

## Dispersion Models

### 2.1 Introduction

In the analysis of correlated data, it is relatively easy to recognize one-dimensional marginal distribution for each of the response vectors. In the example of Indonesian children's health study in Section 1.3.1, the univariate response at a given visit is the infection status, which takes two values with 1 representing the presence of infection and 0 otherwise. Obviously, the marginal distribution of such a binary response variable is Bernoulli or binomial with the size parameter equal to one. In some cases where marginal distributions are subtle to determine, one may apply some model diagnostic tools to check the assumption of marginal distributions. For example, univariate histograms, quantile-quantile plots, and some residual-based model diagnostics in univariate regression analysis, whichever is suitable, could be applied to draw some preliminary understanding of marginal distributions. In the GLMs, the diagnostic analysis of distributional assumptions is carried out through primarily validating the so-called mean and variance relationship. As far as a correlated data analysis concerns, the knowledge of marginal distributions is not yet developed enough to specify a full joint probability model for the data, and a proper statistical inference has to address the correlation among the components of the response vector. Failing to incorporate the correlation in the data analysis will, in general, result in a certain loss of efficiency in the estimation for the model parameters, which may cause misleading conclusions on statistical significance for some covariates.

There are two essential approaches to handling the correlation. One is to construct a full probability model that integrates the marginal distributions and the correlation coherently; within such a framework, the maximum likelihood estimation and inference can be then established. When the joint model is adequately specified, this approach is preferable, because the maximum likelihood method provides a fully efficient inference. Such an approach has been extensively investigated in the class of multivariate normal distributions. However, for many nonnormal data types, constructing a suitable joint

probability distribution is not trivial, and relatively less effort on this matter has been made in the literature in comparison to other areas of research in statistics. In particular, the construction of multivariate discrete distributions, such as multivariate binomial distributions and multivariate Poisson distributions, is still under debate, particularly as to which of many versions of their multivariate extensions is desirable relative to the others. More details concerning this approach will be presented in Chapters 6 and 7. Two major classes of joint probability models are specified via, respectively, Gaussian copulas and random effects.

To avoid the difficulty of specifying a full probability model, the second approach takes a compromise; that is, it only specifies the first two moments of the data distribution. This approach constitutes the minimal requirements for a quasi-likelihood inference procedure. Although the resulting estimation is less efficient than the MLE, it enjoys the robustness against model misspecifications on higher moments. This quasi-likelihood inference would be the choice when robustness appears to be more appealing than efficiency in a given data analysis. A kind of such a quasi-likelihood approach, known as generalized estimating equations (GEE), will be discussed in Chapter 5.

To proceed, it is needed to first outline the marginal parametric distributions that will be used to develop either the full probability model approach or the quasi-likelihood approach. Marginal distributions are the essential pieces to formulate both inference approaches in correlated data analysis. To some extent, the breadth of marginal distributions determines the variety of data types that the proposed inference can handle. This means if one only considers marginal normal distributions, the resulting inference would be merely restricted to continuous data type.

This chapter is devoted to a review of the theory of dispersion models based primarily on Jørgensen's (1997) book, *The theory of dispersion models*. The dispersion models provide a rich class of one-dimensional parametric distributions for various data types, including those commonly considered in the GLM analysis. In effect, error distributions in the GLMs form a special subclass of the dispersion models, which are the *exponential dispersion models*. This means that the GLMs considered in this chapter, as well as in the entire book, encompass a wider scope of GLMs than those outlined in McCullagh and Nelder's (1989) book. Two special examples are the von Mises distribution for directional (circular or angular) data and the simplex distribution for compositional (or proportional) data, both of which are the dispersion models but not the exponential dispersion models.

According to McCullagh and Nelder (1989), the random component of a GLM is specified by an exponential dispersion (ED) family density of the following form:

$$p(y; \theta, \phi) = \exp \left[ \frac{\{y\theta - \kappa(\theta)\}}{a(\phi)} + C(y, \phi) \right], y \in \mathcal{C}, \quad (2.1)$$

with parameters  $\theta \in \Theta$  and  $\phi > 0$ , where  $\kappa(\cdot)$  is the cumulant generating function and  $\mathcal{C}$  is the support of the density. It is known that the first derivative of the cumulant function  $\kappa(\cdot)$  gives the expectation of the distribution, namely  $\mu = E(Y) = \dot{\kappa}(\theta)$ . Table 2.1 lists some ED distributions.

**Table 2.1.** Some commonly used exponential dispersion GLMs.

Distribution Domain		Data type	Canonical link Model	
Normal	$(-\infty, \infty)$	Continuous	Identity	Linear model
Binomial	$\{0, 1, \dots, n\}$	Binary or counts	Logit	Logistic model
Poisson	$\{0, 1, \dots, \}$	Counts	Log	Loglinear model
Gamma	$(0, \infty)$	Positive continuous	Reciprocal	Reciprocal model

The systematic component of a GLM is then assumed to take the form:

$$g(\mu) = \mathbf{x}^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (2.2)$$

where  $g$  is the link function,  $\mathbf{x} = (1, x_1, \dots, x_p)^T$  is a  $(p+1)$ -dimensional vector of covariates, and  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$  is a  $(p+1)$ -dimensional vector of regression coefficients. The *canonical link* function  $g(\cdot)$  is such that  $g(\mu) = \theta$ , the canonical parameter.

The primary statistical tasks include estimation and inference for  $\boldsymbol{\beta}$ . Checking model assumptions is also an important task of regression analysis, which, however, is not the main focus of the book.

## 2.2 Dispersion Models

The normal distribution  $N(\mu, \sigma^2)$  plays the central role in the classical linear regression regression. The density of  $N(\mu, \sigma^2)$  is

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\}, \quad y \in \mathcal{R},$$

where  $(y - \mu)^2$  can be regarded as an Euclidean distance that measures the discrepancy between the observed  $y$  and the expected  $\mu$ . And this discrepancy measure is used to develop many regression analysis methods, such as the  $F$ -statistic for the assessment of goodness-of-fit for nested models.

Mimicking the normal density, Jørgensen (1987) defines a dispersion models (DM) by extending the Euclidean distance  $(y-\mu)^2$  to a general discrepancy function  $d(y; \mu)$ . It is found that many commonly used parametric distributions, such as those in Table 2.1, are included as special cases of this extension. Moreover, each of such distributions will be determined uniquely by the discrepancy function  $d$ , and the resulting distribution is fully parameterized by two parameters  $\mu$  and  $\sigma^2$ .

### 2.2.1 Definitions

A (*reproductive*) *dispersion model*  $\text{DM}(\mu, \sigma^2)$  with *location parameter*  $\mu$  and *dispersion parameter*  $\sigma^2$  is a family of distributions whose probability density functions take the following form:

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \quad y \in \mathcal{C} \quad (2.3)$$

where  $\mu \in \Omega$ ,  $\sigma^2 > 0$ , and  $a \geq 0$  is a suitable normalizing term that is independent of the  $\mu$ . Usually,  $\Omega \subseteq \mathcal{C} \subseteq \mathcal{R}$ . The fact that the normalizing term  $a$  does not involve  $\mu$  will allow to estimate  $\mu$  (or  $\beta$  in the GLM setting) separately from estimating  $\sigma^2$ , which gives rise to great ease in the parameter estimation. Such a nice property, known as the likelihood orthogonality, holds in the normal distribution, and it will remain in the dispersion models.

A bivariate function  $d(\cdot; \cdot)$  is called the *unit deviance* defined on  $(y, \mu) \in \mathcal{C} \times \Omega$  if it satisfies the following two properties:

- i) It is zero when the observed  $y$  and the expected  $\mu$  are equal, namely

$$d(y; y) = 0, \quad \forall y \in \Omega;$$

- ii) It is positive when the observed  $y$  and the expected  $\mu$  are different, namely

$$d(y; \mu) > 0, \quad \forall y \neq \mu.$$

Furthermore, a unit deviance is called *regular* if function  $d(y; \mu)$  is twice continuously differentiable with respect to  $(y, \mu)$  on  $\Omega \times \Omega$  and satisfies

$$\frac{\partial^2 d}{\partial \mu^2}(y; y) = \frac{\partial^2 d}{\partial \mu^2}(y; \mu) \Big|_{\mu=y} > 0, \quad \forall y \in \Omega.$$

For a regular unit deviance, the variance function is defined as follows. The *unit variance function*  $V : \Omega \rightarrow (0, \infty)$  is

$$V(\mu) = \frac{2}{\frac{\partial^2 d}{\partial \mu^2}(y; \mu) \Big|_{y=\mu}}, \quad \mu \in \Omega. \quad (2.4)$$

Some popular dispersion models are given in Table 2.2, in which the unit deviance  $d$  and variance function  $V$  can be found in a similar fashion to that presented in the following two examples.

**Table 2.2.** Unit deviance and variance functions of some dispersion models.

Distribution	Deviance $d$	$\mathcal{C}$	$\Omega$	$V(\mu)$
Normal	$(y - \mu)^2$	$(-\infty, \infty)$	$(-\infty, \infty)$	1
Poisson	$2(y \log \frac{y}{\mu} - y + \mu)$	$\{0, 1, \dots\}$	$(0, \infty)$	$\mu$
Binomial	$2 \left\{ y \log \frac{y}{\mu} + (n - y) \log \frac{n-y}{n-\mu} \right\}$	$\{0, 1, \dots, n\}$	$(0, 1)$	$\mu(1 - \mu)$
Negative binomial	$2 \left\{ y \log \frac{y}{\mu} + (1 - y) \log \frac{1-y}{1-\mu} \right\}$	$\{0, 1, \dots\}$	$(0, \infty)$	$\mu(1 + \mu)$
Gamma	$2 \left( \frac{y}{\mu} - \log \frac{y}{\mu} - 1 \right)$	$(0, \infty)$	$(0, \infty)$	$\mu^2$
Inverse Gaussian	$\frac{(y-\mu)^2}{y\mu^2}$	$(0, \infty)$	$(0, \infty)$	$\mu^3$
von Mises	$2\{1 - \cos(y - \mu)\}$	$(0, 2\pi)$	$(0, 2\pi)$	1
Simplex	$\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$	$(0, 1)$	$(0, 1)$	$\mu^3(1 - \mu)^3$

*Example 2.1 (Normal Distribution).* In the normal distribution  $N(\mu, \sigma^2)$ , first the unit deviance function  $d(y; \mu) = (y - \mu)^2$ ,  $y \in \mathcal{C} = \mathcal{R}$ , and  $\mu \in \Omega = \mathcal{R}$ . It is easy to see that this  $d$  function is non-negative and has the unique minimum 0 when  $y = \mu$ . This unit deviance is regular because it is twice continuously differentiable. Moreover, the first and second order derivatives of the  $d$  function *w.r.t.*  $\mu$  are, respectively,

$$\frac{\partial d}{\partial \mu} = -2(y - \mu), \text{ and } \frac{\partial^2 d}{\partial \mu^2} = 2.$$

It follows that the unit variance function is  $V(\mu) = \frac{2}{2} = 1$ .

