1 INTRODUCTION

There are a large number of maximum likelihood estimation procedures for categorical data available for scientific application. In this dissertation the most commonly used methods are compared with a maximum likelihood estimation procedure under constraints and an exposition of the theory and application of the methods are given.

The more generally used methods of maximum likelihood estimation for categorical data includes the Newton-Raphson and Fisher scoring algorithms for complete data and the EM algorithm for incomplete data. The Newton-Raphson algorithm is an iterative procedure which is employed for solving non-linear equations. It makes use of the vector of first order partial derivatives and matrix of second order partial derivatives of the function to be maximized. The Fisher scoring algorithm is similar to the Newton-Raphson algorithm, the distinction being that Fisher scoring uses the expected value of the second derivative with respect to the parameters in the model.
In the broad class of models referred to as generalized linear models the observations come from an exponential family and a function of their expectation is written as a linear model using a link function. Agresti (1990) shows that when a canonical link function is used the Newton-Raphson and Fisher scoring algorithms are identical.

The EM algorithm can be used for maximum likelihood estimation in incomplete contingency tables. The algorithm makes use of the interdependence between the missing data and the parameters to be estimated. The missing data are filled in based on an initial estimate of the parameters (the E-step). The parameters are then re-estimated based on the observed data and the filled in data (the M-step). The process iterates between the two steps until the estimates converge. The EM algorithm is specifically applied to the exponential family to determine ML estimates in incomplete contingency tables when the missing data mechanism is ignorable. Little and Rubin (1987) describes and uses the EM algorithm to determine the ML estimates of cell probabilities for loglinear models.

Matthews (1995) presents a maximum likelihood estimation procedure for the mean of the exponential family subject to the constraint $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{g}$ is a vector valued function of $\boldsymbol{\mu}$.
For the loglinear model and logistic regression the results obtained from this method are the same as those obtained from the Newton-Raphson algorithm.
The analysis of patterns of symmetry in squared contingency tables are considered by using ML estimation under contraints and a program is given which can be used for any squared contingency table. Results obtained are the same as the special cases considered in literature.
The method is also further developed to determine maximum likelihood estimates for loglinear models when the contingency table is incomplete and the missing data mechanism is ignorable. This also illustrates the elegance with which the method of ML estimation under contraints can be applied.
The method under constraints is conceptually comprehensive, logically clear and at the same time computationally less intensive than the EM and other algorithms.

1.1 THE EXPONENTIAL FAMILY

Let $\mathbf{Y}$ be a $p \times 1$ random vector and $\boldsymbol{\theta}$ a $p \times 1$ vector of parameters. Barndorff-Nielsen (1978) defines the exponential family by

$$p(\mathbf{y}, \boldsymbol{\theta}) = b(\mathbf{y}) \exp\left[\mathbf{y}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})\right], \quad \mathbf{y} \in \Re^p, \quad \boldsymbol{\theta} \in \aleph \tag{1}$$

where $\kappa(\boldsymbol{\theta})$ is referred to as the cumulant generating function and $\aleph$ is the natural parameter space for the canonical parameter $\boldsymbol{\theta}$.

The moment generating function of the exponential family is given by

$$\begin{aligned}
\mathbf{M}_{\mathbf{Y}}(\mathbf{t}) &= E\left[e^{\mathbf{t}'\mathbf{Y}}\right] \\
&= \int \cdots \int b(\mathbf{y}) \exp\left[\mathbf{y}'(\boldsymbol{\theta}+\mathbf{t}) - \kappa(\boldsymbol{\theta})\right] d\mathbf{y} \\
&= \exp\left[-\kappa(\boldsymbol{\theta})\right] \int \cdots \int b(\mathbf{y}) \exp\left[\mathbf{y}'(\boldsymbol{\theta}+\mathbf{t})\right] d\mathbf{y} \\
&= \exp\left[-\kappa(\boldsymbol{\theta})\right] \exp\left[\kappa(\boldsymbol{\theta}+\mathbf{t})\right] \int \cdots \int b(\mathbf{y}) \exp\left[\mathbf{y}'(\boldsymbol{\theta}+\mathbf{t}) - \kappa(\boldsymbol{\theta}+\mathbf{t})\right] d\mathbf{y} \\
&= \exp\left[-\kappa(\boldsymbol{\theta})\right] \exp\left[\kappa(\boldsymbol{\theta}+\mathbf{t})\right].
\end{aligned}$$

