2 MAXIMUM LIKELIHOOD ESTIMATION PROCEDURES

This chapter outlines the theory of the Newton-Raphson, Fisher-Scoring and EM algorithms as procedures for maximum likelihood estimation. The EM algorithm is specifically applied to the exponential family to determine ML estimates for incomplete data when the missing data mechanism is ignorable. A maximum likelihood estimation procedure for the mean of the exponential family, subject to the constraint $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$, is also discussed.

2.1 THE NEWTON-RAPHSON ALGORITHM

The Newton-Raphson method is an iterative procedure to determine the value $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ that maximizes a function $g(\boldsymbol{\beta})$.
Let $\boldsymbol{\beta}^{(r)}$ be the rth approximation of $\widehat{\boldsymbol{\beta}}$ where $r = 0, 1, 2, \ldots$. As described in Agresti (1990), the method requires an initial guess, $\boldsymbol{\beta}^{(0)}$, for the value that maximizes the function. At step r in the iterative process the function $g(\boldsymbol{\beta})$ is approximated by the terms up to the second order in the Taylor series expansion of $g(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^{(r)}$, that is

$$Q^{(r)}(\boldsymbol{\beta}) = g\left(\boldsymbol{\beta}^{(r)}\right) + \mathbf{q}^{(r)\prime}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\right) + \tfrac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\right)' \mathbf{H}^{(r)}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\right) + o\left(\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\right\|\right) \tag{2}$$

where $\mathbf{H}$ is the matrix having elements $\dfrac{\partial^2 g(\boldsymbol{\beta})}{\partial\beta_h \partial\beta_k}$, $\mathbf{q}$ is the vector having elements $\dfrac{\partial g(\boldsymbol{\beta})}{\partial\beta_k}$, and $\mathbf{H}^{(r)}$ and $\mathbf{q}^{(r)}$ are $\mathbf{H}$ and $\mathbf{q}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r)}$.
The next approximation of $\widehat{\boldsymbol{\beta}}$ is in the location of the maximum value of (2).
Solving $\dfrac{\partial Q^{(r)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \mathbf{q}^{(r)} + \mathbf{H}^{(r)}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\right) = \mathbf{0}$ for $\boldsymbol{\beta}$ yields the next approximation of $\widehat{\boldsymbol{\beta}}$,

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left(\mathbf{H}^{(r)}\right)^{-1}\mathbf{q}^{(r)} \tag{3}$$

assuming $\mathbf{H}^{(r)}$ is nonsingular.
Iteration continues until convergence is attained.

EXAMPLE 2.1
Determining ML estimates using the Newton-Raphson algorithm.

The number of accidents per thousand per age group in a certain factory is given in Table 2.1.

TABLE 2.1: Accidents per 1000 per age group.

Age group	I	II	III
Number of accidents	80	15	5

Suppose the elements of $\mathbf{Y} : 3 \times 1$, the number of accidents for each category, are independent Poisson random variables with parameter vector $\boldsymbol{\mu}$ The observed vector is $\mathbf{y}' = (80, 15, 5)$. The model under consideration is $\mu_i = \alpha\gamma^{i-1}$ for $i = 1, 2, 3$. The likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\mu}|\mathbf{y}) &= \frac{\exp\left(-\sum\mu_i\right)\prod\mu_i^{y_i}}{\prod y_i!} \\ &= \frac{\exp(-\alpha)\left(1+\gamma+\gamma^2\right)\alpha^{(y_1+y_2+y_3)}\gamma^{(y_2+2y_3)}}{\prod y_i!}. \end{aligned}$$

The value, $\widehat{\boldsymbol{\beta}}' = (\widehat{\alpha}, \widehat{\gamma})$, that maximizes l will also maximize the log-likelihood function

$$L(\boldsymbol{\beta}|\mathbf{y}) = (-\alpha)\left(1+\gamma+\gamma^2\right) + (y_1+y_2+y_3)\log(\alpha) + (y_2+2y_3)\log(\gamma) - \sum\log(y_i!)$$

and is determined iteratively with the expression

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left(\mathbf{H}^{(r)}\right)^{-1} \mathbf{q}^{(r)} \tag{4}$$

where $\boldsymbol{\beta}^{(r)}$ is the rth approximation of $\widehat{\boldsymbol{\beta}}$, and $\mathbf{q}^{(r)}$ and $\mathbf{H}^{(r)}$ are $\mathbf{q}$ and $\mathbf{H}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r)}$ with

$$\mathbf{q} = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \alpha} \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} -\left(1+\gamma+\gamma^2\right) + \frac{y_1+y_2+y_3}{\alpha} \\ -\alpha(1+2\gamma) + \frac{y_2+2y_3}{\gamma} \end{pmatrix} \tag{5}$$

$$\mathbf{H} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \begin{pmatrix} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \alpha^2} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \alpha \partial \gamma} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \gamma \partial \alpha} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \gamma^2} \end{pmatrix} = \begin{pmatrix} -\frac{(y_1+y_2+y_3)}{\alpha^2} & -(1+2\gamma) \\ -(1+2\gamma) & -2\alpha - \frac{(y_2+2y_3)}{\gamma^2} \end{pmatrix}. \tag{6}$$

From the model to be fitted $\alpha = \mu_1$ and $\gamma = \dfrac{\mu_2}{\alpha} = \dfrac{\mu_2}{\mu_1}$. If the observed data is used as an initial estimate of $\boldsymbol{\mu}$ the first approximation of $\widehat{\boldsymbol{\beta}}$ is

$$\boldsymbol{\beta}^{(0)} = \begin{pmatrix} \alpha^{(0)} \\ \gamma^{(0)} \end{pmatrix} = \begin{pmatrix} 80 \\ 0.1875 \end{pmatrix}$$

and is used to determine $\mathbf{q}^{(0)}$ and $\mathbf{H}^{(0)}$. Substituting $\boldsymbol{\beta}^{(0)}$, $\mathbf{q}^{(0)}$ and $\mathbf{H}^{(0)}$ into (4) gives

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \left(\mathbf{H}^{(0)}\right)^{-1} \mathbf{q}^{(0)}.$$

This is used to determine $\mathbf{q}^{(1)}$ and $\mathbf{H}^{(1)}$.
The process continues until convergence is attained. Table 2.2 shows $\boldsymbol{\beta}^{(r)}$ at different steps of the algorithm.

TABLE 2.2: Values of $\boldsymbol{\beta}^{(r)}$ at different steps of the Newton-Raphson algorithm.

r	$\alpha^{(r)}$	$\gamma^{(r)}$
0	80	0.1875
1	79.294919	0.2153986
2	78.829748	0.2200938
3	78.821827	0.2201973
4	78.821823	0.2201973

The value $\widehat{\boldsymbol{\beta}}$ that maximizes the log-likelihood function is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\gamma} \end{pmatrix} = \begin{pmatrix} 78.821823 \\ 0.2201973 \end{pmatrix}.$$

Substituting this into the model to be fitted, $\mu_i = \alpha\gamma^{i-1}$, gives

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \\ \widehat{\mu}_3 \end{pmatrix} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\alpha}\widehat{\gamma} \\ \widehat{\alpha}\widehat{\gamma}^2 \end{pmatrix} = \begin{pmatrix} 78.821823 \\ 17.356354 \\ 3.8218228 \end{pmatrix}.$$

The program is given in the Appendix.

EXAMPLE 2.2
Determining ML estimates for a loglinear model using the Newton-Raphson algorithm.