*Example 2.2 (Poisson Distribution).* To verify the results of the Poisson distribution given in Table 2.2, express the Poisson density with mean parameter  $\mu$  as follows:

$$p(y; \mu) = \frac{\mu^y}{y!} e^{-\mu}, y \in \{0, 1, \dots\}; \mu \in \Omega = (0, \infty),$$

or equivalently

$$p(y; \mu) = \frac{1}{y!} \exp\{y \log \mu - \mu\}.$$

Note that the exponent  $\{y \log \mu - \mu\}$  is not a deviance function because it does not equal to zero when  $y = \mu$ . To yield a deviance function, a new term  $\{y \log y - y\}$  is added into the exponent, which results in

$$p(y; \mu) = \left\{ \frac{1}{y!} \exp(y \log y - y) \right\} \exp \left\{ -\frac{1}{2} 2(y \log y + y - y \log \mu + \mu) \right\}.$$

Comparing to the DM density in (2.3), one can identify the  $d$  function, the normalizing term, and the dispersion parameter, respectively,

$$\begin{aligned} d(y; \mu) &= 2(y \log \frac{y}{\mu} - y + \mu), \\ a(y) &= \frac{1}{y!} \exp\{y \log y - y\}, \\ \sigma^2 &= 1. \end{aligned}$$

To show this  $d$  function is a regular deviance function, it is sufficient to show it is convex with a unique minimum of zero. First, note that at a given mean value  $\mu$ , the first and second order derivatives of the  $d$  w.r.t.  $y$  are

$$\frac{\partial d}{\partial y} = 2(\log y - \log \mu), \text{ and } \frac{\partial^2 d}{\partial y^2} = \frac{2}{y}.$$

Clearly, the first order derivative is negative when  $y < \mu$  and positive when  $y > \mu$ , implying that the  $d$  is a convex function with a unique minimum 0 at  $y = \mu$ . Thus, the  $d$  function is a regular unit deviance for the Poisson distribution.

To find the unit variance function, note that the second order derivative  $\frac{\partial^2 d}{\partial \mu^2} = 2 \frac{y}{\mu^2}$ , which immediately leads to  $V(\mu) = \mu$  by the definition (2.4).

### 2.2.2 Properties

This section lists some useful properties of the dispersion models.

**Proposition 2.3.** *If a unit deviance  $d$  is regular, then*

$$\frac{\partial^2 d}{\partial y^2}(y; y) = \frac{\partial^2 d}{\partial \mu^2}(y; y) = -\frac{\partial^2 d}{\partial \mu \partial y}(y; y), \quad \forall y \in \Omega. \quad (2.5)$$

*Proof.* By the definition of a unit deviance,

$$d(y; y) = d(\mu; \mu) = 0 \text{ and } d(y; \mu) \geq 0, \quad \forall y, \mu \in \Omega,$$

implying that  $d(y; \cdot)$  has a unique minimum at  $y$  and similarly  $d(\cdot; \mu)$  has a unique minimum at  $\mu$ . Therefore,

$$\frac{\partial d}{\partial \mu}(y; y) = \frac{\partial d}{\partial y}(y; y) = 0. \quad (2.6)$$

The result of (2.5) holds by simply differentiating both equations in (2.6) w.r.t.  $y$ .

**Proposition 2.4.** *Taylor expansion of a regular unit deviance  $d$  near its minimum  $(\mu_0, \mu_0)$  is given by*

$$d(\mu_0 + x\delta; \mu_0 + m\delta) = \frac{\delta^2}{V(\mu_0)}(x - m)^2 + o(\delta^2),$$

where  $V(\cdot)$  is the unit variance function.

*Proof.* It follows from equation (2.6) that

$$\begin{aligned} d(\mu_0 + x\delta; \mu_0 + m\delta) &= d(\mu_0, \mu_0) + \frac{\partial d}{\partial \mu}(\mu_0, \mu_0)(x\delta) + \frac{\partial d}{\partial y}(\mu_0, \mu_0)(m\delta) \\ &\quad + \frac{1}{2} \frac{\partial^2 d}{\partial \mu^2}(\mu_0, \mu_0)(\delta^2 x^2) + \frac{1}{2} 2 \frac{\partial^2 d}{\partial \mu \partial y}(\mu_0, \mu_0)(\delta m) \\ &\quad + \frac{1}{2} \frac{\partial^2 d}{\partial y^2}(\mu_0, \mu_0)(\delta^2 m^2) + o(\delta^2) \\ &= \frac{\delta^2}{V(\mu_0)} x^2 - \frac{\delta^2}{V(\mu_0)} 2xm + \frac{\delta^2}{V(\mu_0)} m^2 + o(\delta^2) \\ &= \frac{\delta^2}{V(\mu_0)} (x - m)^2 + o(\delta^2). \end{aligned}$$

In some cases, the normalizing term  $a(\cdot)$  has no closed form expression, which gives rise to the difficulty of estimating the dispersion parameter  $\sigma^2$ . The following proposition presents an approximation to the normalizing term  $a(\cdot)$ , resulting from the saddlepoint approximation of the density for small dispersion. Notation  $a \simeq b$  exclusively stands for an approximation of  $a$  to  $b$  when the dispersion  $\sigma^2 \rightarrow 0$ , useful for small-dispersion asymptotics.

**Proposition 2.5 (Saddlepoint approximation).** *As the dispersion  $\sigma^2 \rightarrow 0$ , the density of a regular DM model can be approximated to be:*

$$p(y; \mu, \sigma^2) \simeq \{2\pi\sigma^2 V(y)\}^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\},$$

which equivalently says that as  $\sigma^2 \rightarrow 0$ , the normalizing term has a small dispersion approximation,

$$a(y; \sigma^2) \simeq \{2\pi\sigma^2 V(y)\}^{-1/2}, \quad (2.7)$$

with the unit variance function  $V(\cdot)$ .

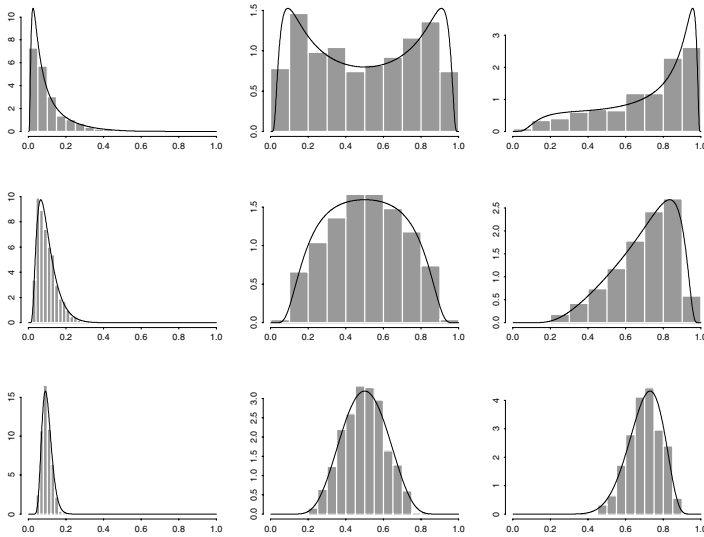
The proof of this proposition is basically an application of the Laplace approximation given in, for example, Barndorff-Nielsen and Cox (1989, p.60). Also see Jørgensen (1997, p.28).

It follows from Propositions 2.4 and 2.5 that the small dispersion asymptotic normality holds, as stated in the following:

**Proposition 2.6 (Asymptotic Normality).** : Let  $Y \sim DM(\mu_0 + \sigma\mu, \sigma^2)$  be a dispersion model with uniformly convergent saddlepoint approximation, namely convergence in (2.7) is uniformly in  $y$ . Then

$$\frac{Y - \mu_0}{\sigma} \xrightarrow{d} N(\mu, V(\mu_0)) \text{ as } \sigma^2 \rightarrow 0.$$

In other words,  $DM(\mu_0 + \sigma\mu, \sigma^2) \stackrel{d}{\simeq} N(\mu_0 + \sigma\mu, \sigma^2 V(\mu_0))$  for small dispersion  $\sigma^2$ .



**Fig. 2.1.** Simplex density functions with mean  $\mu = (0.1, 0.5, 0.7)$  from left to right and dispersion parameter  $\sigma^2 = (4^2, 2^2, 1)$  from top to bottom. The solid lines represent the simplex densities with the histograms as the background. These histograms are based on 500 simulated data from respective densities.

To illustrate this small-dispersion asymptotic normality, Figure 2.1 displays the simplex distributions with different mean  $\mu$  and dispersion  $\sigma^2$  parameters. See the detail of a simplex distribution in Table 2.2. This figure clearly indicates that the smaller the dispersion is, the less deviation the simplex distribution is from the normality.

## 2.3 Exponential Dispersion Models

The class of dispersion models contains two important subclasses, namely *the exponential dispersion (ED) models* and *the proper dispersion (PD) models*.



The PD models are mostly of theoretical interest, so they are not discussed in this book. Readers may refer to the book of Jørgensen (1997) for relevant details.

