3.055  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 343.00  on 167  degrees of freedom
## Residual deviance: 333.55  on 166  degrees of freedom
## AIC: 594.59
##
## Number of Fisher Scoring iterations: 5
```

We can then plot the model as follows.

```
plot(1970 + ((uspolio$time - 1)/12), uspolio$cases, type="h")
lines(1970 + ((uspolio$time - 1)/12), polio.glm$fitted)
```



We can see that this is perhaps unsatisfactory. To improve the model, we can explore a linear trend with seasonal (annual) components.

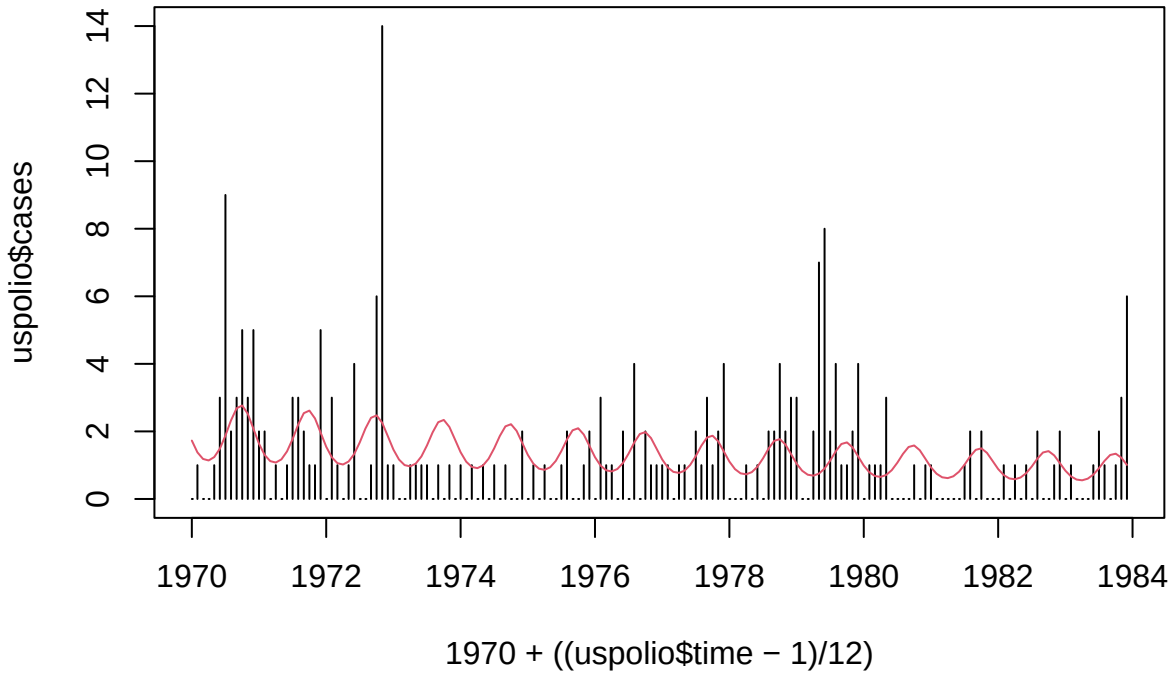
```
# Poisson model with linear trend and annual components
polio1.glm <- glm(cases ~ time + I(cos(2*pi*time/12)) + I(sin(2*pi*time/12)),
family=poisson(link=log), data=uspolio)

summary(polio1.glm)
```

```
##
## Call:
## glm(formula = cases ~ time + I(cos(2 * pi * time/12)) + I(sin(2 *
##     pi * time/12)), family = poisson(link = log), data = uspolio)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.606612   0.124800   4.861 1.17e-06 ***
## time                      -0.004644   0.001401  -3.315 0.000916 ***
## I(cos(2 * pi * time/12))   0.181254   0.096160   1.885 0.059442 .
## I(sin(2 * pi * time/12)) -0.423187   0.097590  -4.336 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 343.00  on 167  degrees of freedom
```