Consider the model in Example 1.2 and Example 2.1. The log-likelihood function is

$$L(\boldsymbol{\mu}|\mathbf{y}) = \sum_i y_i \log \mu_i - \sum_i \mu_i - \sum_i \log y_i!. \tag{7}$$

In Example 1.2 the model $\mu_i = \alpha\gamma^{i-1}$ was written as the generalized linear model

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

with $\beta_1 = \log\alpha$ and $\beta_2 = \log\gamma$, and $\mathbf{X}$ the design matrix.
Using the fact that $\log\mu_i = \sum_j \beta_j x_{ij}$ and $\mu_i = \exp\left(\sum_j \beta_j x_{ij}\right)$ the log-likelihood function in (7) can be written as a function of the elements of $\boldsymbol{\beta}$. That is

$$L(\boldsymbol{\beta}|\mathbf{y}) = \sum_i y_i \sum_j \beta_j x_{ij} - \sum_i \exp\left(\sum_j \beta_j x_{ij}\right) - \sum_i \log y_i!. \tag{8}$$

The value of $\widehat{\boldsymbol{\beta}}$ that maximizes $L(\boldsymbol{\beta}|\mathbf{y})$ can be found iteratively with

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left(\mathbf{H}^{(r)}\right)^{-1}\mathbf{q}^{(r)} \tag{9}$$

where $\mathbf{q}$ is the vector with elements the first order partial derivatives

$$q_k = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k} = -\sum_i x_{ik} \exp\left(\sum_j \beta_j x_{ij}\right) + \sum_i y_i x_{ik}$$

and $\mathbf{H}$ is the matrix of second order partial derivatives having elements

$$h_{hk} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h \partial\beta_k} = -\sum_i x_{ih}x_{ik} \exp\left(\sum_j \beta_j x_{ij}\right) = -\sum_i x_{ih}x_{ik}\mu_i.$$

From this

$$\mathbf{q}^{(r)} = \mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right) \tag{10}$$

$$\mathbf{H}^{(r)} = -\mathbf{X}'\text{diag}\left(\boldsymbol{\mu}^{(r)}\right)\mathbf{X} \tag{11}$$

with $\boldsymbol{\mu}^{(r)} = \exp\left(\mathbf{X}\boldsymbol{\beta}^{(r)}\right)$ the rth approximation of $\widehat{\boldsymbol{\mu}}$, $(r = 0, 1, 2, \ldots)$.
Substituting (10) and (11) into (9) gives

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left[\mathbf{X}'\text{diag}\left(\boldsymbol{\mu}^{(r)}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right). \tag{12}$$

From the model to be fitted $\alpha = \mu_1$ and $\gamma = \dfrac{\mu_2}{\alpha} = \dfrac{\mu_2}{\mu_1}$. Using the observed data as an initial estimate of $\boldsymbol{\mu}$, the approximation of $\widehat{\boldsymbol{\beta}}$ at $r = 0$ is

$$\boldsymbol{\beta}^{(0)} = \begin{pmatrix} \log\alpha^{(0)} \\ \log\gamma^{(0)} \end{pmatrix} = \begin{pmatrix} 1.90309 \\ -0.72700 \end{pmatrix}.$$

This is used to determine $\boldsymbol{\mu}^{(0)} = \exp\left(\mathbf{X}\boldsymbol{\beta}^{(0)}\right)$. Substituting $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\mu}^{(0)}$ in (12) gives the next approximation for $\widehat{\boldsymbol{\beta}}$,

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left[\mathbf{X}'\text{diag}\left(\boldsymbol{\mu}^{(0)}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(0)}\right)$$

which is used to determine $\boldsymbol{\mu}^{(1)}$.
The process continues until convergence is attained and the value $\widehat{\boldsymbol{\beta}}$ that maximizes the log-likelihood function in (8) is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \log\widehat{\alpha} \\ \log\widehat{\gamma} \end{pmatrix} = \begin{pmatrix} 4.3671899 \\ -1.513231 \end{pmatrix}.$$

Substituting this into the model, $\mu_i = \alpha\gamma^{i-1}$, gives

$$\widehat{\boldsymbol{\mu}} = \exp\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \begin{pmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \\ \widehat{\mu}_3 \end{pmatrix} = \begin{pmatrix} 78.821823 \\ 17.356354 \\ 3.8218228 \end{pmatrix}$$

This is the same result as obtained in Example 2.1.
The program is given in the Appendix.

2.2 THE FISHER SCORING ALGORITHM

The Fisher scoring algorithm is similar to the Newton-Raphson algorithm, the distinction being that Fisher scoring uses the information matrix. The information matrix is the negative expected value of the second order derivitave matrix of the function to be maximized. The Newton-Raphson algorithm uses the observed value of the second order derivitave matrix. The formula for Fisher scoring is

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left(\mathbf{Inf}^{(r)}\right)^{-1} \mathbf{q}^{(r)}$$

where $\mathbf{Inf}^{(r)}$ is the rth approximation for the estimated information matrix. The information matrix, $\mathbf{Inf}$, is the negative expected value of the matrix of second order partial derivatives of the log-likelihood and has elements $\text{Inf}_{hk} = -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_k}\right)$.

EXAMPLE 2.3
Determining ML estimates using the Fisher scoring algorithm.

Suppose the elements of $\mathbf{Y} : 3 \times 1$ are independent Poisson random variables with parameter vector $\boldsymbol{\mu}$ and observed vector $\mathbf{y}' = (80, 15, 5)$. The model to be fitted is $\mu_i = \alpha\gamma^{i-1}$. In Example 2.1 the Newton-Raphson algorithm was used to find the ML estimates.
The equation used in the iterative procedure is

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left(\mathbf{Inf}^{(r)}\right)^{-1} \mathbf{q}^{(r)}$$

where $\mathbf{Inf}^{(r)}$ is

$$\mathbf{Inf} = -E\left[\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right] = \begin{pmatrix} -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\alpha^2}\right) & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\alpha\partial\gamma}\right) \\ -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\gamma\partial\alpha}\right) & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\gamma^2}\right) \end{pmatrix} = \begin{pmatrix} E\left(\frac{(y_1+y_2+y_3)}{\alpha}\right) & -(1+2\gamma) \\ -(1+2\gamma) & E\left(2\alpha + \frac{(y_2+2y_3)}{\gamma^2}\right) \end{pmatrix}$$

evaluated at $\boldsymbol{\beta}^{(r)}$.
Table 2.3 gives the values of $\boldsymbol{\beta}^{(r)}$ at different steps of the Fisher scoring algorithm.

TABLE 2.3: Values of $\boldsymbol{\beta}^{(r)}$ at different steps of the Fisher scoring algorithm.

r	$\alpha^{(r)}$	$\gamma^{(r)}$
0	80	0.1875
1	79.294919	0.2153986
2	78.820871	0.2201953
3	78.821823	0.2201973

This is the same result as obtained in Example 2.1 with the Newton-Raphson algorithm.
The program is given in the Appendix.

EXAMPLE 2.4
Determining ML estimates for a loglinear model using the Fisher scoring algorithm.

This example uses the model and data in Example 2.2 where the ML estimates for the GLM were found iteratively with the Newton-Raphson algorithm given by the equation

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left[\mathbf{X}'\text{diag}\left(\boldsymbol{\mu}^{(r)}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right).$$

Since

$$\mathbf{H}^{(r)} = -\mathbf{X}'\text{diag}\left(\boldsymbol{\mu}^{(r)}\right)\mathbf{X}$$

is not a function of the observed data $\mathbf{y}$, the observed and expected second derivative matrices are the same. Thus

$$\mathbf{Inf} = -\mathbf{H}.$$

This happens for all GLMs that use a canonical link function. The Newton-Raphson and Fisher scoring algorithms are identical in such cases.

2.3 IGNORABLE MISSING DATA MECHANISM

The EM algorithm can be used to determine maximum likelihood estimates for incomplete data. Before presenting the theory of the EM algorithm, it is necessary to define an ignorable missing data mechanism.