This section focuses on the ED models, which have already been introduced in Section 2.1 as a family of GLMs' error distributions. The family of ED models includes continuous distributions such as normal, gamma, and inverse Gaussian, and discrete distributions such as Poisson, binomial, negative binomial, among others. To establish the connection of the ED model representation (2.1) to the DM, it is sufficient to show that expression (2.1) is a special form of (2.3). An advantage with the DM type of parametrization for the ED models is that both mean  $\mu$  and dispersion parameters  $\sigma^2$  are explicitly present in the density, whereas expression (2.1) hides the mean  $\mu$  in the first order derivative  $\mu = \dot{\kappa}(\theta)$ . In addition, having a density form similar to the normal enables us to easily borrow the classical normal regression theory to the development of regression analysis for nonnormal data. One example is the analogue of the likelihood ratio test in the GLMs to the F-test for goodness-of-fit in the normal regression model.

To show an ED model, denoted by  $\text{ED}(\mu, \sigma^2)$ , as a special case of the DM, it suffices to find a unit deviance function  $d$  such that the density of the ED model can be expressed in the form of (2.3). First, denote  $\lambda = 1/a(\phi)$ . Then, the density in (2.1) can be rewritten as of the form:

$$p(y; \theta, \lambda) = c(y; \lambda) \exp[\lambda\{\theta y - \kappa(\theta)\}], \quad y \in \mathcal{C} \quad (2.8)$$

where  $c(\cdot)$  is a suitable normalizing term. Parameter  $\lambda = 1/\sigma^2 \in \Lambda \subset (0, \infty)$  is called the *index parameter* and  $\Lambda$  is called the index set. To reparametrize this density (2.1) by the mean  $\mu$  and dispersion  $\sigma^2$ , define the *mean value mapping*:  $\tau : \text{int}\Theta \rightarrow \Omega$ ,

$$\tau(\theta) = \dot{\kappa}(\theta) \equiv \mu,$$

where  $\text{int}(\Theta)$  is the interior of the parameter space  $\Theta$ .

**Proposition 2.7.** *The mean mapping function  $\tau(\theta)$  is strictly increasing.*

*Proof.* The property of the natural exponential family distribution leads to

$$\text{Var}(Y) = \lambda \ddot{\kappa}(\theta) > 0, \quad \theta \in \text{int}\Theta.$$

In the mean time, because  $\dot{\tau}(\theta) = \ddot{\kappa}(\theta)$ ,  $\dot{\tau}(\cdot)$  is positive. This implies that  $\tau(\theta)$  is a strictly increasing function in  $\theta$ .

It follows that the inverse of the mean mapping function  $\tau(\cdot)$  exists, denoted by  $\theta = \tau^{-1}(\mu)$ ,  $\mu \in \Omega$ . Hence, the density in (2.8) can be reparametrized as follows,

$$p(y; \mu, \sigma^2) = c(y; \sigma^{-2}) \exp \left[ \frac{1}{\sigma^2} \{y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))\} \right]. \quad (2.9)$$

**Proposition 2.8.** *The first order derivative of  $\tau^{-1}(\mu)$  with respect to  $\mu$  is  $1/V^*(\mu)$ , where  $V^*(\mu) = \dot{\tau}(\tau^{-1}(\mu))$ .*

*Proof.* Differentiating both sides of equation  $\mu = \tau(\theta)$  gives

$$d\mu = \dot{\tau}(\theta)d\theta = \dot{\tau}(\tau^{-1}(\mu))d\theta = V^*(\mu)d\theta,$$

with  $V^*(\mu) = \dot{\tau}(\tau^{-1}(\mu))$ . This implies immediately that

$$\frac{d\tau^{-1}(\mu)}{d\mu} = \frac{d\theta}{d\mu} = \frac{1}{V^*(\mu)}.$$

Moreover, Proposition 2.9 below shows that the  $V^*(\mu)$  is indeed the same as the unit variance function  $V(\mu)$  given by the definition (2.4). The proof of this result will be given after the unit deviance function of the ED model is derived.

To derive the unit deviance function of the ED model, let

$$f(y; \mu) = y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu)) = y\theta - \kappa(\theta).$$

Obviously, this  $f$  is not the unit deviance function since it does not equal to zero when  $\mu = y$ . One way to resolve this problem is to add a new term so that the resulting function is positive and equal to zero uniquely at  $\mu = y$ . Such a valley point corresponds effectively to the maximum of the density  $p(y; \mu, \sigma^2)$ .

Differentiating  $f$  with respect to  $\mu$  and using Propositions 2.8 and 2.9, one can obtain

$$\dot{f}(y, \mu) = \frac{y - \mu}{V(\mu)}, \quad (2.10)$$

which is positive for  $y > \mu$  and negative for  $y < \mu$ . This means that the  $f$  has a unique maximum, or equivalently, the  $-f$  has a unique minimum at  $\mu = y$ . Therefore, it seems natural to define

$$\begin{aligned} d(y; \mu) &= 2 \left[ \sup_{\mu} \{f(y; \mu)\} - f(y; \mu) \right] \\ &= 2 \left[ \sup_{\theta \in \Theta} \{\theta y - \kappa(\theta)\} - y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu)) \right]. \end{aligned} \quad (2.11)$$

Clearly, this  $d$  function satisfies (i)  $d(y; \mu) \geq 0$  for all  $y \in \mathcal{C}$  and  $\mu \in \Omega$ , and (ii)  $d(y; \mu)$  attains the minimum at  $\mu = y$  because the supremum term is independent of  $\mu$ . Thus, (2.11) gives a proper unit deviance function. Moreover, since it is continuously twice differentiable, it is also regular. As a result, the density of an ED model can be expressed as of the DM form:

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\},$$

with the unit deviance function  $d$  given in (2.11) and the normalizing term given by

$$a(y; \sigma^2) = c(y; \sigma^{-2}) \exp \left[ \sigma^{-2} \sup_{\theta \in \Theta} \{y\theta - \kappa(\theta)\} \right].$$

**Proposition 2.9.** *For the unit deviance function (2.11), the corresponding unit variance function  $V(\mu)$  given in (2.4) is  $V^*(\mu) = \dot{\tau}(\tau^{-1}(\mu))$ ; that is,  $V(\mu) = V^*(\mu)$ .*

*Proof.* It follows from equations (2.10) and (2.11) that

$$\frac{\partial d}{\partial \mu} = -2 \frac{\partial f}{\partial \mu} = -2 \frac{y - \mu}{V^*(\mu)},$$

where  $V^*(\mu) = \dot{\tau}(\tau^{-1}(\mu))$ . Then, according to Proposition 2.3,

$$\frac{\partial^2 d}{\partial \mu^2} = -\frac{\partial^2 d}{\partial y \partial \mu} = \frac{2}{V^*(\mu)}.$$

Plugging this into the definition of the unit variance function (2.4) leads to

$$V(\mu) = \frac{2}{\frac{\partial^2 d}{\partial \mu^2}(y; \mu)|_{y=\mu}} = V^*(\mu).$$

Here are a few remarks for the ED models:

- (1) Parameter  $\mu$  is the mean of the distribution, namely  $E(Y) = \mu$ .
- (2) Variance of the distribution is

$$\text{Var}(Y) = \sigma^2 V(\mu). \quad (2.12)$$

This mean-variance relationship is one of the key properties for the ED models, which will play an important role in the development of quasi-likelihood inference.

- (3) An important variant of the reproductive ED model representation is the so-called *additive exponential dispersion model*, denoted by  $\text{ED}^*(\theta, \lambda)$ , whose density takes the form

$$p^*(z; \theta, \lambda) = c^*(z; \lambda) \exp\{\theta z - \lambda \kappa(\theta)\}, \quad z \in \mathcal{C}. \quad (2.13)$$

Essentially the ED and  $\text{ED}^*$  representations are equivalent under the *duality transformation* that converts one form to the other.

Suppose  $Z \sim \text{ED}^*(\theta, \lambda)$  and  $Y \sim \text{ED}(\mu, \sigma^2)$ . Then, the duality transformation performs

$$\begin{aligned} Z \sim \text{ED}^*(\theta, \lambda) &\Rightarrow Y = Z/\lambda \sim \text{ED}(\mu, \sigma^2), \text{ with } \mu = \tau(\theta), \sigma^2 = 1/\lambda; \\ Y \sim \text{ED}(\mu, \sigma^2) &\Rightarrow Z = Y/\sigma^2 \sim \text{ED}^*(\theta, \lambda), \text{ with } \theta = \tau^{-1}(\mu), \lambda = 1/\sigma^2. \end{aligned}$$

Consequently, the mean and variance of  $\text{ED}^*(\theta, \lambda)$  are, respectively,

$$\mu^* = E(Z) = \lambda\tau(\theta), \quad \text{Var}(Z) = \lambda V(\mu^*/\lambda).$$

Moreover, the normalizing term in the DM density (2.3) is

$$a^*(z; \sigma^2) = c^*(z; \sigma^{-2}) \exp \left[ \sigma^{-2} \sup_{\theta \in \Theta} \{z\theta - \kappa(\theta)\} \right].$$

An important property for the ED models is the closure under convolution operation.

**Proposition 2.10 (Convolution for the  $ED^*$  models).** *Assume  $Z_1, \dots, Z_n$  are independent and  $Z_i \sim ED^*(\theta, \lambda_i)$ ,  $i = 1, \dots, n$ . Then the sum follows still an  $ED^*$  model:*

$$Z_+ = Z_1 + \dots + Z_n \sim ED^*(\theta, \lambda_1 + \dots + \lambda_n).$$

For example, consider two independent and identically distributed (*i.i.d.*) Poisson random variables  $Z_i \sim ED^*(\log \mu, 1)$ ,  $i = 1, 2$ , where  $\mu$  is the mean parameter and the canonical parameter  $\theta = \log(\mu)$ . Then, Proposition 2.10 implies that the sum  $Z_+ = Z_1 + Z_2 \sim ED^*(\log \mu, 2)$ .