```
## Residual deviance: 310.72  on 164  degrees of freedom
## AIC: 575.77
##
## Number of Fisher Scoring iterations: 5
```

```
plot(1970 + ((uspolio$time - 1)/12), uspolio$cases, type="h")
lines(1970 + ((uspolio$time - 1)/12), polio1.glm$fitted, col=2)
```



We can also add six-monthly components into the model.

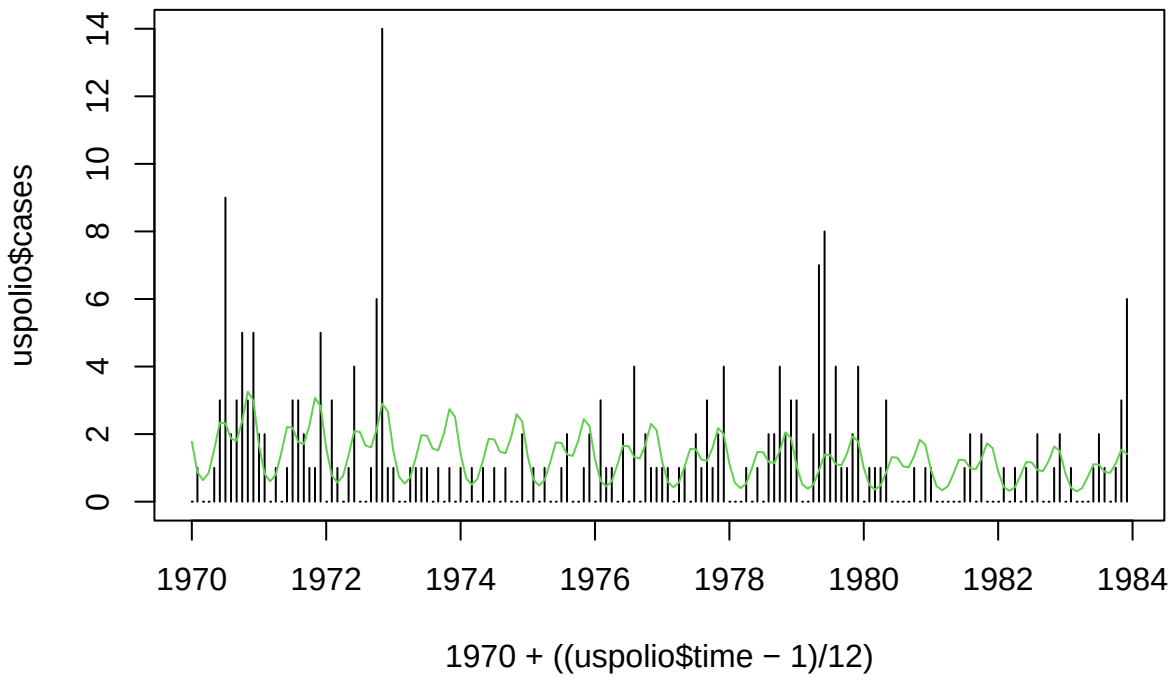
```
# Poisson model with linear trend and seasonal (annual + six-monthly) components
polio2.glm <- glm(cases ~ time + I(cos(2*pi*time/12)) + I(sin(2*pi*time/12))
+ I(cos(2*pi*time/6)) + I(sin(2*pi*time/6)), family=poisson(link=log),
data=uspolio)

summary(polio2.glm)
```

```
##
## Call:
## glm(formula = cases ~ time + I(cos(2 * pi * time/12)) + I(sin(2 *
##     pi * time/12)) + I(cos(2 * pi * time/6)) + I(sin(2 * pi *
##     time/6)), family = poisson(link = log), data = uspolio)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.557241   0.127303   4.377 1.20e-05 ***
## time                       -0.004799   0.001403  -3.421 0.000625 ***
## I(cos(2 * pi * time/12))    0.137132   0.089479   1.533 0.125384
## I(sin(2 * pi * time/12)) -0.534985   0.115476  -4.633 3.61e-06 ***
## I(cos(2 * pi * time/6))     0.458797   0.101467   4.522 6.14e-06 ***
```

