Cairo University
Faculty of Economics and Political Science
Department of Statistics

Mathematical Programming Approach to Multinomial Logistic Regression

Prepared by

Aya Anas Aly

Supervised by

Dr. Ramadan Hamed

Prof. of Statistics

Dept. of Statistics

Dr. Ali El-Hefnawy

Associate Prof. of Statistics

Dept. of Statistics

A Thesis Submitted to the Department of Statistics, Faculty of Economics and Political Science in Partial Fulfillment of the Requirements for the M.Sc. Degree in Statistics

2009

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Ramadan Hamed, who guided me throughout the work in the thesis despite of his responsibilities.

My deep gratitude goes to my second supervisor, Dr. Ali El-Hefnawy, who was supportive and helpful in each step of my research.

I would like also to mention how cooperative Dr. Hesham Abd-Allah was, especially in the simulation study presented in this thesis.

I would like also to thank all my professors and colleagues at the faculty of Economics and Political science for their help and encouragement.

My special thanks go to my family, specially my parents, for their care and support during my work in this thesis.

الإجازة

أجازت لجنة المناقشة هذه الرسالة للحصول على درجة الماجستير فى الاحصاء بتقدير/امتياز
بتاريخ ٢٠٠٩/٦/٣٠ بعد استيفاء جميع المتطلبات

اللجنة

الاسم	الدرجة العلمية	التوقيع
ا.د/ نادية مكاري جرجس	**أستاذ الإحصاء غير المتفرغ بكلية الاقتصاد و العلوم السياسية**	
ا.د/ عبد الفتاح محمد قنديل	**أستاذ الإحصاء وعميد كلية التجارة– جامعة بنها**	
ا.د/ رمضان حامد محمد	**أستاذ الإحصاء بكلية الاقتصاد و العلوم السياسية**	
ا.د/ على الحفناوى الدسوقى	**أستاذ الإحصاء المساعد بكلية الاقتصاد و العلوم السياسية**	

Mathematical Programming Approach to Multinomial Logistic Regression

Abstract

Mathematical programming methods can handle various statistical problems already solved by classical methods. In this study, the multinomial logistic regression problem is considered and a non-linear goal programming approach is proposed to estimate the parameters of the multinomial logistic regression model. The study suggests two non-linear goal programming models for this purpose. The performance of the suggested models is compared to that of the maximum likelihood approach using a simulation study. The comparison is based on three criteria: the bias of the estimated parameters, their mean squared error and the overall percentage of correct classification. One of the suggested models proved its superiority in terms of the overall percentage of correct classification. In addition, this model is useful when the outcome groups are of unequal importance with respect to the decision maker. A numerical example explaining the model application in such situations is also presented in this study.

Key Words

Multinomial Logistic Regression *Non-Linear Goal Programming*

Sequential Goal Programming *Monte Carlo Simulation*

Supervised by

Dr.Ramadan Hamed **Dr. Ali El-Hefnawy**

Prof. of Statistics *Associate Prof. of Statistics*

Dept. of Statistics *Dept. of Statistics*

A Thesis submitted to the Department of Statistics, Faculty of Economics and Political Science in Partial Fulfillment of the Requirements for the M.Sc. Degree in Statistics

2009

Table of Contents

List of Tables

List of Figures

Introduction

The multinomial logistic regression model is an extension of the binary logistic regression model. It is used to model the relationship between a categorical outcome variable with more than two categories and a set of covariates. Logistic regression modeling is now employed in various fields such as epidemiology, health policy, business, finance and criminology [13]. Mathematical programming (MP) methods can handle various statistical problems already solved by classical methods [3]. In this study, the multinomial logistic regression problem will be considered using a mathematical programming approach.

The application of MP to statistics was introduced by Charnes et al [7]. Charnes paper considered the linear regression problem. In addition to the application of MP in linear regression analysis, other applications exist in the discriminant and cluster analysis areas (as stated in [16]).

MP techniques have various advantages over the classical methods. For example, MP methods provide techniques to deal with multiple objectives. They can also deal with complex problem formulations with flexibility. Moreover, some MP methods, especially linear programming methods, lend themselves to sensitivity analysis [9]. In addition, MP models can deal with additional inequality or non-negative constraints on the model parameters with flexibility [3].

MP techniques are also efficient tools for solving models based on minimizing the sum of absolute errors. Previous research has shown that these methods can be attractive alternatives to the existing statistical methods if the data is highly skewed or outlier-contaminated [23].