**Proposition 2.11 (Convolution for the ED models).** *Assume  $Y_1, \dots, Y_n$  are independent and*

$$Y_i \sim ED(\mu, \frac{\sigma^2}{w_i}), i = 1, \dots, n,$$

where  $w_i$ s are certain positive weights. Let  $w_+ = w_1 + \dots + w_n$ . Then the weighted average follows still an ED model; that is,

$$\frac{1}{w_+} \sum_{i=1}^n w_i Y_i \sim ED(\mu, \frac{\sigma^2}{w_+}).$$

In particular, with  $w_i = 1$ ,  $i = 1, \dots, n$  the sample average

$$\frac{1}{n} \sum_{i=1}^n Y_i \sim ED(\mu, \frac{\sigma^2}{n}).$$

For the example of two *i.i.d.* Poisson random variables with  $Y_i \sim ED(\mu, 1)$ ,  $i = 1, 2$ , their average  $(Y_1 + Y_2)/2 \sim ED(\mu, \frac{1}{2})$ . Note that the resulting  $ED(\mu, \frac{1}{2})$  is no longer a Poisson distribution but it is still an ED distribution.

It is noticeable that although the class of the ED models is closed under the convolution operation, it is in general not closed under scale transformation. That is,  $cY$  may not follow an ED model even if  $Y \sim ED(\mu, \sigma^2)$ , for a constant  $c$ . However, a subclass of the ED models, termed as the *Tweedie class*, is closed under this type of scale transformation. The Tweedie models will be discussed in Section 2.5.

Finally, the following property concerns sufficient and necessary conditions for the de-convolution for the ED models.

**Definition 2.12 (Infinite Divisibility).**  $X$  is said to be infinitely divisible, if for any integer  $n \in \{1, 2, \dots\}$ , there exist i.i.d. random variables  $X_1, \dots, X_n$  such that  $X \stackrel{d}{=} X_1 + \dots + X_n$ . Notation  $U \stackrel{d}{=} V$  means that two random variables  $U$  and  $V$  are identically distributed.

**Proposition 2.13 (Deconvolution for the ED\*).** Suppose  $Z \sim ED^*(\theta, \lambda)$ . Then,  $Z$  is infinitely divisible if and only if the index parameter set  $\Lambda = (0, \infty)$ .

This result holds simply because by Proposition 2.10 there exist  $X_i \sim ED^*(\theta, \lambda/n), i = 1, \dots, n$  such that

$$ED^*(\theta, \lambda) = ED^*(\theta, \lambda/n) + \dots + ED^*(\theta, \lambda/n).$$

It is easy to see that gamma models are infinitely divisible, but binomial models are not infinitely divisible.

## 2.4 Residuals

Residual analysis is an important part of regression analysis. In the context of the dispersion models where the unit deviance functions  $d$  are highly non-linear in comparison to the square normal deviance  $(y - \mu)^2$  of the normal model, there are several other types of residuals besides the traditional Pearson residual  $(y - \mu)$ . Table 2.3 lists some proposed residuals in the GLMs. Among them, the Pearson and deviance residuals are most commonly used in practice, which are in fact implemented in statistical softwares such as SAS. For example, SAS PROC GENMOD uses the deviance residual in the analysis of outliers and influential data cases.

**Table 2.3.** Some types of residuals in the GLMs.

Type	Notation	Definition
Pearson residual	$r_p$	$\frac{y - \mu}{V^{1/2}(\mu)}$
Score residual	$r_s$	$-\frac{\partial d}{2\partial \mu} V^{1/2}(\mu)$
Dual score residual	$r_d$	$\frac{\partial d}{2\partial y} V^{1/2}(\mu)$
Deviance residual	$r$	$\pm d^{1/2}(y; \mu)$
Modified deviance residual	$r^*$	$\frac{r}{\sigma} + \frac{\sigma}{r} \log \frac{r_d}{r}$

Besides the residual analysis for model diagnosis, another important application of residuals is in the approximation of tail area probabilities with small dispersion. Calculating tail probabilities is often encountered, such as in the calculation of  $p$ -values. Most of cumulative distribution functions (CDFs) of the ED models have no closed form expressions, so a certain approximation to their CDF is useful.

Let  $F(y; \mu, \sigma^2)$  be the CDF of an  $ED(\mu, \sigma^2)$ . By Proposition 2.6, the small dispersion asymptotic normality gives

$$F(y; \mu, \sigma^2) \simeq \Phi(r_p/\sigma) \text{ for } \sigma^2 \text{ small,}$$

where  $\Phi$  is the CDF of the standard normal  $N(0, 1)$ . This result is based on the Pearson residual  $r_p$ . Because it is a first-order linear approximation, this approximation may not be satisfactorily accurate when the unit deviance  $d$  is highly nonlinear.

Two formulas based on the so-called third-order approximation provide much more accurate approximations for the CDF of the DM model. One is Barndorff-Nielsen's formula given by,

$$F(y; \mu, \sigma^2) = \Phi(r^*)\{1 + O(\sigma^3)\},$$

where  $r^*$  is the modified deviance residual given in Table 2.3. The other is Lugannani-Rice's formula

$$F(y; \mu, \sigma^2) = \Phi^*(y; \mu, \sigma^2)\{1 + O(\sigma^3)\},$$

where

$$\Phi^*(y; \mu, \sigma^2) = \Phi\left(\frac{r}{\sigma}\right) + \sigma\phi\left(\frac{r}{\sigma}\right)\left(\frac{1}{r} - \frac{1}{r_d}\right),$$

where  $r$  is the deviance residual and  $\phi$  is the density of the standard normal  $N(0, 1)$ .

## 2.5 Tweedie Class

Tweedie class is an important subclass of the ED models, which is closed under the scale transformation. Tweedie models are characterized by the unit variance functions in the form of the power function:

$$V_p(\mu) = \mu^p, \mu \in \Omega_p, \quad (2.14)$$

where  $p \in R$  is a *shape* parameter.

It is shown that the ED model with the power unit variance function (2.14) always exists except  $0 < p < 1$ . A Tweedie model is denoted by  $Y \sim Tw_p(\mu, \sigma^2)$  with mean  $\mu$  and variance

$$\text{Var}(Y) = \sigma^2 \mu^p.$$

The following proposition gives the characterization of the Tweedie models.

**Proposition 2.14 (Tweedie Characterization).** *Let  $ED(\mu, \sigma^2)$  be a reproductive ED model satisfying  $V(1) = 1$  and  $1 \in \Omega$ . If the model is closed with respect to scale transformation, such that there exists a function  $f : R_+ \times \Lambda^{-1} \rightarrow \Lambda^{-1}$  for which*

$$cED(\mu, \sigma^2) \sim ED[c\mu, f(c, \sigma^2)], \forall c > 0,$$

then

- (a)  $ED(\mu, \sigma^2)$  is a Tweedie model for some  $p \in R \setminus (0, 1)$ ;
- (b)  $f(c, \sigma^2) = c^{2-p}\sigma^2$ ;
- (c) the main domain  $\Omega = R$  for  $p = 0$  and  $\Omega = (0, \infty)$  for  $p \neq 0$ ;
- (d) the model is infinitely divisible.

It follows immediately from Proposition 2.14 that

$$cTw_p(\mu, \sigma^2) = Tw_p(c\mu, c^{2-p}\sigma^2).$$

The importance of the Tweedie class is that it serves as a class of limiting distributions of the ED models, as described in the following proposition.

**Definition 2.15.** *The unit variance function  $V$  is said to be regular of order  $p$  at 0 (or at  $\infty$ ), if  $V(\mu) \sim c_0\mu^p$  as  $\mu \rightarrow 0$  (or  $\mu \rightarrow \infty$ ) for certain  $p \in \mathcal{R}$  and  $c_0 > 0$ .*

**Proposition 2.16.** *Suppose the unit variance function  $V$  is regular of order  $p$  at 0 or at  $\infty$ , with  $p \notin (0, 1)$ . For any  $\mu > 0$  and  $\sigma^2 > 0$ ,*

$$c^{-1}ED(c\mu, c^{2-p}\sigma^2) \xrightarrow{d} TW_p(\mu, c_0\sigma^2), \text{ as } c \rightarrow 0 \text{ or } \infty,$$

where the convergence is through values of  $c$  such that  $c\mu \in \Omega$  and  $c^{p-2}/\sigma^2 \in \Lambda$ .

Refer to Jørgensen et al. (1994) for the proof of this result.

## 2.6 Maximum Likelihood Estimation

This section is devoted to maximum likelihood estimation in the GLMs based on the dispersion models. Therefore, the MLE theory given in, for example, McCullagh and Nelder (1989) are the special cases, because the ED family is a subclass of the DM family.

### 2.6.1 General Theory

Consider a cross-sectional dataset,  $(y_i, \mathbf{x}_i), i = 1, \dots, K$ , where the  $y_i$ 's are *i.i.d.* realizations of  $Y_i$ 's according to  $\text{DM}(\mu_i, \sigma^2)$  and  $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ . Let  $\mathbf{y} = (y_1, \dots, y_K)^T$  and  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T$ . The likelihood for the parameter vector  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$  is given by

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^K a(y_i; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y_i; \mu_i) \right\}, \quad \boldsymbol{\beta} \in \mathcal{R}^{p+1}, \sigma^2 > 0.$$

The log-likelihood is then

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{y}) &= \sum_{i=1}^K \log a(y_i; \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^K d(y_i; \mu_i) \\ &= \sum_{i=1}^K \log a(y_i; \sigma^2) - \frac{1}{2\sigma^2} D(\mathbf{y}; \boldsymbol{\mu}), \end{aligned} \quad (2.15)$$

where  $\mu_i = \mu_i(\boldsymbol{\beta})$  is a nonlinear function in  $\boldsymbol{\beta}$  and  $D(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^K d(y_i; \mu_i)$  is the sum of deviances depending on  $\boldsymbol{\beta}$  only. This  $D$  is analogous to the sum of squared residuals in the linear regression model.

The score function for the regression coefficient  $\boldsymbol{\beta}$  is

$$s(\mathbf{y}; \boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} \sum_{i=1}^K \frac{\partial d(y_i; \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

Denote the  $i$ -th linear predictor by  $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ , and denote the *deviance scores* by

$$\delta(y_i; \mu_i) = -\frac{1}{2} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i}, \quad i = 1, \dots, K. \quad (2.16)$$

Note that

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \{\dot{g}(\mu_i)\}^{-1} \mathbf{x}_i,$$

where  $\dot{g}(\mu)$  is the first order derivative of link function  $g$  w.r.t  $\mu$ . Table 2.4 lists some commonly used link functions and their derivatives.