```
## I(sin(2 * pi * time/6))  -0.069627   0.098123  -0.710 0.477957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 343.00  on 167  degrees of freedom
## Residual deviance: 288.85  on 162  degrees of freedom
## AIC: 557.9
##
## Number of Fisher Scoring iterations: 5
```

```
plot(1970 + ((uspolio$time - 1)/12), uspolio$cases, type="h")
lines(1970 + ((uspolio$time - 1)/12), polio2.glm$fitted, col=3)
```

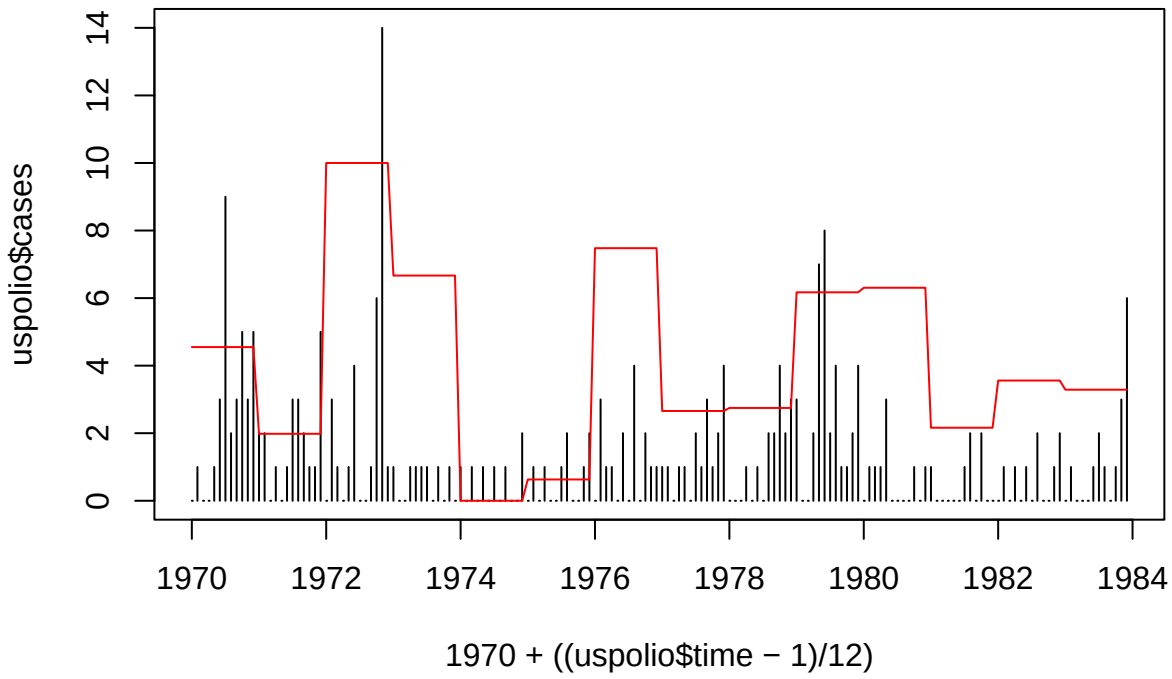


Assuming we have annual temperature data over the 14 years, we can add them into the model to investigate their effects.