Due to the various advantages of using the MP methods in dealing with statistical problems, this study proposes a non-linear goal programming approach to estimate the parameters of the multinomial logistic regression model. The new approach comes out with two models, namely, Model (A) and Model (B).

Model (A) is based on grouping the data into categories according to the values of the explanatory variables. This makes the application of the model possible if the raw data is not available to the researcher. Unlike Model (A), Model (B) uses the individual observations. A major advantage of Model (B) is its ability to give a

high percentage of correct classification in a certain specified outcome group if it is solved using sequential goal programming. This is considered a major advantage, as in many situations the outcome groups are of un-equal importance and the decision maker might be interested in obtaining a high percentage of correct classification in a certain outcome group [9].

The performance of the suggested models is compared to that of the maximum likelihood (ML) approach using a simulation study. The results of the simulation are based on 1000 replications. Three criteria are used in the comparison; the bias of the resulting estimators, their mean squared error and the overall percentage of correct classification.

Study Objectives

MP methods proved to be advantageous in dealing with various statistical problems. However, reviewing the available literature reveals that there is a lack of studies on the MP approaches to the logistic regression problem. Even the existing studies are interested in estimating the parameters of the binary logistic regression model. This makes the researcher interested in tackling the current topic and introducing a new MP approach for estimating the parameters of the multinomial logistic regression model. Thus, this study has the following objectives:

1-Presenting a new non-linear goal programming approach for estimating the parameters of the multinomial logistic regression model.

2-Comparing the performance of the proposed approach to that of the maximum likelihood estimation using a simulation study.

3-Studying the possibility of using the suggested approach in increasing the percentage of correct classification in a certain outcome group.

To achieve the previous objectives, this study is organized as follows:

Chapter 1:

Introduces the multinomial logistic regression model along with its formulation, method of estimating parameters and classification tables.

Chapter 2:

Presents the suggested models, their uses and advantages. In addition, a numerical example explaining the procedure of applying each of the suggested models is presented. The presented example will show the possibility of using Model (B) to increase the percentage of correct classification in a certain specified outcome group.

Chapter 3:

Shows the simulation design, data generation method and the software packages that are used to produce the results. In addition, this chapter reports the parameters and classification results.

***Chapter 4*:**

Gives the most important conclusions and some points for future work.

Commands Appendix:

Displays some of the SPSS and GAMS commands used in this study**.**

Chapter 1

Multinomial Logistic Regression

The use of logistic regression modeling has become increasingly popular during the past decade. This modeling approach is now employed in various fields such as epidemiology, health policy, business, finance, criminology as well as other fields [13]

As stated in [13], McFadden (1974) introduced the multinomial logistic regression model as a modification of the binary logistic regression model. A binary logistic regression model uses a binary outcome variable. However, a multinomial logistic regression model uses nominal outcome variables with more than two categories. One category of the outcome variable is set as the reference category and the probability of membership in the other categories is compared with the probability of membership in the reference category. The multinomial logistic regression model is also known as polychotomous or polytomous logistic regression model ([13] and [18]). In this study, the term multinomial will be used.

The first section of this chapter introduces the multinomial logistic regression model. Estimating the parameters of the multinomial logistic regression model is the issue of the second section of this chapter. In this section, the discussion is limited to the case of an outcome variable with three categories; however, generalization to more than three categories is a matter of notation more than of concept. In the third section of this chapter, the classification tables are considered.

1.1- The Multinomial Logistic Regression Model

Logistic regression analysis can be extended beyond the analysis of binary outcome variables to the analysis of nominal outcome variables with more than two categories. McFadden (1974) introduced this type of model and called it a discrete choice model (as stated in [13]). In the literature on logistic regression, this model is called a discrete choice model in business and econometric literature and multinomial, polychotomous or polytomous logistic regression model in the health and life sciences ([13] and [18]).

Logistic Regression has various applications including the examination of occupational choices, modes of transportation, consumer preferences and other categorical outcomes [2]. For example, people's occupational choices can be affected by their parents' occupations and their own education level. One can study the relationship between one's occupation choice on one hand and his education level and father's occupation on the other. The outcome variable in this problem is the occupational choices consisting of more than two categories. The explanatory variables in this example are the education level and parent's occupation.