Then the score function for  $\boldsymbol{\beta}$  takes the form

$$s(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^K \mathbf{x}_i \frac{1}{\dot{g}(\mu_i)} \delta(y_i; \mu_i). \quad (2.17)$$

Moreover, the score equation leading to the maximum likelihood estimate of the  $\boldsymbol{\beta}$  is

$$\sum_{i=1}^K \mathbf{x}_i \frac{1}{\dot{g}(\mu_i)} \delta(y_i; \mu_i) = 0. \quad (2.18)$$



**Table 2.4.** Some common link functions and derivatives. NB and IG stand for Negative binomial and Inverse Gaussian, respectively.

Model	Link $g$	Derivative $\dot{g}$	Domain $\Omega$
Binomial or simplex	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{\mu(1-\mu)}$	$\mu \in (0, 1)$
Poisson, NB, gamma, or IG	$\log(\mu)$	$\frac{1}{\mu}$	$\mu \in (0, \infty)$
Gamma	$\frac{1}{\mu}$	$-\frac{1}{\mu^2}$	$\mu \in (0, \infty)$
von Mises	$\tan(\mu/2)$	$\frac{1}{2}\sec^2(\mu/2)$	$\mu \in [-\pi, \pi)$

Note that this equation does not involve the dispersion parameter  $\sigma^2$ . Under some mild regularity conditions, the resulting ML estimator  $\hat{\beta}_K$ , which is the solution to the score equation (2.18), is consistent

$$\hat{\beta}_K \xrightarrow{P} \beta \text{ as } K \rightarrow \infty,$$

and asymptotically normal with mean 0 and covariance matrix  $\mathbf{i}^{-1}(\theta)$ . Here  $\mathbf{i}(\theta)$  is the Fisher information matrix given by

$$\begin{aligned}
 \mathbf{i}(\theta) &= -\mathbf{E}\{\dot{s}(\mathbf{Y}; \beta)\} \\
 &= \frac{1}{\sigma^2} \sum_{i=1}^K \mathbf{x}_i \frac{1}{\{\dot{g}(\mu_i)\}^2} \mathbf{E}\{-\dot{\delta}(Y_i; \mu_i)\} \mathbf{x}_i^T \\
 &= \frac{1}{\sigma^2} \sum_{i=1}^K \mathbf{x}_i u_i^{-1} \mathbf{x}_i^T \\
 &= \mathbf{X}^T U^{-1} \mathbf{X} / \sigma^2,
 \end{aligned} \tag{2.19}$$

where  $\mathbf{X}$  is a  $K \times (p+1)$  matrix with the  $i$ -th row being the  $\mathbf{x}_i^T$ , and  $U$  is a diagonal matrix with the  $i$ -th diagonal element  $u_i$  given by

$$u_i = \frac{\{\dot{g}(\mu_i)\}^2}{\mathbf{E}\{-\dot{\delta}(Y_i; \mu_i)\}}, \quad i = 1, \dots, K. \tag{2.20}$$

When the dispersion parameter  $\sigma^2$  is present in the model, the ML estimation for the dispersion parameter  $\sigma^2$  can be derived similarly, if the normalizing term  $a(y; \sigma^2)$  is simple enough to allow such a derivation, such as the case of the normal distribution. However, in many cases, the term  $a(\cdot)$  has no closed form expression and its derivative *w.r.t.*  $\sigma^2$  may appear too complicated to be numerically solvable. In this case, two methods have been suggested to acquire the estimation for  $\sigma^2$ . The first method is to invoke the small dispersion asymptotic normality (Proposition 2.5), where subject to a constant,

$$\log a(y; \sigma^2) \simeq -\frac{1}{2} \log \sigma^2.$$

Applying this approximation in the log-likelihood (2.15) and differentiating the resulting approximate log-likelihood *w.r.t.*  $\sigma^2$ , one can obtain an equation as follows,

$$-\frac{K}{2\sigma^2} + \frac{1}{2\sigma^4} D(\mathbf{y}; \boldsymbol{\mu}) = 0.$$

Solution to this equation gives an estimator of the dispersion parameter  $\sigma^2$ ,

$$\hat{\sigma}^2 = \frac{1}{K} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \frac{1}{K} \sum_{i=1}^K d(y_i; \hat{\mu}_i). \quad (2.21)$$

This book refers this estimator to as *the Jørgensen estimator* of the dispersion parameter, which in fact is an average of the estimated unit deviances.

However, the Jørgensen estimator is not, in general, unbiased even if the adjustment on the degrees of freedom,  $K - (p + 1)$  is made to replace  $K$ . Moreover, this formula is recommended when the dispersion parameter  $\sigma^2$  is small, say less than 5.

To obtain an unbiased estimator of the dispersion parameter  $\sigma^2$ , the second method utilizes a moment property given in the following proposition.

**Proposition 2.17.** *Let  $Y \sim DM(\mu, \sigma^2)$  with a regular unit deviance  $d(y; \mu)$ . Then,*

$$\begin{aligned} E\{\delta(Y; \mu)\} &= 0, \\ \text{Var}\{\delta(Y; \mu)\} &= \sigma^2 E\{-\dot{\delta}(Y; \mu)\}, \end{aligned}$$

where  $\dot{\delta}$  is the first order derivative of the deviance score given in (2.16) *w.r.t.*  $\mu$ .

*Proof.* Differentiating both sides of equation  $\int p(y; \mu, \sigma^2) dy = 1$  *w.r.t.*  $\mu$  gives

$$-\frac{1}{2\sigma^2} \int \dot{d}(y; \mu) p(y; \mu, \sigma^2) dy = 0,$$

or  $E\{\dot{d}(Y; \mu)\} = 0$ . Differentiating the above equation again *w.r.t.*  $\mu$ , we obtain

$$-\frac{1}{2\sigma^2} \int \{\dot{d}(y; \mu)\}^2 p(y; \mu, \sigma^2) dy + \int \ddot{d}(y; \mu) p(y; \mu, \sigma^2) dy = 0,$$

or equivalently

$$E\{\ddot{d}(Y; \mu)\} = \frac{1}{2\sigma^2} E\{\dot{d}(Y; \mu)\}^2 = \frac{1}{2\sigma^2} \text{Var}\{\dot{d}(Y; \mu)\}.$$

According to (2.16), this relation can be rewritten as follows,

$$\text{Var}\{\delta(Y; \mu)\} = \sigma^2 E\{-\dot{\delta}(Y; \mu)\}.$$

Based on this result, one can consistently estimate the dispersion parameter  $\sigma^2$  by the method of moments:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^K (\delta_i - \bar{\delta})^2}{\sum_{i=1}^K (-\dot{\delta}_i)}, \quad (2.22)$$

where  $\delta_i = \delta(y_i; \hat{\mu}_i)$ ,  $\dot{\delta}_i = \dot{\delta}(y_i; \hat{\mu}_i)$  and  $\bar{\delta} = \frac{1}{K} \sum_i \delta_i$ .

### 2.6.2 MLE in the ED Models

Now return to the special case of the GLMs based on the ED models. For the unit deviance of the ED model given in (2.11), it is easy to see

$$\delta(y; \mu) = \frac{y - \mu}{V(\mu)}. \quad (2.23)$$

It follows that the score equation (2.18) becomes

$$\sum_{i=1}^K \mathbf{x}_i \frac{1}{\dot{g}(\mu_i) V(\mu_i)} (y_i - \mu_i) = 0.$$

Let  $w_i = \dot{g}(\mu_i) V(\mu_i)$ . Then the score equation can be re-expressed as of the form

$$\sum_{i=1}^K \mathbf{x}_i w_i^{-1} (y_i - \mu_i) = 0,$$

or in the matrix notation,

$$\mathbf{X}^T W^{-1} (\mathbf{y} - \boldsymbol{\mu}) = 0,$$

where  $W = \text{diag}(w_1, \dots, w_K)$ . The following result is useful to calculate the Fisher information.

**Proposition 2.18.** *Suppose  $Y \sim ED(\mu, \sigma^2)$ . Then,*

$$E\{-\dot{\delta}(Y; \mu)\} = \frac{1}{V(\mu)},$$

where  $\dot{\delta}(y; \mu)$  is the first order derivative of the deviance score  $\delta(y; \mu)$  w.r.t.  $\mu$ .

*Proof.* Differentiating  $\delta$  in (2.16) w.r.t.  $\mu$  gives

$$-\dot{\delta}(y; \mu) = \frac{1}{V(\mu)} + \frac{(y - \mu) \dot{V}(\mu)}{V^2(\mu)},$$

which leads to

$$E\{-\dot{\delta}(Y; \mu)\} = \frac{1}{V(\mu)},$$

because  $E(Y) = \mu$  in the ED model.

In the Fisher information matrix  $\mathbf{i}(\boldsymbol{\theta})$  of (2.19),  $\mathbf{i}(\boldsymbol{\theta}) = \mathbf{X}^T U^{-1} \mathbf{X} / \sigma^2$ ,  $U$  is a diagonal matrix whose  $i$ -th diagonal element can be simplified as

$$u_i = \{\dot{g}(\mu_i)\}^2 V(\mu_i).$$

Furthermore, if the canonical link function  $g = \tau^{-1}(\cdot)$  is chosen, then a further simplification leads to  $w_i = 1$  and  $u_i = 1/V(\mu_i)$  because in this case,  $\dot{g}(\mu_i) = 1/V(\mu_i)$ . So, the matrix  $W$  becomes the identity matrix and the matrix  $U$  is determined by the reciprocals of the variance functions.

It is interesting to note that the choice of the canonical link simplifies both score function and Fisher information. In summary, under the canonical link function, the score equation of an ED GLM is

$$\sum_{i=1}^K \mathbf{x}_i (y_i - \mu_i) = 0, \text{ or } \mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu}) = 0,$$

and the Fisher information takes the form

$$\mathbf{i}(\boldsymbol{\theta}) = \mathbf{X}^T U^{-1} \mathbf{X} / \sigma^2$$

where  $U = \text{diag}(u_1, \dots, u_K)$ , a diagonal matrix with variance function  $V(\mu_i)$  as the  $i$ -th diagonal element.