```
# Average annual temperature data over the 14 years
temp_data <- rep(c(5.195, 5.138, 5.316, 5.242, 5.094, 5.108, 5.260, 5.153,
                   5.155, 5.231, 5.234, 5.142, 5.173, 5.167), each=12)

# Scale the data so that it plots nicely
scaled_temp = 10 * (temp_data - min(temp_data))/(max(temp_data) - min(temp_data))
uspolio$temp = scaled_temp

# Plot temperature data against cases data to see interest
plot(1970 + ((uspolio$time - 1)/12), uspolio$cases, type="h")
lines(1970 + ((uspolio$time - 1)/12), uspolio$temp, col="red")
```



Poisson GLM with temperature data.

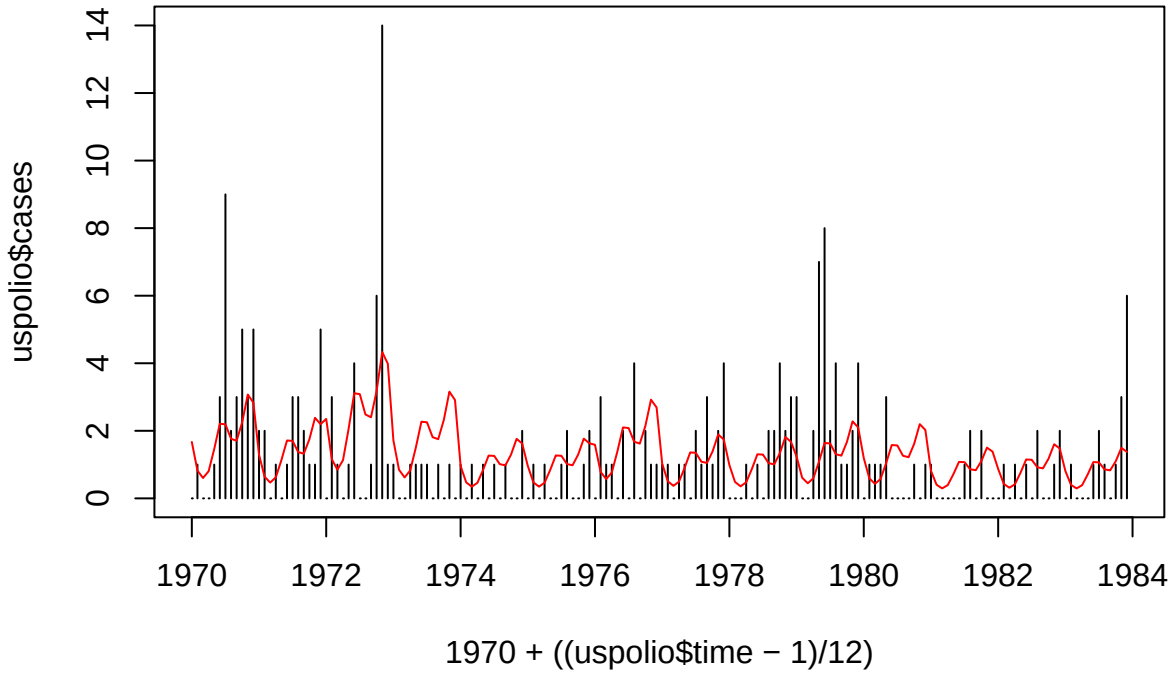
```
# Poisson model with additional temperature covariate
polio3.glm <- glm(cases ~ time + temp + I(cos(2*pi*time/12)) + I(sin(2*pi*time/12))
+ I(cos(2*pi*time/6)) + I(sin(2*pi*time/6)) , family=poisson(link=log),
data=uspolio)

summary(polio3.glm)
```

```
##
## Call:
## glm(formula = cases ~ time + temp + I(cos(2 * pi * time/12)) +
##     I(sin(2 * pi * time/12)) + I(cos(2 * pi * time/6)) + I(sin(2 *
##     pi * time/6)), family = poisson(link = log), data = uspolio)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.129643   0.186352   0.696 0.486623
## time                      -0.003972   0.001439  -2.761 0.005770 **
## temp                       0.080308   0.023139   3.471 0.000519 ***
## I(cos(2 * pi * time/12))   0.136094   0.089489   1.521 0.128314
## I(sin(2 * pi * time/12)) -0.531668   0.115466  -4.605 4.13e-06 ***
## I(cos(2 * pi * time/6))    0.457487   0.101435   4.510 6.48e-06 ***
## I(sin(2 * pi * time/6))   -0.068345   0.098149  -0.696 0.486218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 343.00  on 167  degrees of freedom
## Residual deviance: 276.84  on 161  degrees of freedom
## AIC: 547.88
##
## Number of Fisher Scoring iterations: 5
```