Suppose the data of interest is denoted by $\mathbf{Y} = (Y_{ij}) : n \times p$ matrix of n observations measured for p variables. The data is assumed to be generated by a model with probability function $f(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of unknown parameters. In the case of incomplete data let $\mathbf{Y}' = (\mathbf{Y}'_{obs}, \mathbf{Y}'_{mis})$ where $\mathbf{Y}_{obs}$ represents the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ denotes the missing values. The joint probability function of $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ is given by $f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta})$.
An indicator random variable is included in the model which indicates whether each component of $\mathbf{Y}$ is observed or missing. Define a response indicator $\mathbf{R} = (R_{ij})$ such that

$$R_{ij} = \begin{cases} 1, & y_{ij} \text{ observed}, \\ 0, & y_{ij} \text{ missing}. \end{cases}$$

The joint probability function of $\mathbf{R}$ and $\mathbf{Y}$ can be written as

$$f(\mathbf{y}, \mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}|\boldsymbol{\theta}) f(\mathbf{r}|\mathbf{y}, \boldsymbol{\psi}) \tag{13}$$

where $f(\mathbf{r}|\mathbf{y}, \boldsymbol{\psi})$ is the distribution of the missing data mechanism. This mechanism depends on $\mathbf{Y}$ and some unknown vector of parameters $\boldsymbol{\psi}$. In the case where the distribution of the missing data mechanism does not depend on the missing values $\mathbf{Y}_{mis}$, the data is said to be missing at random (MAR) and

$$f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\psi}) = f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\psi}). \tag{14}$$

MAR requires only that the missing values behave like a random sample within subclasses defined by the observed data. If the missing data values are a random sample of all data values the data is said to be missing completely at random (MCAR).
The observed data consist of the values of the variables $(\mathbf{Y}_{obs}, \mathbf{R})$ and its probability function is obtained by integrating out the missing data $\mathbf{Y}_{mis}$:

$$f(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta}) f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\psi}) \, d\mathbf{y}_{mis}. \tag{15}$$

The likelihood of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is proportional to (15), that is

$$l(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{obs}, \mathbf{r}) \propto f(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\psi}). \tag{16}$$

If the data is missing at random, that is if (14) holds, the probability function of the observed data, given in (15), can be written as

$$\begin{aligned} f(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta}) f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\psi}) \, d\mathbf{y}_{mis} \\ &= f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\psi}) \times \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta}) \, d\mathbf{y}_{mis} \\ &= f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\psi}) f(\mathbf{y}_{obs}|\boldsymbol{\theta}). \end{aligned} \tag{17}$$

The likelihood of the observed data under MAR can thus be factored into two pieces, one pertaining to the parameter of interest $\boldsymbol{\theta}$, and the other to $\boldsymbol{\psi}$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct if the joint parameter space of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is the product of the parameter space of $\boldsymbol{\theta}$ and the parameter space of $\boldsymbol{\psi}$. If both MAR and distinctness hold, the missing data mechanism is said to be ignorable (Little and Rubin, 1987) and likelihood based inferences about $\boldsymbol{\theta}$ will be unaffected by $\boldsymbol{\psi}$ or $f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\psi})$.
From equation (17) it follows that

$$f(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) \propto f(\mathbf{y}_{obs}|\boldsymbol{\theta})$$

and thus

$$l(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{obs}, \mathbf{r}) \propto l(\boldsymbol{\theta}|\mathbf{y}_{obs})$$

which means that all relevant statistical information about the parameters is contained in the observed data likelihood, $l(\boldsymbol{\theta}|\mathbf{y}_{obs})$.

EXAMPLE 2.5
Incomplete univariate data with an ignorable missing data mechanism.

Let $\mathbf{Y} : n \times 1$ denote a vector of n independent identically distributed random variables. Let $\mathbf{Y}' = (\mathbf{Y}'_{obs}, \mathbf{Y}'_{mis})$ with $\mathbf{Y}'_{obs} = (Y_1, Y_2, \ldots, Y_m)$ and $\mathbf{Y}'_{mis} = (Y_{m+1}, Y_{m+2}, \ldots, Y_n)$. That is, m units are observed and $n-m$ are missing. Let $\mathbf{R}' = (R_1, R_2, \ldots, R_n)$ denote the response indicators, where $R_i = 1$ if y_i is observed and $R_i = 0$ if y_i is missing. Suppose that each unit is observed with probability ψ. The missing data mechanism is

$$f(\mathbf{r}|\mathbf{y},\psi) = \prod_{i=1}^{n} \psi^{r_i} (1-\psi)^{1-r_i} = \psi^m (1-\psi)^{n-m}$$

and since it does not depend on $\mathbf{Y}_{mis}$ the data is MAR. If $\boldsymbol{\theta}$ and ψ are distinct, inferences about $\boldsymbol{\theta}$ can be based on the observed data likelihood

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}_{obs}) &= \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta})\, d\mathbf{y}_{mis} \\ &= \int \cdots \int \prod_{i=1}^{m} f(y_i|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(y_i|\boldsymbol{\theta})\, dy_{m+1} \cdots dy_n. \\ &= \prod_{i=1}^{m} f(y_i|\boldsymbol{\theta}) \end{aligned}$$

which is a complete data likelihood based on the reduced sample $(Y_1, Y_2, \ldots, Y_m)'$.

EXAMPLE 2.6
Bivariate data with one variable subject to nonresponse if the missing data mechanism is ignorable.

Consider a dataset with variables Y_1 and Y_2 where Y_1 is observed for units $1, 2, \ldots, n$ and Y_2 is observed only for units $1, 2, \ldots, m < n$. The missing data will be MAR if the probability that Y_2 is missing does not depend on Y_2, although it may possibly depend on Y_1. Let y_{i1} and y_{i2} denote the values of Y_1 and Y_2, respectively, for unit i. Since

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta}) = f(\mathbf{y}_{obs}|\boldsymbol{\theta})\, f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})$$

the observed data likelihood can be written as

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}_{obs}) &= \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta})\, d\mathbf{y}_{mis} \\ &= \int f(\mathbf{y}_{obs}|\boldsymbol{\theta})\, f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})\, d\mathbf{y}_{mis} \\ &= \int \prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(y_{i1}|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(y_{i2}|y_{i1}, \boldsymbol{\theta})\, d\mathbf{y}_{mis} \\ &= \prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(y_{i1}|\boldsymbol{\theta}) \int \prod_{i=m+1}^{n} f(y_{i2}|y_{i1}, \boldsymbol{\theta})\, d\mathbf{y}_{mis} \\ &= \prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(y_{i1}|\boldsymbol{\theta}). \end{aligned}$$

This is the product of the joint likelihood for Y_1 and Y_2 where Y_1 and Y_2 are both observed, and the likelihood of Y_1 where only Y_1 is observed.

2.4 THE EM ALGORITHM

2.4.1 Theory of the EM Algorithm

Assuming that the ignorability assumption is correct, all relevant statistical information about the parameters is contained in the observed data likelihood, $l(\boldsymbol{\theta}|\mathbf{y}_{obs})$. The EM algorithm uses the interdependence that exists between the missing data $\mathbf{Y}_{mis}$ and the parameters $\boldsymbol{\theta}$. An initial estimate of $\boldsymbol{\theta}$ is obtained from the observed data $\mathbf{Y}_{obs}$. The missing data is filled in based on this initial estimate of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ is then re-estimated based on $\mathbf{Y}_{obs}$ and the filled in $\mathbf{Y}_{mis}$. The process iterates until the estimates converge.
Suppose the density function of the complete data $\mathbf{y}$ is given by $f(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown parameter. Let $\mathbf{Y}' = (\mathbf{Y}'_{obs}, \mathbf{Y}'_{mis})$ where $\mathbf{Y}_{obs}$ represents the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ denotes the missing values. The distribution of the complete data can be factored as

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta}) = f(\mathbf{y}_{obs}|\boldsymbol{\theta})\, f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}). \tag{18}$$

The objective is to maximize the likelihood function for the observed data, that is maximize

$$l(\boldsymbol{\theta}|\mathbf{y}_{obs}) \propto \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\theta})\, d\mathbf{y}_{mis}$$

with respect to $\boldsymbol{\theta}$ or, alternatively, to maximize the log-likelihood

$$L(\boldsymbol{\theta}|\mathbf{y}_{obs}) = \log\left[l(\boldsymbol{\theta}|\mathbf{y}_{obs})\right].$$