Each ED model holds the so-called mean-variance relation, *i.e.*  $\text{Var}(Y) = \sigma^2 V(\mu)$ , which may be used to obtain a consistent estimator of the dispersion parameter  $\sigma^2$  given as follows:

$$\hat{\sigma}^2 = \frac{1}{K - p - 1} \sum_{i=1}^K \hat{r}_{p,i}^2 = \frac{1}{K - p - 1} \sum_{i=1}^K \left\{ \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right\}^2,$$

where  $\hat{r}_p$  is the Pearson residual listed in Table 2.3. This estimator is referred to as the Pearson estimator of the dispersion parameter  $\sigma^2$ . In fact, the relation given in Proposition 2.17 is equivalent to this mean-variance relation for the ED models, simply because of Proposition 2.18.

### 2.6.3 MLE in the Simplex GLM

The GLM for binary data or logistic regression model, the GLM for count data or log-linear regression model, and the GLM for positive continuous data or gamma regression model have been extensively illustrated in the literature. Interested readers can find examples of these ED GLMs easily in many references such as McCullagh and Nelder's (1989). This section supplies two non-ED GLMs based, respectively, on the simplex distribution and the von Mises distribution. Both are not available in the classical theory of GLMs.

In the ED GLMs, both score equation and Fisher information can be treated as a special case of weighted least squares estimation, due to the fact

that its first order derivative of the unit deviance is  $(y - \mu)/V(\mu)$ , which is linear in  $y$ . However, this linearity no longer holds for a DM GLM outside the class of the ED GLMs. The simplex distribution is one of such examples. A simplex model  $S^-(\mu; \sigma^2)$  has the density given by

$$p(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}d(y; \mu)\right\}, \quad y \in (0, 1), \mu \in (0, 1),$$

with the unit deviance function

$$d(y; \mu) = \frac{(y - \mu)^2}{y(1-y)\mu^2(1-\mu)^2}, \quad y \in (0, 1), \mu \in (0, 1),$$

where  $\mu = E(Y)$  is the mean. The unit variance function is  $V(\mu) = \mu^3(1-\mu)^3$ , obtained from (2.4).

For a non-ED GLM, the canonical link function no longer helps to simplify the weights  $u_i$  or the  $w_i$ , because the density does not explicitly involve the cumulant generating function  $\kappa(\cdot)$  as in the ED GLM. For the simplex distribution, since  $\mu \in (0, 1)$ , one may take the logit as the link function to formulate the systematic component:

$$\log \frac{\mu}{1-\mu} = \mathbf{x}^T \boldsymbol{\beta}.$$

According to Table 2.4,  $\dot{g}(\mu) = \{\mu(1-\mu)\}^{-1}$ . It follows from (2.18) that the score equation for the regression parameter  $\boldsymbol{\beta}$  is

$$\sum_{i=1}^K \mathbf{x}_i \{\mu_i(1-\mu_i)\} \delta(y_i; \mu_i) = 0, \quad (2.24)$$

where the deviance score is

$$\begin{aligned} \delta(y; \mu) &= -\frac{1}{2} \dot{d}(y; \mu) \\ &= \frac{y - \mu}{\mu(1-\mu)} \left\{ d(y; \mu) + \frac{1}{\mu^2(1-\mu)^2} \right\}. \end{aligned} \quad (2.25)$$

It is clear that this  $\delta$  function is nonlinear in both  $y$  and  $\mu$ . Solving nonlinear equation (2.24) can be done iteratively by the Newton-Raphson algorithm or quasi-Newton algorithm. The calculation of the Fisher information requires the knowledge of  $E\{-\ddot{\delta}(Y_i; \mu_i)\}$ . It is equivalent to deriving  $\frac{1}{2}E\ddot{d}(Y_i; \mu_i)$ .

Differentiating  $\dot{d}$  w.r.t.  $\mu$  gives

$$\begin{aligned} \frac{1}{2} \ddot{d}(y; \mu) &= \frac{1}{\mu(1-\mu)} d(y; \mu) + \frac{1-2\mu}{\mu^2(1-\mu)^2} (y-\mu) d(y; \mu) \\ &\quad + \frac{1}{\mu^3(1-\mu)^3} + \frac{1-2\mu}{\mu^4(1-\mu)^4} (y-\mu) \\ &\quad - \frac{1}{\mu(1-\mu)} (y-\mu) \dot{d}(y; \mu) - \frac{2(2\mu-1)}{\mu^4(1-\mu)^4} (y-\mu). \end{aligned} \quad (2.26)$$

Hence,

$$\begin{aligned}
\frac{1}{2}E\{\ddot{d}(Y; \mu)\} &= \frac{1}{\mu(1-\mu)} \left[ E\{d(Y; \mu)\} - E\{(Y - \mu)\dot{d}(Y; \mu)\} \right] \\
&\quad + \frac{1-2\mu}{\mu^2(1-\mu)^2} E\{(Y - \mu)d(Y; \mu)\} + \frac{1}{\mu^3(1-\mu)^3} \\
&= \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3},
\end{aligned} \tag{2.27}$$

where the last equation holds by applying part (e) of Proposition 2.19 below. Therefore, the Fisher information is

$$\mathbf{i}(\beta) = \frac{1}{\sigma^2} \sum_{i=1}^K \mathbf{x}_i u_i^{-1} \mathbf{x}_i^T,$$

where

$$u_i = \frac{\mu_i(1-\mu_i)}{1 + 3\sigma^2\{\mu_i(1-\mu_i)\}^2}, \quad i = 1, \dots, K.$$

As seen in (2.26), the first order derivative of the deviance score  $\dot{\delta}$  appears tedious, but its expectation in (2.27) is much simplified. Therefore, it is appealing to implement the Fisher-scoring algorithm in the search for the solution to the score equation (2.24). One complication in the application of Fisher-scoring algorithm is the involvement of the dispersion parameter  $\sigma^2$ . This can be resolved by replacing  $\sigma^2$  with a  $\sqrt{K}$ -consistent estimate,  $\hat{\sigma}^2$ . A consistent estimate of such a type can be obtained by the method of moments. For example, the property (a) in Proposition 2.19 is useful to establish an estimate of  $\sigma^2$  as follows:

$$\hat{\sigma}^2 = \frac{1}{K - (p+1)} \sum_{i=1}^K d(y_i; \hat{\mu}_i). \tag{2.28}$$

**Proposition 2.19.** *Suppose  $Y \sim S^-(\mu; \sigma^2)$  with mean  $\mu$  and dispersion  $\sigma^2$ . Then,*

- (a)  $E\{d(Y; \mu)\} = \sigma^2$ ;
- (b)  $E\{(Y - \mu)\dot{d}(Y; \mu)\} = -2\sigma^2$ ;
- (c)  $E\{(Y - \mu)d(Y; \mu)\} = 0$ ;
- (d)  $E\{\dot{d}(Y; \mu)\} = 0$ ;
- (e)  $\frac{1}{2}E\{\ddot{d}(Y; \mu)\} = \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}$ ;
- (f)  $\text{Var}\{d(Y; \mu)\} = 2(\sigma^2)^2$ ;
- (g)  $\text{Var}\{\delta(Y; \mu)\} = \frac{3\sigma^4}{\mu(1-\mu)} + \frac{\sigma^2}{\mu^3(1-\mu)^3}$ .

The following lemma is needed in order to prove Proposition 2.19.

**Lemma 2.20 (Jørgensen, 1997, P.191).** *Consider a dispersion model  $DM(\mu, \sigma^2)$  whose density takes the form:*

$$f(y; \mu, \lambda) = c_\alpha(\mu, \lambda) y^{\alpha-1} \exp \left\{ -\frac{\lambda(y - \mu)^2}{2y\mu^{1-2\alpha}} \right\},$$

where  $\lambda = 1/\sigma^2$  and the normalization constant is defined by

$$\frac{1}{c_\alpha(\mu, \lambda)} = 2K_\alpha(\lambda\mu^{2\alpha}) e^{\lambda\mu^{2\alpha}} \mu^\alpha.$$

Then the asymptotic expansion of  $1/c_\alpha(\mu, \lambda)$  is given by, for large  $\lambda$ ,

$$\left\{ \frac{2\pi}{\lambda} \right\}^{\frac{1}{2}} \left\{ 1 + \frac{4\alpha^2 - 1}{8\lambda\mu^{2\alpha}} + \frac{(4\alpha^2 - 1)(4\alpha^2 - 9)}{2!(8\lambda\mu^{2\alpha})^2} + \frac{(4\alpha^2 - 1)(4\alpha^2 - 9)(4\alpha^2 - 25)}{3!(8\lambda\mu^{2\alpha})^3} + \dots \right\}.$$

The proof of Proposition 2.19 is given as follows.

*Proof.* First prove part (b). Note that

$$0 = E[(Y - \mu)] = \int_0^1 (y - \mu) p(y; \mu, \sigma^2) dy,$$

and differentiating both sides of the equation with respect to  $\mu$  gives

$$0 = -1 - \frac{1}{2\sigma^2} E[(Y - \mu) \dot{d}(Y; \mu)],$$

and hence  $E[(Y - \mu) \dot{d}(Y; \mu)] = -2\sigma^2$ .