```
plot(1970 + ((uspolio$time - 1)/12), uspolio$cases, type="h")
lines(1970 + ((uspolio$time - 1)/12), polio3.glm$fitted, col="red")
```



2.10 Estimation of Dispersion Parameter ϕ

Because the dispersion ϕ cancels from the score equation $\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0$, there is no need to estimate ϕ in order to estimate $\boldsymbol{\beta}$. However, $\text{Var}[\hat{\boldsymbol{\beta}}]$ does depend on ϕ, as one might expect. Thus, if necessary or of interest, ϕ can be estimated via:

$$\hat{\phi} = \frac{1}{n-p} \sum_i m_i \frac{(y_i - \hat{\mu}_i)^2}{\mathcal{V}(\hat{\mu}_i)} \tag{2.88}$$

where p is the number of parameters of the model. The motivation for the above estimation is that:

$$\text{Var}[y_i] = \text{E}[(y_i - \mu_i)^2] = \phi_i \mathcal{V}(\mu_i) = \frac{\phi}{m_i} \mathcal{V}(\mu_i), \tag{2.89}$$

which can be rearranged to

$$\phi = \frac{m_i}{\mathcal{V}(\mu_i)} \text{E}[(y_i - \mu_i)^2] = \text{E}\left[m_i \frac{(y_i - \mu_i)^2}{\mathcal{V}(\mu_i)} \right]. \tag{2.90}$$

Thus, after estimating $\hat{\boldsymbol{\beta}}$, we can use its value and Equation (2.88) to estimate $\hat{\phi}$.

2.10.1 Special Cases

2.10.1.1 Gaussian

When $Y|\boldsymbol{\beta}, x \sim \mathcal{N}(\mu, \sigma^2)$ with $m_i = 1$, we have $\mathcal{V}(\mu_i) = 1$ and thus,

$$\hat{\phi} = \frac{1}{n-p}\sum_i (y_i - \hat{\mu}_i)^2 = \hat{\sigma}^2. \tag{2.91}$$

2.10.1.2 Gamma

Recall from Exercise 7.1 last term that we can parameterise the Gamma function in terms of its mean μ and variance σ^2, and we found that $\mathcal{V}(\mu) = \mu^2$. Thus, when $Y|\boldsymbol{\beta}, x \sim \text{Gamma}(\mu, \sigma^2)$, we have

$$\frac{1}{\hat{\nu}} = \hat{\phi} = \frac{1}{n-p}\sum_i m_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}. \tag{2.92}$$

2.10.2 Practical Example: Hospital Stay Data

In this example, we will use the Hospital Stay data introduced last term to fit a Gamma GLM and estimate its dispersion parameter.

```
library(npmlreg)
data(hosp)

# Fit the GLM and print the summary
hosp.glm <- glm(duration ~ age + temp1, data=hosp, family=Gamma(link=log))
summary(hosp.glm)
```

```
##
## Call:
## glm(formula = duration ~ age + temp1, family = Gamma(link = log),
##     data = hosp)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.654096  16.621018  -1.724   0.0987 .
## age           0.014900   0.005698   2.615   0.0158 *
## temp1         0.306624   0.168141   1.824   0.0818 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2690233)
##
##     Null deviance: 8.1722  on 24  degrees of freedom
```

```
## Residual deviance: 5.7849  on 22  degrees of freedom
## AIC: 142.73
##
## Number of Fisher Scoring iterations: 6
```

```
# From the summary, note the line:
# (Dispersion parameter for Gamma family taken to be 0.2690233)