From this the cumulant generating function can be derived.

$$\begin{aligned}
\log \mathbf{M}_{\mathbf{Y}}(\mathbf{t}) &= \kappa(\boldsymbol{\theta}+\mathbf{t}) - \kappa(\boldsymbol{\theta}) \\
&= \kappa(\boldsymbol{\theta}) + \left[\frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\right]' \mathbf{t} + \tfrac{1}{2}\mathbf{t}' \left[\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta})\right] \mathbf{t} + r(\mathbf{t}) - \kappa(\boldsymbol{\theta}) \\
&= \left[\frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\right]' \mathbf{t} + \tfrac{1}{2}\mathbf{t}' \left[\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta})\right] \mathbf{t} + r(\mathbf{t}).
\end{aligned}$$

The mean vector and covariance matrix of $\mathbf{Y}$ are given by

$$E(\mathbf{Y}) = \frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \boldsymbol{\mu} \quad \text{and} \quad \mathrm{Cov}(\mathbf{Y}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta}) = \mathbf{V}.$$

EXAMPLE 1.1

The Poisson distribution as a member of the exponential family.

Let Y_i, $i = 1, 2, \ldots, p$ be independent Poisson random variables with $E(Y_i) = \mu_i$. The joint probability function of $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$ is

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\mu}) = \frac{\exp\left(-\sum \mu_i\right) \prod \mu_i^{y_i}}{\prod y_i!} = \exp\left[\sum y_i \log \mu_i - \sum \mu_i\right] \exp\left[-\sum \log y_i!\right]$$

which is a member of the exponential family since it has the form

$$p(\mathbf{y}, \boldsymbol{\theta}) = b(\mathbf{y}) \exp\left[\mathbf{y}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})\right]$$

with $b(\mathbf{y}) = \exp\left[-\sum \log y_i!\right]$
$\boldsymbol{\theta}$ a $p \times 1$ vector with $\theta_i = \log \mu_i$, that is $\mu_i = e^{\theta_i}$
$\kappa(\boldsymbol{\theta}) = \sum \mu_i = \sum \exp(\theta_i)$.

The mean vector of $\mathbf{Y}$ is given by

$$\begin{aligned}
E(\mathbf{Y}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \\
&= \begin{pmatrix} \frac{\partial\kappa(\boldsymbol{\theta})}{\partial\theta_1} \\ \frac{\partial\kappa(\boldsymbol{\theta})}{\partial\theta_2} \\ \vdots \\ \frac{\partial\kappa(\boldsymbol{\theta})}{\partial\theta_p} \end{pmatrix} = \begin{pmatrix} e^{\theta_1} \\ e^{\theta_2} \\ \vdots \\ e^{\theta_p} \end{pmatrix} \\
&= \boldsymbol{\mu}.
\end{aligned}$$

The covariance matrix of $\mathbf{Y}$ is

$$\begin{aligned}
\mathrm{Cov}(\mathbf{Y}) &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta}) \\
&= \begin{pmatrix} \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_1} & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_p} \\ \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_2\partial\theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_1} & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_2} & \cdots & \frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_p} \end{pmatrix} = \begin{pmatrix} e^{\theta_1} & 0 & \cdots & 0 \\ 0 & e^{\theta_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & e^{\theta_p} \end{pmatrix} \\
&= \mathrm{Diag}(\boldsymbol{\mu}).
\end{aligned}$$

1.2 COMPONENTS OF A GENERALIZED LINEAR MODEL

Suppose that $\mathbf{Y} : p \times 1$ is a random vector and that the joint probability function is a member of the natural exponential family with $E(\mathbf{Y}) = \boldsymbol{\mu}$. Let $\boldsymbol{\theta}$ be a $p \times 1$ vector of natural parameters.
A generalized linear model (GLM) consists of the following three components:

1. *The random component.*
   The random component, $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$, refers to the vector with response variables from a distribution in the natural exponential family. That is, the joint probability function is of the form given in (1).

2. *The systematic component.*
   The systematic component relates parameters $\{\eta_i\}$ to the explanatory variables using a linear predictor
   $$\eta_i = \sum_j \beta_j x_{ij} \quad i = 1, 2, \ldots, p.$$
   In matrix form
   $$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$
   where $\boldsymbol{\eta} : p \times 1$, $\boldsymbol{\beta} : m \times 1$ are model parameters, and $\mathbf{X} : p \times m$ is the design matrix consisting of the values of the explanatory variables for the p observations.