To prove part (a) and part (c), take the following transformations for both  $y$  and  $\mu$ ,

$$x = \frac{y}{1 - y}, \quad \xi = \frac{\mu}{1 - \mu}$$

and rewrite the two expectations in the following forms:

$$\begin{aligned} E[d(Y; \mu)] &= \int_0^1 d(y; \mu) p(y; \mu, \sigma^2) dy \\ &= \sqrt{\frac{\lambda}{2\pi}} \frac{(1 + \xi)^2}{\xi^2} \int_0^\infty \left\{ x^{\frac{1}{2}} + (1 - 2\xi)x^{-\frac{1}{2}} \right. \\ &\quad \left. + \xi(\xi - 2)x^{-\frac{3}{2}} + \xi^2 x^{-\frac{5}{2}} \right\} f(x; \xi, \lambda) dx, \end{aligned}$$

and

$$\begin{aligned}
E[(Y - \mu)d(Y; \mu)] &= \int_0^1 (y - \mu)d(y; \mu)p(y; \mu, \sigma^2)dy \\
&= \sqrt{\frac{\lambda}{2\pi}} \frac{1 + \xi}{\xi^2} \int_0^\infty \left\{ x^{\frac{1}{2}} - 3\xi x^{-\frac{1}{2}} \right. \\
&\quad \left. + 3\xi^2 x^{-\frac{3}{2}} - \xi^3 x^{-\frac{5}{2}} \right\} f(x; \xi, \lambda) dx,
\end{aligned}$$

where  $\lambda = 1/\sigma^2$  and

$$f(x; \xi, \lambda) = \exp \left\{ -\frac{\lambda}{2} \frac{(1 + \xi)^2}{\xi^2} \frac{(x - \xi)^2}{x} \right\}.$$

Applying Lemma 2.20 leads to

$$\begin{aligned}
\int_0^\infty x^{\frac{1}{2}} f(x; \xi, \lambda) dx &= \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{\xi^3 + \lambda \xi^2 (1 + \xi)^2}{\lambda (1 + \xi)^3}, \\
\int_0^\infty x^{-\frac{1}{2}} f(x; \xi, \lambda) dx &= \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{\xi}{1 + \xi}, \\
\int_0^\infty x^{-\frac{3}{2}} f(x; \xi, \lambda) dx &= \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{1}{1 + \xi},
\end{aligned}$$

and

$$\int_0^\infty x^{-\frac{5}{2}} f(x; \xi, \lambda) dx = \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{\xi + \lambda(1 + \xi)^2}{\lambda \xi (1 + \xi)^3}.$$

Plugging these results into the expressions above leads to

$$E\{d(Y; \mu)\} = 1/\lambda = \sigma^2 \quad \text{and} \quad E\{(Y - \mu)d(Y; \mu)\} = 0.$$

Part (d) is given by applying part (c) to (2.25) and then taking expectation. Also, part (e) is proved by applying parts (a), (b), and (c) to (2.27).

By part (a), to prove part (f), it is sufficient to show that

$$E\{d^2(Y; \mu)\} = 3(\sigma^2)^2.$$

Simple algebra leads to

$$\begin{aligned}
E\{d^2(Y; \mu)\} &= \int_0^1 d^2(y; \mu)p(y; \mu, \sigma^2)dy \\
&= \sqrt{\frac{\lambda}{2\pi}} \frac{(1 + \xi)^4}{\xi^4} \int_0^\infty \left\{ x^{\frac{3}{2}} + (1 - 4\xi)x^{\frac{1}{2}} \right. \\
&\quad \left. + 2\xi(3\xi - 2)x^{-\frac{1}{2}} + 2\xi^2(3 - 2\xi)x^{-\frac{3}{2}} \right. \\
&\quad \left. + \xi^3(\xi - 4)x^{-\frac{5}{2}} + \xi^4 x^{-\frac{7}{2}} \right\} f(x; \xi, \lambda) dx. \tag{2.29}
\end{aligned}$$



An application of Lemma 2.20 again results in

$$\int_0^\infty x^{\frac{3}{2}} f(x; \xi, \lambda) dx = \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{\lambda^2 \xi^3 (1 + \xi)^4 + 3\lambda \xi^4 (1 + \xi)^2 + 3\xi^5}{\lambda^2 (1 + \xi)^5}$$

and

$$\int_0^\infty x^{-\frac{7}{2}} f(x; \xi, \lambda) dx = \left( \frac{2\pi}{\lambda} \right)^{\frac{1}{2}} \frac{\lambda^2 (1 + \xi)^4 + 3\lambda \xi (1 + \xi)^2 + 3\xi^2}{\lambda^2 \xi^2 (1 + \xi)^5}.$$

Based on these results, the integration (2.29) can be simplified as

$$E\{d^2(Y; \mu)\} = 3(\sigma^2)^2.$$

Part (g) can be proved by applying part (e) in the relation between  $\hat{d}$  and  $\delta$  from Proposition 2.17.

In the application of the simplex GLM, one issue that deserves some attention is whether there is much difference between the normal linear model based on logit-transformed data,  $\log\{y_i/(1 - y_i)\}$ , and the direct simplex GLM. The difference between the two models is the former models  $E[\log\{Y_i/(1 - Y_i)\}]$  as a linear function of covariates, and the latter models  $\mu_i = E(Y_i)$  via  $\log\{\mu_i/(1 - \mu_i)\}$  as a linear function of covariates. Apparently the direct GLM approach gives rise to much ease in interpretation.

The following simulation study suggests that when the dispersion parameter  $\sigma^2$  is large, the performance of the logit-transformed analysis may be questionable, if the data are really from a simplex distributed population.

The simulation study assumes the proportional data are generated independently from the following simplex distribution,

$$Y_i \sim S^-(\mu_i, \sigma^2), \quad i = 1, \dots, 150,$$

where the mean follows a GLM of the following form:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i.$$

Covariates  $T$  and  $S$  are presumably drug dosage levels indicated by  $\{-1, 0, 1\}$  for each 50 subjects and illness severity score ranged in  $\{0, 1, 2, 3, 4, 5, 6\}$  that is randomly assumed to each subject by a binomial distribution  $B(7, 0.5)$ . The true values of regression coefficients are set as  $\beta_0 = 0.5, \beta_1 = -0.5, \beta_2 = 0.5$ , and the dispersion parameter  $\sigma^2 = 0.5, 50, 200, 400$ .

For each combination of parameters, the same simulated data was fit by the simplex GLM for the original responses and the normal linear model for logit-transformed responses. Two hundred replications were done for each case. Results are summarized in Table 2.5, including the averaged estimates, standard deviations of 200 replicated estimates, and standard errors of estimates calculated from the Fisher information.

**Table 2.5.** Summary of the simulation results for the comparison between the direct simplex GLM analysis and logit-transformed linear model analysis.

Parameter	Simplex GLM			Logit-Trans LM		
True	Mean	Std Dev	Std Err	Mean	Std Dev	Std Err
$\sigma^2 = 0.5$						
$\beta_0(0.5)$	.4996	.0280	.0254	.5089	.0288	.0263
$\beta_1(-0.5)$	-.5023	.0330	.0308	-.5110	.0345	.0322
$\beta_2(0.5)$	.5015	.0195	.0205	.5101	.0199	.0222
$\sigma^2 = 50$						
$\beta_0(0.5)$	.5062	.0983	.0960	.8057	.1769	.1752
$\beta_1(-0.5)$	-.5068	.1141	.1185	-.7998	.2065	.2148
$\beta_2(0.5)$	.5170	.0860	.0835	.8153	.1366	.1483
$\sigma^2 = 200$						
$\beta_0(0.5)$	.5060	.1145	.1021	1.0162	.2741	.2541
$\beta_1(-0.5)$	-.5262	.1346	.1263	-1.0479	.3218	.3114
$\beta_2(0.5)$	.5238	.0971	.0899	1.0430	.1919	.2150
$\sigma^2 = 400$						
$\beta_0(0.5)$	.5253	.0963	.1032	1.2306	.2767	.2980
$\beta_1(-0.5)$	-.5001	.1486	.1275	-1.1336	.3888	.3652
$\beta_2(0.5)$	.5165	.1000	.0909	1.1686	.2286	.2521

This simulation study indicates that (i) when the dispersion parameter  $\sigma^2$  is small, the logit-transformed analysis appears fine, with little bias and little loss of efficiency, because of small-dispersion asymptotic normality; (ii) when the dispersion parameter is large, the estimation based on the logit-transformed analysis is unacceptable, in which bias increases and efficiency drops when the  $\sigma^2$  increases.

One may try to make a similar comparison by simulating data from the normal distribution as well as from the beta distribution. Our simulation study suggested that in the case of normal data, the direct simplex GLM performed nearly as well as the normal model, with only a marginal loss of efficiency; in the case of beta data, the simplex GLM clearly outperformed the normal linear model. Interested readers can verify the findings through their own simulation studies.

*Example 2.21 (Body Fat Index).*

Penrose et al. (1985) reports a dataset consisting of 19 variables, including percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) for 252 men. This dataset is available at [http://www.amstat.org/publications/jse/jse\\_data\\_archive.html](http://www.amstat.org/publications/jse/jse_data_archive.html). Body fat, a measure of health, is estimated through an underwater weighing technique. Percentage of body fat may be then calculated by either Brozek's equation or Siri's equation. Fitting body fat to the other measurements using GLM provides a convenient way of estimating body fat for men using only a scale and a measuring tape.

In this example, the simplex GLM is illustrated simply by fitting the the body fat index as a function of covariate **age**. Suppose the percentage of body fat  $Y_i \sim S^-(\mu_i, \sigma^2)$ , where

$$\log \frac{\mu_i}{1 - \mu_i} = \beta_0 + \beta_1 \text{age}.$$

The Fisher-scoring algorithm was applied to obtain the estimates of the regression coefficients and the standard errors were calculated from the Fisher information. The results were summarized in Table 2.6, in which the dispersion parameter is estimated by the method of moments in (2.28). Clearly, from the results given in Table 2.6, age is an important predictor to the percentage of body fat in both Brozek's and Siri's equations. The dispersion  $\sigma^2$  is found not small in this study, so it might be worrisome for the appropriateness of either a direct linear model analysis (with no transformation on the response) or logit-transformed linear model analysis. Some further investigations are needed to elucidate the choice of modeling approach in this data analysis.

**Table 2.6.** Results in the regression analysis of body fat percentage using the simplex GLM.

Parameter			
Body-fat measure	Intercept (Std Err)	Age (Std Err)	$\sigma^2$
Brozek's	-2.7929(0.3304)	0.0193(0.0070)	55.9759
Siri's	-2.8258(0.3309)	0.0202(0.0070)	57.0353

### 2.6.4 MLE in the von Mises GLM

Angular data are a special case of circular data. Mardia (1972) has presented a general framework of estimation and inference in the models for circular

data. Fisher (1993) gave an overview of the state-of-art of research in this field. Although the analysis of circular data is an old topic, there have been recent developments in applied areas, such as multi-dimensional circular data (Fisher, 1993; Rivest, 1997; Breckling, 1989), time series of circular observations (Accardi et al. 1987; Fisher and Lee 1994; Coles 1998), and longitudinal circular data (Artes et al. 2000; D’Elia et al. 2001).