The log-likelihood that corresponds to (18) is

$$L(\boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{y}_{mis}) = L(\boldsymbol{\theta}|\mathbf{y}_{obs}) + \log\left[f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})\right]$$

and can be written as

$$L(\boldsymbol{\theta}|\mathbf{y}_{obs}) = L(\boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{y}_{mis}) - \log\left[f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})\right] \tag{19}$$

where $L(\boldsymbol{\theta}|\mathbf{y}_{obs})$ is the observed log-likelihood to be maximized, $L(\boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{y}_{mis})$ is the complete data log-likelihood and $\log\left[f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})\right]$ is the missing part of the complete data log-likelihood.
The expectation of both sides of (19) over the distribution of the missing data $\mathbf{Y}_{mis}$, given $\mathbf{Y}_{obs}$ and a current estimate of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(r)}$ is

$$L(\boldsymbol{\theta}|\mathbf{y}_{obs}) = Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) - H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) \tag{20}$$

where

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) = \int \left[L(\boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{y}_{mis})\right] f\left(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}^{(r)}\right) d\mathbf{y}_{mis} \tag{21}$$

and

$$H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) = \int \left\{\log\left[f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta})\right]\right\} f\left(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}^{(r)}\right) d\mathbf{y}_{mis}. \tag{22}$$

From Jensen's inequality (Rao 1972)

$$H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) \leq H\left(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r)}\right) \tag{23}$$

and therefore maximization of $L(\boldsymbol{\theta}|\mathbf{y}_{obs})$ is equivalent to maximization of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ with respect to $\boldsymbol{\theta}$.
Each step of the EM algorithm consists of an E-step (expectation step) and an M-step (maximization step):

- In the E-step the function $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ is calculated by averaging the complete data log-likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ over $f\left(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}^{(r)}\right)$.
- In the M-step $\boldsymbol{\theta}^{(r+1)}$ is found by maximizing $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$. That is $Q\left(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{\theta}^{(r)}\right) \geq Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ for all $\boldsymbol{\theta}$.

2.4.2 The EM Algorithm for exponential families

Little and Rubin (1987) presents a simple characterization of the EM algorithm when $f(\mathbf{y}|\boldsymbol{\theta})$ has the form for the regular exponential family defined by

$$f(\mathbf{y}|\boldsymbol{\theta}) = b(\mathbf{y}) \exp\left(\mathbf{s}(\mathbf{y})' \boldsymbol{\theta}\right) / a(\boldsymbol{\theta}) \tag{24}$$

where $\boldsymbol{\theta}$ is the parameter vector and $\mathbf{s}(\mathbf{Y})$ is the vector of complete data sufficient statistics. For regular exponential families the complete data MLE can be found as a solution to the likelihood equations

$$E(\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta}) = \mathbf{s} \tag{25}$$

where $\mathbf{s}$ is the realized value of the vector $\mathbf{s}(\mathbf{Y})$.
Suppose $\boldsymbol{\theta}^{(r)}$ denotes the current value $\boldsymbol{\theta}$ after r cycles of the algorithm. The next cycle can be described in two steps, as follows:

- E-step: Estimate the complete data sufficient statistics $\mathbf{s}(\mathbf{Y})$ by finding

$$\mathbf{s}^{(r)} = E\left(\mathbf{s}(\mathbf{Y})|\mathbf{Y}_{obs}, \boldsymbol{\theta}^{(r)}\right). \tag{26}$$

- M-step: The M-step determines the new estimate $\boldsymbol{\theta}^{(r+1)}$ of $\boldsymbol{\theta}$ as the solution of the equations

$$E(\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta}) = \mathbf{s}^{(r)} \tag{27}$$

  which are the likelihood equations for the complete data with $\mathbf{s}(\mathbf{Y})$ replaced by $\mathbf{s}^{(r)}$ as obtained in the E-step in (26).

EXAMPLE 2.7
Incomplete univariate normal data. EM algorithm for the regular exponential family.

Suppose Y_i, $i = 1, 2, \ldots, n$ are independent identically distributed random variables from a $N(\mu, \sigma^2)$ distribution. Let $\boldsymbol{\theta}' = (\mu, \sigma^2)$. The log-likelihood function for the complete data is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right] \end{aligned}$$

which is linear in the sufficient statistics $\mathbf{s}(\mathbf{Y}) = (s_1(\mathbf{Y}), s_2(\mathbf{Y})) = \left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2\right)$.
With no missing data the ML estimates of μ and σ^2 are

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2 \end{aligned}$$

Suppose now that only the first m components of the data vector $\mathbf{Y}$ are observed and that the data are missing at random (MAR).

The E-step of the EM algorithm calculates

$$s_1^{(r)} = E\left(s_1(\mathbf{Y})|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right) = E\left(\sum_{i=1}^{n} Y_i|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right) = \sum_{i=1}^{m} y_i + (n-m)\,\mu^{(r)}$$

$$s_2^{(r)} = E\left(s_2(\mathbf{Y})|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right) = \sum_{i=1}^{m} y_i^2 + (n-m)\left[\left(\mu^{(r)}\right)^2 + \sigma^{2(r)}\right]$$

for current estimates $\boldsymbol{\theta}^{(r)} = \left(\mu^{(r)}, \sigma^{2(r)}\right)$ of the parameters. In the M-step the expectations of the sufficient statistics calculated in the E-step are substituted in the expressions for the ML estimates giving

$$\begin{aligned}\mu^{(r+1)} &= \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right)\\ &= \frac{1}{n}\left[\sum_{i=1}^{m} y_i + (n-m)\,\mu^{(r)}\right]\end{aligned}$$

and

$$\begin{aligned}\sigma^{2(r+1)} &= \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i^2|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right) - \left(\mu^{(r+1)}\right)^2\\ &= \frac{1}{n}\left[\sum_{i=1}^{m} y_i^2 + (n-m)\left[\left(\mu^{(r)}\right)^2 + \sigma^{2(r)}\right]\right] - \left(\mu^{(r+1)}\right)^2.\end{aligned}$$

Numerical Example

Suppose Y_i, $i = 1,2,\ldots,10$ are independent identically distributed random variables from a $N(12,9)$ distribution and that Y_i are observed for $i = 1,2,\ldots,6$ and missing for $i = 7,\ldots,10$. The 6 observed values are 12.893, 7.012, 12.165, 12.274, 14.657 and 8.644.

The initial values of $\mu^{(0)} = 10$ and $\sigma^{2(0)} = 10$ were chosen arbitrarily. Table 2.4 displays the results at different steps of the algorithm until convergence. The results are the same as the mean and variance for the six observed data points, that is

$$\begin{aligned}\widehat{\mu} &= \frac{1}{6}\sum_{i=1}^{6} y_i\\ \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{6} y_i^2}{6} - \widehat{\mu}^2\end{aligned}$$

TABLE 2.4: Iterations of the EM algorithm for incomplete univariate normal data, $n = 10$ and $m = 6$.

	M-Step		E-Step	
r	$\mu^{(r)}$	$\sigma^{2(r)}$	$E\left(\sum_{i=1}^{n} Y_i\|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right)$	$E\left(\sum_{i=1}^{n} Y_i^2\|\mathbf{Y}_{obs},\boldsymbol{\theta}^{(r)}\right)$
0	10	10	107.645	1243.582
1	10.765	8.4884	110.703	1301.015
2	11.070	7.550	111.926	1323.988
3	11.193	7.124	112.415	1333.178
4	11.242	6.945	112.611	1336.853
5	11.261	6.873	112.689	1338.324
6	11.269	6.843	112.721	1338.912
7	11.272	6.831	112.733	1339.147
8	11.273	6.827	112.738	1339.241
9	11.274	6.825	112.740	1339.279
10	11.274	6.824	112.741	1339.294
11	11.274	6.824	112.741	1339.300
∞	11.274	6.824	112.741	1339.300

EXAMPLE 2.8

EM algorithm for data from a multinomial distribution.

