UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Maximum likelihood estimation procedures for categorical data

by

René Ehlers

Submitted in partial fulfilment
of the requirements for the degree
Magister Scientiae (Mathematical Statistics)

in the
Faculty of Natural and Agricultural Sciences
University of Pretoria

July 2002

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor N.A.S. Crowther, for his guidance and motivation. I also wish to thank my parents for their continuous support and for making this study possible.

ABSTRACT

Maximum likelihood estimation procedures for categorical data

by

René Ehlers

Supervisor: Professor N.A.S. Crowther
Department of Statistics
University of Pretoria

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, namely the Newton-Raphson, Fisher scoring and EM algorithms are compared with a maximum likelihood estimation procedure under constraints. An exposition of the theory and application of the methods are given.

Chapter 1 gives a brief overview of the exponential family, the generalized linear model and measures of goodness of fit.

In Chapter 2 the theory of the Newton-Raphson, Fisher scoring and EM algorithms and the method of maximum likelihood estimation under constraints is discussed.
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 matrix of second order partial derivatives 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 is a very general iterative algorithm for ML estimation in incomplete data problems and is described in detail by Dempster, Laird and Rubin (1977). 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.
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 $\boldsymbol{\mu}$. If $\mathbf{Y}$ is a random vector with probability function belonging to the exponential family with $E(\mathbf{Y}) = \boldsymbol{\mu}$, then the ML estimate of $\boldsymbol{\mu}$ subject to the constraint $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$, is given by

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

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.
Comparative examples of all the above procedures are given in Chapter 2.

In Chapter 3 ML estimation of parameters for loglinear and logistic regression models is discussed. The results obtained by using the method under constraints are the same as those obtained by using the Newton-Raphson algorithm.

In Chapter 4 different patterns of symmetry in squared contingency tables are discussed and illustrated with an example from Agresti (1990). Results obtained are the same as the special cases considered in literature.

In Chapter 5 the method of ML estimation under constraints is used to determine ML estimates of cell probabilities in an incomplete contingency table for any loglinear model. It is assumed that the data are missing at random (MAR) and that the missing data mechanism is ignorable. It is shown that results are asymptotically the same as those obtained with the EM algorithm, the advantage being that the method under constraints is computationally less intensive.

CONTENTS

1 INTRODUCTION **1**

1.1 THE EXPONENTIAL FAMILY 2

1.2 COMPONENTS OF A GENERALIZED LINEAR MODEL 3

1.3 MEASURES OF GOODNESS OF FIT 4

2 MAXIMUM LIKELIHOOD ESTIMATION PROCEDURES **6**

2.1 THE NEWTON-RAPHSON ALGORITHM 6

2.2 THE FISHER SCORING ALGORITHM 9

2.3 IGNORABLE MISSING DATA MECHANISM 10

2.4 THE EM ALGORITHM 12

2.4.1 Theory of the EM Algorithm 12

2.4.2 The EM Algorithm for exponential families 13

2.5 A MAXIMUM LIKELIHOOD ESTIMATION PROCEDURE WHEN MODELLING IN TERMS OF CONSTRAINTS 16

3 CATEGORICAL DATA ANALYSIS **23**

3.1 LOGLINEAR ANALYSIS 23

3.1.1 The Model 23

3.1.2 Newton-Raphson algorithm for ML estimation 24

3.1.3 Maximum likelihood estimation under constraints 25

3.2 LOGISTIC REGRESSION 26

3.2.1 The Model 28

3.2.2 Newton-Raphson algorithm for ML estimation 30

3.2.3 Maximum likelihood estimation under constraints 31

4 SYMMETRY MODELS FOR SQUARE CONTINGENCY TABLES WITH ORDERED CATEGORIES 35

4.1 SYMMETRY MODEL 35

4.2 CONDITIONAL SYMMETRY 36

4.3 DIAGONALS-PARAMETER SYMMETRY 36

4.4 LINEAR DIAGONALS-PARAMETER SYMMETRY 37

4.5 ANOTHER LINEAR DIAGONALS-PARAMETER SYMMETRY MODEL 37

4.6 2-RATIOS-PARAMETER SYMMETRY 38

4.7 QUASI SYMMETRY 39

4.8 EXAMPLE 39

5 INCOMPLETE CONTINGENCY TABLES 42

5.1 ML ESTIMATION IN INCOMPLETE CONTINGENCY TABLES 42

5.1.1 The EM Algorithm 42

5.1.2 ML Estimation under constraints 45

5.2 LOGLINEAR MODELS FOR INCOMPLETE CONTINGENCY TABLES 49

5.2.1 The EM Algorithm 49

5.2.2 ML Estimation under constraints 49

5.3 CONCLUSION 50

6 REFERENCES 55

7 APPENDIX 56

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.