3. *The link between the random and systematic components.*
   The link function h, connects the expected values of the random component, μ_i, to the linear predictor by
   $$h(\mu_i) = \eta_i$$
   where h is a monotonic differentiable function.
   A GLM links μ_i to the explanatory variables through the equation
   $$h(\mu_i) = \eta_i = \sum_j \beta_j x_{ij} \quad i = 1, 2, \ldots, p.$$
   The link function that transforms μ_i, to the natural parameter θ_i, is called the canonical link, for which
   $$h(\mu_i) = \eta_i = \theta_i = \sum_j \beta_j x_{ij}.$$

EXAMPLE 1.2
The components of a GLM for a loglinear model.

Suppose the elements of $\mathbf{Y}: 3 \times 1$ are independent Poisson random variables with parameter vector $\boldsymbol{\mu}$. The model to be fitted is $\mu_i = \alpha\gamma^{i-1}$ or, as a loglinear model

$$\log \mu_i = \log \alpha + (i-1) \log \gamma.$$

The generalized linear model is

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

The three components of the GLM are:

1. The random component $\mathbf{Y}$.
   In Example 1.1 it was shown that the joint probability function of $\mathbf{Y}$ is a member of the natural exponential family.

2. The systematic component

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

   with $\boldsymbol{\beta}' = (\beta_1, \beta_2)$ where $\beta_1 = \log \alpha$ and $\beta_2 = \log \gamma$.

3. The link function, which is also a canonical link for this example, is given by

$$\eta_i = h(\mu_i) = \log \mu_i = \theta_i = \sum_j \beta_j x_{ij}.$$

1.3 MEASURES OF GOODNESS OF FIT

Suppose that $\{\widehat{\mu}_i\}$ are the estimated frequencies for the contingency table on fitting an appropriate model to the data The following statistics can be used to test the goodness of fit of a model:

- The Pearson Chi-squared Statistic

$$\chi^2 = \sum_{i=1}^{p} \frac{(\mu_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

- The Deviance
  A saturated GLM has as many parameters as observations, giving a perfect fit. In a saturated model all variation is consigned to the systematic component. For a given unsaturated model the ratio

$$-2\log\left(\frac{\text{maximum likelihood under model}}{\text{maximum likelihood under saturated model}}\right)$$

  describes lack of fit.
  The deviance, as defined by Nelder and Wedderburn (1972), is given by

$$D = 2\left[L(\widehat{\boldsymbol{\mu}}, \mathbf{y}) - L(\mathbf{y}, \mathbf{y})\right]$$

  where $L(\widehat{\boldsymbol{\mu}}, \mathbf{y})$ is the log-likelihood maximized over some vector of parameters and $L(\mathbf{y}, \mathbf{y})$ is the maximum likelihood achievable in the saturated model.
  As an example consider the form of the deviance for the Poisson distribution.
  Let $Y_1, Y_2, \ldots, Y_n$ be n independent Poisson random variables with $E(Y_i) = \mu_i$. The log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\mu}, \mathbf{y}) &= \log\left[\prod \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!}\right] \\ &= \sum y_i \log \widehat{\mu}_i - \sum \widehat{\mu}_i - \sum \log y_i! \end{aligned}$$

The deviance for a model with fitted values $\widehat{\mu}_i$ is

$$\begin{aligned} D &= 2\left[\sum y_i \log \widehat{\mu}_i - \sum \widehat{\mu}_i - \sum \log y_i! - \left\{\sum y_i \log y_i - \sum y_i - \sum \log y_i!\right\}\right] \\ &= 2\left[\sum y_i \log \frac{\widehat{\mu}_i}{y_i} + \sum (y_i - \widehat{\mu}_i)\right]. \end{aligned}$$

- The Wald Statistic

  If the model under consideration is formulated in terms of the constraints $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$ and $\mathbf{G} = \frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}|_{\boldsymbol{\mu}=\mathbf{y}}$ then the Wald statistic is

$$W = \mathbf{g}'(\mathbf{y})\left(\mathbf{G_y V_y G'_y}\right)^{-1}\mathbf{g}(\mathbf{y}).$$