This example, discussed by Dempster, Laird and Rubin (1977) gives the data in which 197 animals are distributed multinomially into five categories. The complete data, $\mathbf{Y}' = (Y_1, Y_2, Y_3, Y_4, Y_5)$, are the counts for each category and the cell probabilities in this model are given as

$$\pi' = \left(\tfrac{1}{2}, \tfrac{1}{4}p, \tfrac{1}{4}(1-p), \tfrac{1}{4}(1-p), \tfrac{1}{4}p\right) \text{ for some } 0 \leq p \leq 1.$$

For the complete data the density function is

$$f(\mathbf{y}|p) = \frac{(y_1+y_2+y_3+y_4+y_5)!}{y_1!y_2!y_3!y_4!y_5!}\left(\tfrac{1}{2}\right)^{y_1}\left(\tfrac{1}{4}p\right)^{y_2}\left(\tfrac{1}{4}-\tfrac{1}{4}p\right)^{y_3}\left(\tfrac{1}{4}-\tfrac{1}{4}p\right)^{y_4}\left(\tfrac{1}{4}p\right)^{y_5}.$$

The ML estimate of p for the complete data is given by

$$\widehat{p} = \frac{y_2+y_5}{y_2+y_3+y_4+y_5}. \tag{28}$$

The kernel of the complete data log-likelihood is

$$L(p|\mathbf{y}) = y_1 \log \tfrac{1}{2} + (y_2+y_5)\log \tfrac{1}{4}p + (y_3+y_4)\log\left(\tfrac{1}{4}-\tfrac{1}{4}p\right)$$

and the counts are the sufficient statistics.
The observed data is $\mathbf{y}'_{obs} = (y_1+y_2, y_3, y_4, y_5) = (125, 18, 20, 34)$. Only the total of Y_1 and Y_2 is observed. In the E-step the conditional expectations of the sufficient statistics, Y_i, $i = 2,3,4,5$, given the observed values and a current estimate of p, are calculated. At step r $(r = 0, 1, 2, \ldots)$

$$\begin{aligned}
E(Y_2|\mathbf{Y}_{obs}, p^{(r)}) &= 125\frac{\frac{1}{4}p^{(r)}}{\frac{1}{2}+\frac{1}{4}p^{(r)}} \\
E(Y_3|\mathbf{Y}_{obs}, p^{(r)}) &= 18 \\
E(Y_4|\mathbf{Y}_{obs}, p^{(r)}) &= 20 \\
E(Y_5|\mathbf{Y}_{obs}, p^{(r)}) &= 34.
\end{aligned}$$

In the M-step the conditional expectations of Y_i as calculated in the E-step are substituted in expression (28) giving the next estimate of $\widehat{p}$ in the iterative process

$$\begin{aligned}
p^{(r+1)} &= \frac{E(Y_2|\mathbf{Y}_{obs}, p^{(r)}) + 34}{E(Y_2|\mathbf{Y}_{obs}, p^{(r)}) + 18 + 20 + 34} \\
&= \frac{125\frac{\frac{1}{4}p^{(r)}}{\frac{1}{2}+\frac{1}{4}p^{(r)}} + 34}{125\frac{\frac{1}{4}p^{(r)}}{\frac{1}{2}+\frac{1}{4}p^{(r)}} + 18 + 20 + 34}.
\end{aligned}$$

The process iterates between the E-step and the M-step until convergence is attained.
Table 2.5 shows that, starting from $p^{(0)} = 0.5$, the EM algorithm converges after seven steps.

TABLE 2.5: Iterations of the EM algorithm.

	M-step	E-step
r	$p^{(r)}$	$E(Y_2\|\mathbf{Y}_{obs}, p^{(r)})$
0	0.5	25
1	0.608247	29.15020
2	0.624321	29.73727
3	0.626489	29.82589
4	0.626777	29.82634
5	0.626816	29.82773
6	0.626821	29.82792
7	0.626821	29.82794
∞	0.626821	29.82794

2.5 A MAXIMUM LIKELIHOOD ESTIMATION PROCEDURE WHEN MODELLING IN TERMS OF CONSTRAINTS

Proposition 1
Suppose $\mathbf{Y}$ is a random vector with probability function belonging to the exponential family and with $E(\mathbf{Y}) = \boldsymbol{\mu}$. Matthews (1995) derives a ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$, as

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{G}_\mu \mathbf{V})' \left(\mathbf{G}_\mathbf{y} \mathbf{V} \mathbf{G}_\mu'\right)^{-1} g(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|) \tag{29}$$

where $\mathbf{g}(\boldsymbol{\mu})$ is a continuous vector valued function of $\boldsymbol{\mu}$ for which the first order partial derivatives exist, $\mathbf{G}_\mu = \dfrac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$, $\mathbf{G}_\mathbf{y} = \dfrac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}|_{\mu=\mathbf{y}}$ and $\mathbf{V}$ is the covariance matrix which could be known or could be some function of $\boldsymbol{\mu}$, say $\mathbf{V}_\mu$. This result implies that the ML estimate must be obtained iteratively.

Matthews (1995) gives the following proof of this result.

Proof:
Let $\boldsymbol{\gamma}$ be a vector of Lagrange multipliers. To find the ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$, we maximize

$$\frac{\partial}{\partial \boldsymbol{\mu}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \ln b(\mathbf{y}) + \mathbf{y}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta}) + \boldsymbol{\gamma}' \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\theta})).$$

Hence we find

$$\frac{\partial}{\partial \boldsymbol{\mu}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\theta}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\right].$$

Since we set $\dfrac{\partial}{\partial \boldsymbol{\mu}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \mathbf{0}$ for a maximum, and since $\left[\dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\right]$ is invertible for a regular exponential family, we need further only consider $\dfrac{\partial}{\partial \boldsymbol{\theta}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma})$.
Thus

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) &= \mathbf{y} - \frac{\partial}{\partial \boldsymbol{\theta}} \kappa(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}} \{\boldsymbol{\gamma}' \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\theta}))\} \\
&= \mathbf{y} - \boldsymbol{\mu} + \left\{\frac{\partial}{\partial \boldsymbol{\mu}} \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\theta})) \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]\right\}' \boldsymbol{\gamma} \\
&= \mathbf{y} - \boldsymbol{\mu} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu' \boldsymbol{\gamma}.
\end{aligned}$$

Setting $\dfrac{\partial}{\partial \boldsymbol{\theta}} \omega(\mathbf{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \mathbf{0}$, we get

$$\boldsymbol{\mu} = \mathbf{y} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu' \boldsymbol{\gamma}. \tag{30}$$

Using the linear Taylor expansion of $\mathbf{g}(\boldsymbol{\mu})$ about $\mathbf{y}$, we get

$$\begin{aligned}
\mathbf{g}(\boldsymbol{\mu}) &= \mathbf{g}\left(\mathbf{y} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu' \boldsymbol{\gamma}\right) \\
&= \mathbf{g}(\mathbf{y}) + \mathbf{G}_y \left(\mathbf{y} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu' \boldsymbol{\gamma} - \mathbf{y}\right) + o(\|\mathbf{y} - \boldsymbol{\mu}\|) \\
&= \mathbf{g}(\mathbf{y}) + \mathbf{G}_y \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu' \boldsymbol{\gamma} + o(\|\mathbf{y} - \boldsymbol{\mu}\|).
\end{aligned}$$

Setting $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$ and solving for $\boldsymbol{\gamma}$, gives

$$\boldsymbol{\gamma} = -\left(\mathbf{G}_y \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}_\mu'\right)^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|).$$

Substituting γ in (30) we get

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \left(\mathbf{G}_{\boldsymbol{\mu}} \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]\right)' \left(\mathbf{G}_{\mathbf{y}} \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' \mathbf{G}'_{\boldsymbol{\mu}}\right)^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|).$$

Now

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left\{\frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}\right\} = \frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}.$$

Hence

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \left[\frac{\partial \mu_i}{\partial \theta_j}\right] = \frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}$$

and

$$\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' = \left[\frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}\right] = \mathbf{V}.$$

Therefore

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{G}_{\boldsymbol{\mu}} \mathbf{V})' (\mathbf{G}_{\mathbf{y}} \mathbf{V} \mathbf{G}'_{\boldsymbol{\mu}})^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|)$$

which is the required result.