The von Mises distribution is another example of the DM model but not of an ED model. The density of a von Mises distribution takes the form

$$p(y; \mu, \sigma^2) = \frac{1}{2\pi I_0(\lambda)} \exp\{\lambda \cos(y - \mu)\}, \quad y \in [-\pi, \pi), \quad (2.30)$$

where  $\mu \in [-\pi, \pi)$  is the mean,  $\lambda = 1/\sigma^2 > 0$  is the index parameter, and  $I_0(\lambda)$  is the modified Bessel function of the first kind of order 0, given by

$$I_0(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\{\lambda \cos(y)\} dy.$$

It is easy to rewrite the von Mises density in the form of DM model with the unit deviance function given by

$$d(y; \mu) = 2\{1 - \cos(y - \mu)\}, \quad y, \mu \in [-\pi, \pi),$$

whose first and second order derivatives *w.r.t.*  $\mu$  are, respectively,

$$\dot{d} = -2 \sin(y - \mu), \quad \ddot{d} = 2 \cos(y - \mu).$$

It follows that the unit variance function is  $V(\mu) = 1$  for  $\mu \in [-\pi, \pi)$  and the deviance score  $\delta(y; \mu) = \sin(y - \mu)$ .

Now consider a GLM for directional (circular or angular) data, where  $Y_i \sim vM(\mu_i, \sigma^2)$ , associated with  $p$ -element vector of covariates  $\mathbf{x}_i$ . According to Fisher and Lee (1992) or Fisher (1993, Section 6.4), a GLM for the mean direction  $\mu_i = E(Y_i|\mathbf{x}_i)$  may be formulated as follows:

$$\mu_i = \mu_0 + 2\arctan(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p), \quad (2.31)$$

where  $\mu_0$  is an offset mean parameter representing the origin. If  $Y_i^* = Y_i - \mu_0$  is taken as a surrogate response, then the corresponding mean direction is  $\mu_i^* = \mu_i - \mu_0 = 2\arctan(\eta_i)$  with the origin of  $0^\circ$ . This implies

$$\tan(\mu_i^*/2) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta},$$

where the intercept term is not included, because of the  $0^\circ$  origin. Clearly, in this GLM, the link function  $g(z) = \tan(z/2)$  and  $\dot{g}(z) = \frac{1}{2}\sec^2(z/2)$ , as shown in Table 2.4. To estimate the regression parameter  $\boldsymbol{\beta}$ , formula (2.18) is applied here to yield the following score equation:

$$\begin{aligned}
s(\mathbf{y}; \boldsymbol{\beta}) &= \lambda \sum_{i=1}^K \mathbf{x}_i \frac{1}{\hat{g}(\mu_i^*)} \delta(y_i^*; \mu_i^*) \\
&= 2\lambda \sum_{i=1}^K \mathbf{x}_i \left( \frac{1}{1 + \eta_i^2} \right) \sin(y_i^* - \mu_i^*) \\
&= 2\lambda \sum_{i=1}^K \mathbf{x}_i \left( \frac{1}{1 + \eta_i^2} \right) \sin(y_i - \mu_0 - 2\arctan(\mathbf{x}_i^T \boldsymbol{\beta})),
\end{aligned}$$

where the identity of  $\sec^2(\arctan(a)) = 1 + a^2$  is used. The MLE of  $\boldsymbol{\beta}$  is the solution to the score equation

$$s(\mathbf{y}; \boldsymbol{\beta}) = 0. \quad (2.32)$$

To find the Fisher Information for  $\hat{\boldsymbol{\beta}}$ , first note that the surrogate response  $Y_i^* \sim vM(\mu_i^*, \sigma^2)$ , and then

$$\mathbb{E}\{-\dot{\delta}(Y_i^*; \mu_i^*)\} = \mathbb{E}\{\cos(Y_i^* - \mu_i^*)\} = \frac{I_1(\lambda)}{I_0(\lambda)},$$

where  $I_1(\lambda)$  is the first order modified Bessel function of the first kind given by

$$I_1(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(y) \exp\{\lambda \cos(y)\} dy.$$

Denote the mean resultant length by  $A_1(\sigma^2) = I_1(\sigma^{-2})/I_0(\sigma^{-2})$ . Then the weights  $u_i$  in (2.20) are found as

$$u_i = \frac{(1 + \eta_i^2)^2}{4A_1(\sigma^2)}, \quad i = 1, \dots, K.$$

Moreover, the Fisher Information for the  $\hat{\boldsymbol{\beta}}$  is  $\mathbf{i}(\boldsymbol{\beta}) = \mathbf{X}^T U^{-1} \mathbf{X} / \sigma^2$ , with  $U = \text{diag}(u_1, \dots, u_K)$ .

To estimate the parameter  $\mu_0$  and the dispersion parameter  $\sigma^2$ , the MLE may be also employed. The log likelihood is proportional to

$$\ell(\boldsymbol{\theta}) \propto -K \log I_0(\lambda) + \lambda \sum_{i=1}^K \cos(y_i - \mu_0 - 2\arctan(\eta_i)),$$

and the scores for  $\mu_0$  and  $\lambda$  are, respectively,

$$\begin{aligned}
s(\mathbf{y}; \mu_0) &= \lambda \sum_{i=1}^K \sin(y_i - \mu_0 - 2\arctan(\eta_i)), \\
s(\mathbf{y}; \lambda) &= -K \frac{\dot{I}_0(\lambda)}{I_0(\lambda)} + \sum_{i=1}^K \cos(y_i - \mu_0 - 2\arctan(\eta_i)).
\end{aligned}$$

Also,

$$\begin{aligned}
-E\{\dot{s}_{\mu_0}(\mathbf{y}; \mu_0)\} &= K\lambda A_1(\lambda), \\
-E\{\dot{s}_{\boldsymbol{\beta}}(\mathbf{y}; \mu_0)\} &= 4\lambda A_1(\lambda) \sum_{i=1}^K \mathbf{x}_i \frac{1}{(1 + \eta_i^2)^2}, \\
-E\{\dot{s}_{\lambda}(\mathbf{y}; \lambda)\} &= K \left\{ \frac{\dot{I}_1(\lambda)}{I_0(\lambda)} - A_1^2(\lambda) \right\}, \\
-E\{\dot{s}_{\boldsymbol{\beta}}(\mathbf{y}; \lambda)\} &= 0.
\end{aligned}$$

It is easy to show that  $\dot{I}_0(\lambda) = I_1(\lambda)$  and  $\dot{I}_1(\lambda) = \frac{1}{2}\{I_1(\lambda) + I_0(\lambda)\}$ . Let  $\hat{\mu}_0$  and  $\hat{\lambda}$  be the MLE. Then  $(\hat{\mu}_0, \hat{\boldsymbol{\beta}}, \hat{\lambda})$  will be the solution to the following joint score equations:

$$\begin{pmatrix} s(\mathbf{y}; \mu_0) \\ s(\mathbf{y}; \boldsymbol{\beta}) \\ s(\mathbf{y}; \lambda) \end{pmatrix} = \begin{pmatrix} \lambda \sum_{i=1}^K \text{diag}[1, \mathbf{x}_i][1, 2(1 + \eta_i^2)^{-1}]^T \sin(y_i - \mu_0 - 2\arctan(\eta_i)) \\ -K A_1(\lambda) + \sum_{i=1}^K \cos(y_i - \mu_0 - 2\arctan(\eta_i)) \\ 0 \end{pmatrix} = \mathbf{0}. \quad (2.33)$$

The corresponding Fisher information matrix is

$$\mathbf{i}(\mu_0, \boldsymbol{\beta}, \lambda) = \begin{pmatrix} \lambda \sum_{i=1}^K \text{diag}[1, \mathbf{x}_i][1, u_i^{-1}]^T [1, u_i^{-1}] \text{diag}^T[1, \mathbf{x}_i] & 0 \\ 0 & K \left\{ \frac{1}{2}(A_1(\lambda) + 1) - A_1^2(\lambda) \right\} \end{pmatrix}.$$

One may use the iterative Fisher-scoring algorithm to solve jointly the score equation (2.33) for the MLE, which involves inverting the above Fisher information matrix at current values of the parameters. Alternatively, one may solve equations (2.32), and the following (2.34) and (2.35) in cycle,

$$\hat{\mu}_0 = \arctan(\bar{S}/\bar{C}) \quad (2.34)$$

$$\hat{\lambda} = A_1^{-1} \left\{ \frac{1}{K} \sum_{i=1}^K \cos(y_i - \mu_0 - 2\arctan(\eta_i)) \right\}, \quad (2.35)$$

where  $A_1^{-1}\{\cdot\}$  is the inverse function of  $A_1$ , and

$$\begin{aligned}
\bar{S} &= \frac{1}{K} \sum_{i=1}^K \sin(y_i - 2\arctan(\eta_i)), \\
\bar{C} &= \frac{1}{K} \sum_{i=1}^K \cos(y_i - 2\arctan(\eta_i)).
\end{aligned}$$

It is known in the literature that when the sample size  $K$  is small, the MLE of  $\sigma^2$  or  $\lambda$  appears to have some noticeable bias. Alternatively, one may

use the moment property,  $\text{Var}(Y) = 1 - A_1(\lambda)$ , to obtain a consistent moment estimator,

$$\hat{\lambda}_{\text{mom}} = A_1^{-1} \left\{ 1 - \frac{1}{K - p - 1} \sum_{i=1}^K (y_i - \hat{\mu}_i)^2 \right\}. \quad (2.36)$$

An R package **CircStats** provides functions to plot circular data (e.g., function `circ.plot`) and compute many quantities given above, such as  $I_0(\lambda)$ ,  $I_1(\lambda)$ , and even  $I_p(\lambda)$  for any integer  $p$ . In this package, another useful function is `circ.kappa`, which provides a bias correction for the MLE estimation for the index parameter  $\lambda = 1/\sigma^2$ . Interested readers can follow Problem 2.5 in Problem Set 2 (available at the book webpage) to gain some numerical experience with the analysis of circular data.



<http://www.springer.com/978-0-387-71392-2>

Correlated Data Analysis: Modeling, Analytics, and  
Applications

Song, P.X.-K.

2007, XVI, 352 p., Hardcover

ISBN: 978-0-387-71392-2