# Compute by hand
1/(hosp.glm$df.res)*sum((hosp$duration-hosp.glm$fitted)^2/(hosp.glm$fitted^2))
```

```
## [1] 0.2690233
```

2.11 Asymptotic Properties of $\hat{\boldsymbol{\beta}}$

In our context, *asymptotic* means that $M = \sum_{i=1}^{n} m_i \to \infty$. This could be because $n \to \infty$, or because the $m_i \to \infty$, or a combination of both.

Let us denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^*$. In the following, we assume consistency of $\hat{\boldsymbol{\beta}}$, i.e., $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}^*$, meaning that $P(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| \geq \varepsilon) \to 0$ as $n \to \infty$. We will denote this by $\hat{\boldsymbol{\beta}} \stackrel{a}{=} \boldsymbol{\beta}^*$. We will also abuse this notation to mean "tends to asymptotically" for expectations, i.e., if we write $E[Z] \stackrel{a}{=} z$, that means $E[Z] \stackrel{n\to\infty}{\longrightarrow} z$.

From the consistency assumption, $\hat{\boldsymbol{\beta}}$ will be close to $\boldsymbol{\beta}^*$ asymptotically, and we can expand $\boldsymbol{S}$ around it:

$$\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0 \stackrel{a}{=} \boldsymbol{S}(\boldsymbol{\beta}^*) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \tag{2.93}$$

$$= \boldsymbol{S}(\boldsymbol{\beta}^*) - \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \tag{2.94}$$

or equivalently,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \stackrel{a}{=} \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{S}(\boldsymbol{\beta}^*). \tag{2.95}$$

2.11.1 Fisher Scoring

In Section 2.8, we stated that we often use the (expected) Fisher Information in place of the Observed Fisher Information (known as the Fisher Scoring method). Doing so in the context of asymptotic arguments is acceptable. We can roughly see this as follows. For any $\boldsymbol{\beta}$,

$$\frac{1}{n}\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = -\frac{1}{n}\frac{\partial l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}) \to -\text{E}\left[\frac{\partial l_1}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta})\right] = F_1(\boldsymbol{\beta}) \tag{2.96}$$

where $F_1(\boldsymbol{\beta})$ is the expected Fisher Information for a sample of size 1 and we are using the law of large numbers as $n \to \infty$ here. It can be shown (see exercise section) that $\boldsymbol{F}(\boldsymbol{\beta}) = nF_1(\boldsymbol{\beta})$, thus justifying use of $\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) \stackrel{a}{=} \boldsymbol{F}(\boldsymbol{\beta})$ in the forthcoming asymptotic arguments.

2.11.2 Expectation

From Equation (2.95), we have:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \stackrel{a}{=} \boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{S}(\boldsymbol{\beta}^*) \stackrel{a}{=} \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{S}(\boldsymbol{\beta}^*). \tag{2.97}$$

Because convergence in probability implies convergence in distribution, this in turn implies that

$$E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*] \stackrel{a}{=} \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}E[\boldsymbol{S}(\boldsymbol{\beta}^*)] = 0. \tag{2.98}$$

In other words, $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased.

2.11.3 Variance

Since $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*] \stackrel{a}{=} 0$, we have that

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*] &\stackrel{a}{=} \text{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T] && (2.99)\\
&\stackrel{a}{=} \text{E}[\boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{S}(\boldsymbol{\beta}^*)\boldsymbol{S}(\boldsymbol{\beta}^*)^T\boldsymbol{F}(\boldsymbol{\beta}^*)^{-T}] && (2.100)\\
&= \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\text{E}[\boldsymbol{S}(\boldsymbol{\beta}^*)\boldsymbol{S}(\boldsymbol{\beta}^*)^T]\boldsymbol{F}(\boldsymbol{\beta}^*)^{-T} && (2.101)\\
&= \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\text{Var}[\boldsymbol{S}(\boldsymbol{\beta}^*)]\boldsymbol{F}(\boldsymbol{\beta}^*)^{-T} && (2.102)\\