The iterative procedure

The process is a double iteration over $\mathbf{y}$ and $\boldsymbol{\mu}$. Let $\boldsymbol{\mu}^{(i,j)}$ denote the (i,j)th approximation obtained for the ML estimate $\widehat{\boldsymbol{\mu}}_c$ of $\boldsymbol{\mu}$, where i $(i = 0, 1, 2, \ldots)$ refers to iteration over $\boldsymbol{\mu}$, and j $(j = 0, 1, 2, \ldots)$ refers to iteration over $\mathbf{y}$. Note that $j = 0$ at the start of every iteration over $\mathbf{y}$.

The initial value for $\boldsymbol{\mu}$ is $\boldsymbol{\mu}^{(0,0)} = \mathbf{y}$, the vector of observed values. Iteration then takes place over $\mathbf{y}$ and the value of $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{V}_{\boldsymbol{\mu}}$ is kept constant at $\boldsymbol{\mu}^{(0,0)} = \mathbf{y}$. The first approximation of $\widehat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\mu}^{(0,1)} = \mathbf{y} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,0)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{G}_{\mathbf{y}} \mathbf{V}_{\boldsymbol{\mu}^{(0,0)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,0)}}\right)^{-1} \mathbf{g}(\mathbf{y}).$$

If convergence over $\mathbf{y}$ is not attained at this step, $\mathbf{y}$ is replaced by $\boldsymbol{\mu}^{(0,1)}$ to obtain the next approximation of $\widehat{\boldsymbol{\mu}}_c$, whilst the estimated value for $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{V}_{\boldsymbol{\mu}}$ is kept constant at $\boldsymbol{\mu}^{(0,0)} = \mathbf{y}$. Thus,

$$\boldsymbol{\mu}^{(0,2)} = \boldsymbol{\mu}^{(0,1)} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,0)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,1)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,0)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,0)}}\right)^{-1} \mathbf{g}\left(\boldsymbol{\mu}^{(0,1)}\right).$$

This is repeated until convergence over $\mathbf{y}$ is attained, say at $j = k$.

The value at convergence, $\boldsymbol{\mu}^{(0,k)}$, is used as the next estimate for $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{V}_{\boldsymbol{\mu}}$. The procedure again iterates over $\mathbf{y}$, starting with the vector of observed values, $\mathbf{y}$, and keeping the estimated value for $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{V}_{\boldsymbol{\mu}}$ constant at $\boldsymbol{\mu}^{(0,k)}$. That is

$$\boldsymbol{\mu}^{(1,1)} = \mathbf{y} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,k)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,k)}}\right)' \left(\mathbf{G}_{\mathbf{y}} \mathbf{V}_{\boldsymbol{\mu}^{(0,k)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,k)}}\right)^{-1} \mathbf{g}(\mathbf{y}).$$

If convergence over $\mathbf{y}$ is not obtained at this step, the next approximation of $\widehat{\boldsymbol{\mu}}_c$ is

$$\boldsymbol{\mu}^{(1,2)} = \boldsymbol{\mu}^{(1,1)} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,k)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,k)}}\right)' \left(\mathbf{G}_{\boldsymbol{\mu}^{(1,1)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,k)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,k)}}\right)^{-1} \mathbf{g}\left(\boldsymbol{\mu}^{(1,1)}\right).$$

At convergence the iteration over $\mathbf{y}$ yields the next estimate for $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{V}_{\boldsymbol{\mu}}$. The process continues until convergence over $\boldsymbol{\mu}$ is attained.

In certain cases the iterative procedure simplifies to an iteration only over $\mathbf{y}$ or only over $\boldsymbol{\mu}$.

- If $\mathbf{g}$ is a linear function of $\boldsymbol{\mu}$, say $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\mu}$ then $\mathbf{G}_{\mu} = \mathbf{A} = \mathbf{G}_{\mathbf{y}}$ and

$$\boldsymbol{\mu}^{(0,1)} = \mathbf{y} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{y}. \tag{31}$$

For the iteration over $\mathbf{y}$ convergence is immediately attained since substitution of $\boldsymbol{\mu}^{(0,1)}$ into $\mathbf{y}$ in equation (31) gives

$$\begin{aligned}
& \boldsymbol{\mu}^{(0,1)} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}'_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A}\boldsymbol{\mu}^{(0,1)} \\
= \; & \mathbf{y} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{y} - \\
& \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}'_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A} \left[\mathbf{y} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{y}\right] \\
= \; & \mathbf{y} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\right)' \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{y} \\
= \; & \boldsymbol{\mu}^{(0,1)}.
\end{aligned}$$

The process simplifies to iteration only over $\boldsymbol{\mu}$ with $\mathbf{y}$ remaining constant.
At step $i+1 \quad (i = 0, 1, 2, \ldots)$ the approximation of $\widehat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\mu}^{(i+1)} = \mathbf{y} - \left(\mathbf{A}\mathbf{V}_{\boldsymbol{\mu}^{(i)}}\right)' \left(\mathbf{A}\mathbf{V}'_{\boldsymbol{\mu}^{(i)}}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{y}$$

with $\boldsymbol{\mu}^{(0)} = \mathbf{y}$. The process converges to the ML estimate $\widehat{\boldsymbol{\mu}}_c$.

- Let $\mathbf{D}_{\mu}$ be a diagonal matrix with the elements of $\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$ on the principal diagonal and $\mathbf{V} = \mathbf{D}_{\mu}$. Suppose $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{A}\log(\boldsymbol{\mu})$. Then

$$\begin{aligned}
\mathbf{G}_{\mu} &= \frac{\partial}{\partial \boldsymbol{\mu}} \mathbf{A}\log(\boldsymbol{\mu}) = \mathbf{A}\mathbf{D}_{\mu}^{-1} \\
\mathbf{G}_{\mathbf{y}} &= \mathbf{A}\mathbf{D}_{\mathbf{y}}^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - \left(\mathbf{G}_{\mu}\mathbf{V}_{\mu}\right)' \left(\mathbf{G}_{\mathbf{y}}\mathbf{V}_{\mu}\mathbf{G}'_{\mu}\right)^{-1} \mathbf{A}\log(\mathbf{y}) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right) \\
&= \mathbf{y} - \left(\mathbf{A}\mathbf{D}_{\mu}^{-1}\mathbf{D}_{\mu}\right)' \left(\mathbf{A}\mathbf{D}_{\mathbf{y}}^{-1}\mathbf{D}_{\mu}\mathbf{D}_{\mu}^{-1}\mathbf{A}'\right)^{-1} \mathbf{A}\log(\mathbf{y}) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right) \\
&= \mathbf{y} - \mathbf{A}'\left(\mathbf{A}\mathbf{D}_{\mathbf{y}}^{-1}\mathbf{A}'\right)^{-1} \mathbf{A}\log(\mathbf{y}) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right).
\end{aligned}$$

Iteration is only over $\mathbf{y}$. At step $j+1 \quad (j = 0, 1, 2, \ldots)$ the approximation of $\widehat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\mu}^{(j+1)} = \boldsymbol{\mu}^{(j)} - \mathbf{A}'\left(\mathbf{A}\mathbf{D}_{\boldsymbol{\mu}^{(j)}}^{-1}\mathbf{A}'\right)^{-1} \mathbf{A}\log\left(\boldsymbol{\mu}^{(j)}\right)$$

with $\boldsymbol{\mu}^{(0)} = \mathbf{y}$. The process converges to the ML estimate $\widehat{\boldsymbol{\mu}}_c$.

Proposition 2
The asymptotic covariance matrix of $\widehat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\Sigma}_c = \mathbf{V}_{\mu} - \left(\mathbf{G}_{\mu}\mathbf{V}_{\mu}\right)' \left(\mathbf{G}_{\mu}\mathbf{V}_{\mu}\mathbf{G}'_{\mu}\right)^{-1} \mathbf{G}_{\mu}\mathbf{V}_{\mu}.$$

with the MLE obtained by replacing $\boldsymbol{\mu}$ with $\widehat{\boldsymbol{\mu}}_c$.

EXAMPLE 2.9

Determining ML estimates under constraints with iteration over $\mathbf{y}$ *and* $\boldsymbol{\mu}$.

The number of accidents per thousand per age group in a certain factory is given in Table 2.6.

TABLE 2.6: Accidents per 1000 per age group.

Age group	I	II	III
Number of accidents	80	15	5

The model under consideration is $\mu_i = \alpha\gamma^{i-1}$ for $i = 1, 2, 3$, and independent Poisson sampling is assumed.
This model implies the constraint

$$\mathbf{g}(\boldsymbol{\mu}) = \mu_1\mu_3 - \mu_2^2 = 0.$$

In this case

$\mathbf{V}_\mu = \mathbf{D}_\mu$

$\mathbf{G}_\mu = (\ \mu_3,\ \ -2\mu_2,\ \ \mu_1\ )$

$\mathbf{G}_\mathbf{y} = (\ y_3,\ \ -2y_2,\ \ y_1\ )$

$\mathbf{G}_\mu\mathbf{D}_\mu = (\ \mu_1\mu_3,\ \ -2\mu_2^2,\ \ \mu_1\mu_3\ )$

$\mathbf{G}_\mathbf{y}\mathbf{D}_\mu\mathbf{G}'_\mu = (y_1 + y_3)\,\mu_1\mu_3 + 4y_2\mu_2^2.$

The ML estimate of $\boldsymbol{\mu}$ is found iteratively from

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{G}_\mu\mathbf{D}_\mu)'\left(\mathbf{G}_\mathbf{y}\mathbf{D}_\mu\mathbf{G}'_\mu\right)^{-1}\mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|). \tag{32}$$

Iteration is over $\mathbf{y}$ and $\boldsymbol{\mu}$. The process converges after eight steps.

Table 2.7 gives the approximation of $\widehat{\boldsymbol{\mu}}_c$ at different steps of the iterative procedure.These are the same results as obtained by the Newton-Raphson and Fisher scoring algorithms (see Examples 2.1, 2.2 and 2.3).

TABLE 2.7: Approximation of $\widehat{\boldsymbol{\mu}}_c$ at different steps of the iterative procedure.

i	$\mu_1^{(i,j)}$	$\mu_2^{(i,j)}$	$\mu_3^{(i,j)}$	j	$\mu_1^{(i,j)}$	$\mu_2^{(i,j)}$	$\mu_3^{(i,j)}$
0	80	15	5	0	80	15	5
				1	78.526316	16.657895	3.5263158
				2	78.531142	16.652465	3.5311418
1	78.531142	78.531142	3.5311418	0	80	15	5
				1	78.793103	17.413793	3.7931034
				2	78.821807	17.356387	3.8218065
				3	78.821823	17.356354	3.8218228
2	78.821823	17.356354	3.8218228	0	80	15	5
				1	78.793103	17.413793	3.7931034
				2	78.821807	17.356387	3.8218065
				3	78.821823	17.356354	3.8218228

Description of the procedure:

- Both $\mathbf{y}$ and $\boldsymbol{\mu}$ in equation (32) are initially estimated by the observed data, that is $\mathbf{y} = \boldsymbol{\mu}^{(0,0)}$. The first approximation of $\widehat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\mu}^{(0,1)} = \mathbf{y} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{D}_{\boldsymbol{\mu}^{(0,0)}}\right)'\left(\mathbf{G}_\mathbf{y}\mathbf{D}_{\boldsymbol{\mu}^{(0,0)}}\mathbf{G}'_{\boldsymbol{\mu}^{(0,0)}}\right)^{-1}\mathbf{g}(\mathbf{y}).$$

  The process iterates over $\mathbf{y}$ until convergence is attained at $(i, j) = (0, 2)$. At this stage the approximation of $\widehat{\boldsymbol{\mu}}_c$ is

$$\boldsymbol{\mu}^{(0,2)} = \begin{pmatrix} 78.531142 \\ 16.652465 \\ 3.5311418 \end{pmatrix}$$

  This becomes the next estimate of $\boldsymbol{\mu}$ in $\mathbf{G}_\mu$ and $\mathbf{D}_\mu$.

- The process again iterates over $\mathbf{y}$ with the initial value of $\mathbf{y} = \begin{pmatrix} 80 \\ 15 \\ 5 \end{pmatrix}$, the vector of observed data.
  For $(i, j) = (1, 0)$

$$\boldsymbol{\mu}^{(1,0)} = \mathbf{y} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,2)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,2)}}\right)' \left(\mathbf{G}_{\mathbf{y}} \mathbf{V}_{\boldsymbol{\mu}^{(0,2)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,2)}}\right)^{-1} \mathbf{g}(\mathbf{y})$$

  and for $(i, j) = (1, 1)$

$$\boldsymbol{\mu}^{(1,1)} = \boldsymbol{\mu}^{(1,0)} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(0,2)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,2)}}\right)' \left(\mathbf{G}_{\boldsymbol{\mu}^{(1,0)}} \mathbf{V}_{\boldsymbol{\mu}^{(0,2)}} \mathbf{G}'_{\boldsymbol{\mu}^{(0,2)}}\right)^{-1} \mathbf{g}\left(\boldsymbol{\mu}^{(1,0)}\right).$$

  Convergence is attained at $(i, j) = (1, 3)$. The vector $\boldsymbol{\mu}^{(1,3)}$ becomes the next estimate of $\boldsymbol{\mu}$ in $\mathbf{G}_{\boldsymbol{\mu}}$ and $\mathbf{D}_{\boldsymbol{\mu}}$.

- The process again iterates over $\mathbf{y}$ with the initial value of $\mathbf{y}$ the vector of observed data.This iteration over $\mathbf{y}$ converges at $(i, j) = (2, 3)$ and at this stage

$$\boldsymbol{\mu}^{(2,3)} = \boldsymbol{\mu}^{(2,2)} - \left(\mathbf{G}_{\boldsymbol{\mu}^{(1,3)}} \mathbf{V}_{\boldsymbol{\mu}^{(1,3)}}\right)' \left(\mathbf{G}_{\boldsymbol{\mu}^{(2,2)}} \mathbf{V}_{\boldsymbol{\mu}^{(1,3)}} \mathbf{G}'_{\boldsymbol{\mu}^{(1,3)}}\right)^{-1} \mathbf{g}\left(\boldsymbol{\mu}^{(2,2)}\right).$$

  Since $\boldsymbol{\mu}^{(2,3)} = \boldsymbol{\mu}^{(1,3)}$ convergence over $\boldsymbol{\mu}$ is also attained at this step and the process stops.

The program is given in the Appendix.

EXAMPLE 2.10

Determining ML estimates under constraints with iteration over $\mathbf{y}$.

Consider the same data as in Example 2.9 but using the constraint

$$\mathbf{g}(\boldsymbol{\mu}) = \log(\mu_1\mu_3) - 2\log(\mu_2) = 0$$

In this case

$\mathbf{V} = \mathbf{D}_\mu$

$\mathbf{G}_\mu = \left(\frac{1}{\mu_1}, \quad \frac{-2}{\mu_2}, \quad \frac{1}{\mu_3} \right)$

$\mathbf{G}_\mathbf{y} = \left(\frac{1}{y_1}, \quad \frac{-2}{y_2}, \quad \frac{1}{y_3} \right)$

$\mathbf{G}_\mu\mathbf{D}_\mu = \left(1, \quad -2, \quad 1 \right)$

$\mathbf{G}_\mathbf{y}\mathbf{D}_\mu\mathbf{G}'_\mu = \frac{1}{y_1} + \frac{4}{y_2} + \frac{1}{y_3}.$

The ML estimate of $\boldsymbol{\mu}$ is found iteratively from

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_\mu\mathbf{D}_\mu)' \left(\mathbf{G}_\mathbf{y}\mathbf{D}_\mu\mathbf{G}'_\mu\right)^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|) \\ &= \mathbf{y} - \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \frac{\log(y_1y_3) - 2\log(y_2)}{\frac{1}{y_1} + \frac{4}{y_2} + \frac{1}{y_3}} + o(\|\mathbf{y} - \boldsymbol{\mu}\|). \end{aligned}$$

Iteration is only over $\mathbf{y}$.