&= \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1} && (2.103)
\end{aligned}$$

where we have used symmetry of $\boldsymbol{F}$ and the fact that $\boldsymbol{F}(\boldsymbol{\beta}^*) = \text{Var}[\boldsymbol{S}(\boldsymbol{\beta}^*)]$.

Thus,

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \text{Var}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*] \stackrel{a}{=} \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}. \tag{2.104}$$

2.11.4 Asymptotic Normality

The following is a sketch of the argument of asymptotic normality for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, i.e., $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ converges asymptotically to a normal distribution. We start from

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_i \boldsymbol{S}_i(\boldsymbol{\beta}) \tag{2.105}$$

where $\boldsymbol{S}_i(\boldsymbol{\beta})$ is defined in Section 2.6. This is a sum of independent random variables with zero mean and finite variance. As the number of terms in the sum tends to infinity, then under a certain condition, the distribution of the sum converges in distribution to a normal distribution. Since $\text{E}[\boldsymbol{S}(\boldsymbol{\beta})] = 0$ and $\text{Var}[\boldsymbol{S}(\boldsymbol{\beta})] = \boldsymbol{F}(\boldsymbol{\beta})$, we have:

$$\boldsymbol{S}(\boldsymbol{\beta}) \stackrel{a}{\sim} \mathcal{N}(0, \boldsymbol{F}(\boldsymbol{\beta})). \tag{2.106}$$

Hence,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \stackrel{a}{=} \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{S}(\boldsymbol{\beta}^*) \stackrel{a}{\sim} \mathcal{N}(0, \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{F}(\boldsymbol{\beta}^*)\boldsymbol{F}(\boldsymbol{\beta}^*)^{-T}). \tag{2.107}$$

Since $\boldsymbol{F}$ is symmetric and convergence in probability implies convergence in distribution, we have:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \stackrel{a}{\sim} \mathcal{N}(0, \boldsymbol{F}(\boldsymbol{\beta}^*)^{-1}). \tag{2.108}$$

This also implies that the square of Mahalanobis distance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ is asymptotically chi-square distributed:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{F}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \stackrel{a}{\sim} \chi^2(p) \tag{2.109}$$

where p is the number of parameters.

2.11.5 Closing The Circle

At the beginning of this section, we assumed that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}^*$. We may want to justify that this assumption is reasonable. Note that under some regularity conditions,

$$\boldsymbol{F}(\boldsymbol{\beta})^{-1} = \left(\sum_i m_i \ldots\right)^{-1} \to 0 \tag{2.110}$$

as $M \to \infty$. Thus $\hat{\boldsymbol{\beta}}$ converges in distribution to a constant random variable, which means that it converges in probability too. This is what we were assuming.

Equations (2.104), (2.108), and (2.109) remain valid when $\boldsymbol{F}(\boldsymbol{\beta}^*)$ is replaced by $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$.

2.11.6 Next Step

Now that we have seen how to estimate the parameters and some of their sampling properties (asymptotically), we can move on to use these estimates to make inferences: predictions about new values and confidence intervals.

2.12 Exercises

1. Show that $\boldsymbol{F}(\boldsymbol{\beta}) = nF_1(\boldsymbol{\beta})$, where $\boldsymbol{F}(\boldsymbol{\beta})$ and $F_1(\boldsymbol{\beta})$ are the Fisher Information on datasets of size n and 1 respectively.
2. Using the notation and argument in Section 2.11, show that $\boldsymbol{F}(\boldsymbol{\beta}^*) \stackrel{a}{=} \boldsymbol{F}(\hat{\boldsymbol{\beta}})$ and thus we can replace $\boldsymbol{F}(\boldsymbol{\beta}^*)$ by $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$ in Equations (2.104), (2.108), and (2.109).

Bibliography

P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *JRSSB*, 46(2):149–192, 1984.

John A Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.