Table 2.8 gives the estimates of $\widehat{\boldsymbol{\mu}}_c$ at different steps of the iterative procedure.

TABLE 2.8: Approximation of $\widehat{\boldsymbol{\mu}}_c$ at different steps of the iterative procedure.

	Approximation of $\widehat{\boldsymbol{\mu}}_c$ by $\boldsymbol{\mu}^{(r)}$		
r	$\mu_1^{(r)}$	$\mu_2^{(r)}$	$\mu_3^{(r)}$
0	80	15	5
1	78.79924	17.40152	3.79924
2	78.821801	17.356397	3.8218013
3	78.821823	17.356354	3.8218228

Alternatively, the constraint can also be set up in terms of the GLM given in Example 1.2. The model is

$$\log\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

with $\boldsymbol{\beta}' = (\beta_1, \beta_2)$ where $\beta_1 = \log\alpha$ and $\beta_2 = \log\gamma$, and $\mathbf{X}$ the design matrix given in Example 1.2. Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'$. The model can be written in terms of the implied constraints as

$$\mathbf{g}(\boldsymbol{\mu}) = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}']\log\boldsymbol{\mu} = \mathbf{P}\log\boldsymbol{\mu} = \mathbf{0}.$$

The ML estimate for $\boldsymbol{\mu}$ subject to the constraint $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$ is found iteratively from

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{G}_\mu\mathbf{V}_\mu)' \left(\mathbf{G}_\mathbf{y}\mathbf{V}_\mu\mathbf{G}'_\mu\right)^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|)$$

with $\mathbf{V}_\mu = \mathbf{D}_\mu$

$\mathbf{G}_\mu = \mathbf{P}\mathbf{D}_\mu^{-1}$

$\mathbf{G}_\mathbf{y} = \mathbf{P}\mathbf{D}_\mathbf{y}^{-1}$

$\mathbf{G}_\mu\mathbf{V} = \mathbf{P}$

$\mathbf{G}_\mathbf{y}\mathbf{V}\mathbf{G}'_\mu = \mathbf{P}\mathbf{D}_\mathbf{y}^{-1}\mathbf{P}.$

Hence, the estimation procedure is

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \mathbf{P}\left(\mathbf{P}\mathbf{D}_\mathbf{y}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}\log\mathbf{y} + o(\|\mathbf{y} - \boldsymbol{\mu}\|).$$

Iteration is only over $\mathbf{y}$. The estimates of $\widehat{\boldsymbol{\mu}}_c$ at different steps of the iterative procedure is exactly the same as given in Table 2.8. The programs with these two restrictions are given in the Appendix.

EXAMPLE 2.11

Determination of maximum likelihood estimates under constraints. An example for incomplete data.

Example 2.8 gives data in which 197 animals are distributed multinomially into five categories.
The complete data, $\mathbf{Y}' = (Y_1, Y_2, Y_3, Y_4, Y_5)$, are the counts for each category and the cell probabilities in this model are given as

$$\boldsymbol{\pi}' = \left(\tfrac{1}{2}, \tfrac{1}{4}p, \tfrac{1}{4}(1-p), \tfrac{1}{4}(1-p), \tfrac{1}{4}p\right) \text{ for some } 0 \leq p \leq 1.$$

The random vector of complete data is $\mathbf{Y}'=(Y_1, Y_2, Y_3, Y_4, Y_5)$ and the random vector of observed data is $\mathbf{Y}'_{obs} = (Y_1 + Y_2, Y_3, Y_4, Y_5)$ where only the sum of Y_1 and Y_2 is observed. The observed data is $\mathbf{y}'_{obs} = (125, 18, 20, 34)$.
The distributions of $\mathbf{Y}$ and $\mathbf{Y}_{obs}$ are both multinomial and can be written as

$$\mathbf{Y} \sim Mult(n, \boldsymbol{\pi})$$

with

$$\begin{aligned} \boldsymbol{\pi}' &= (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) \\ &= \left(\tfrac{1}{2}, \tfrac{1}{4}p, \tfrac{1}{4}(1-p), \tfrac{1}{4}(1-p), \tfrac{1}{4}p\right) \text{ for some } 0 \leq p \leq 1 \end{aligned}$$

and

$$\mathbf{Y}_{obs} \sim Mult(n, \boldsymbol{\pi}_{obs})$$

with

$$\begin{aligned} \boldsymbol{\pi}'_{obs} &= (\pi_1 + \pi_2, \pi_3, \pi_4, \pi_5) \\ &= \left(\tfrac{1}{2} + \tfrac{1}{4}p, \tfrac{1}{4}(1-p), \tfrac{1}{4}(1-p), \tfrac{1}{4}p\right) \text{ for some } 0 \leq p \leq 1. \end{aligned} \tag{33}$$

The ML estimate of p must be obtained from the observed data, $\mathbf{Y}_{obs}$. For the multinomial distribution

$$E(\mathbf{Y}_{obs}) = n\boldsymbol{\pi}_{obs} = \boldsymbol{\mu}_{obs}.$$

From the cell probabilities given in (33) the constraint $\mathbf{g}(\boldsymbol{\mu}_{obs}) = \mathbf{0}$ can be written as

$$\mathbf{g}(\boldsymbol{\mu}_{obs}) = \mathbf{X}\boldsymbol{\mu}_{obs} = \begin{pmatrix} 1 & -1 & -1 & -3 \\ 0 & 1 & -1 & 0 \end{pmatrix} \boldsymbol{\mu}_{obs}$$

where $\boldsymbol{\mu}'$ is the vector of expected cell counts.
The ML estimate, $\widehat{\boldsymbol{\mu}}_{obs,c}$, of the expected cell counts $\boldsymbol{\mu}_{obs}$ are obtained by solving

$$\widehat{\boldsymbol{\mu}}_{obs,c} = \mathbf{y}_{obs} - \left(\mathbf{G}_{\boldsymbol{\mu}_{obs}}\mathbf{V}_{\boldsymbol{\mu}_{obs}}\right)' \left(\mathbf{G}_{\mathbf{y}_{obs}}\mathbf{V}_{\boldsymbol{\mu}_{obs}}\mathbf{G}'_{\boldsymbol{\mu}_{obs}}\right)^{-1} \mathbf{g}(\mathbf{y}_{obs}) + o(\|\mathbf{y}_{obs} - \boldsymbol{\mu}_{obs}\|)$$

where $\mathbf{V}_{\boldsymbol{\mu}_{obs}} = \text{Diag}(\mathbf{y}_{obs}) - \frac{1}{n}\mathbf{y}_{obs}\mathbf{y}'_{obs}$
$\mathbf{G}_{\boldsymbol{\mu}_{obs}} = \mathbf{X} = \mathbf{G}_{\mathbf{y}_{obs}}$
$\mathbf{g}(\mathbf{y}_{obs}) = \mathbf{X}\mathbf{y}_{obs}$.
Since $\mathbf{g}(\boldsymbol{\mu}_{obs})$ is a linear function of $\boldsymbol{\mu}_{obs}$ iteration is only over $\boldsymbol{\mu}_{obs}$.
The ML estimate of p is then determined from $\widehat{\boldsymbol{\mu}}_{obs,c}$ by

$$\widehat{p} = 4\frac{\widehat{\mu}_{obs,4}}{n}.$$

The process converges after 4 steps and $\widehat{\boldsymbol{\mu}}_{obs,c} = \begin{pmatrix} 129.37096 \\ 18.379041 \\ 18.379041 \\ 30.870959 \end{pmatrix}$ giving $\widehat{p} = 0.6268215$. This is the same result as obtained with the EM algorithm in Example 2.8.
The program is given in the Appendix.