Suppose that the nominal outcome variable Y has $k+1$ categories coded as 0, 1, 2, ….., k, $\mathbf{x}^T=(x_0, x_1,\ldots ,x_p)$; $x_0=1$ denotes a $1\times (p+1)$ observation vector on p covariates, and $\beta^T = (\beta_1^T, \beta_2^T, \ldots, \beta_k^T)$ denotes the vector of $k(p+1)$ unknown parameters, where $\beta_j^T = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp}); j = 1,2,\ldots,k$. In a multinomial logistic regression model, the categories 1,2,….,k are compared to the reference category 0. Thus, for an outcome variable with $k+1$ categories, the multinomial logistic regression model is given by [15]:

$$g_j(\mathbf{x}) = Ln\left[\frac{p(Y=j|\mathbf{x})}{p(Y=0|\mathbf{x})}\right] = \beta_j^T\mathbf{x} \quad ; \; j=1,2,\ldots..,k \tag{1.1}$$

This means that, for an outcome variable with $k+1$ categories, k equations are needed to describe the relationship between the outcome variable and the explanatory variables. More specifically, one equation for each category relative to the reference category [18].

Moreover, the conditional probability of each category of the outcome variable is given by [15]:

$$p(Y=j|\mathbf{x}) = \pi_j(\mathbf{x}) = \left[\frac{e^{\boldsymbol{\beta}_j^T\mathbf{x}}}{1+\sum_{s=1}^{k} e^{\boldsymbol{\beta}_s^T\mathbf{x}}}\right] \quad ; j=0,1,\ldots,k \text{ and } \boldsymbol{\beta}_0 = 0. \tag{1.2}$$

It can be noticed that when k=1, the multinomial logistic regression model will reduce to the binary logistic regression model [18].

1.2-Parameters Estimation

Major statistical packages use the method of maximum likelihood for estimating the parameters of the logistic regression model. The method of maximum likelihood produces values for the unknown parameters that maximize the probability of obtaining the observed dataset [13].

Suppose that the outcome variable of interest has three categories coded as 0, 1, and 2. Then, the logit functions $g_j(\mathbf{x}); j=1,2$ are given by:

$$\begin{aligned} g_1(\mathbf{x}) &= Ln\left[\frac{p(Y=1|\mathbf{x})}{p(Y=0|\mathbf{x})}\right] \\ &= \boldsymbol{\beta}_1^T\mathbf{x} \\ &= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots.. + \beta_{1p}x_p \end{aligned} \tag{1.3}$$

and

$$\begin{aligned} g_2(\mathbf{x}) &= Ln\left[\frac{p(Y=2|\mathbf{x})}{p(Y=0|\mathbf{x})}\right] \\ &= \boldsymbol{\beta}_2^T\mathbf{x} \\ &= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots.. + \beta_{2p}x_p \end{aligned} \tag{1.4}$$

And the three conditional probabilities are given by:

$$\pi_0(\mathbf{x}) = \left[\frac{1}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}}\right], \tag{1.5}$$

$$\pi_1(\mathbf{x}) = \left[\frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \right], \tag{1.6}$$

and

$$\pi_2(\mathbf{x}) = \left[\frac{e^{g_2(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \right] \tag{1.7}$$

To construct the likelihood function, three binary variables, Y_0, Y_1, and Y_2, are created as follows:

if Y=0 $\longrightarrow$ Y_0=1, Y_1=0 and Y_2 =0

if Y=1 $\longrightarrow$ Y_0=0, Y_1=1 and Y_2=0

if Y=2 $\longrightarrow$ Y_0=0, Y_1=0 and Y_2=1

It follows that the likelihood function for a sample of n independent observations is given by:

$$l(\beta) = \prod_{i=1}^{n} \left[\pi_0(\mathbf{x}_i)^{y_{oi}} \pi_1(\mathbf{x}_i)^{y_{1i}} \pi_2(\mathbf{x}_i)^{y_{2i}} \right] \tag{1.8}$$

And since $\sum_{j=0}^{2} y_{ji} = 1$ for each i, then the log likelihood function can be written as:

$$\begin{aligned} L(\beta) &= \ln[l(\beta)] \\ &= \sum_{i=1}^{n} \left\{ y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - \ln(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}) \right\} \end{aligned} \tag{1.9}$$

Afterwards, the likelihood equations can be obtained by making partial differentiation to $L(\beta)$ with respect to each of the 2(p+1) unknown parameters. The resulting partial derivatives are given by:

$$\frac{\partial L(\beta)}{\partial \beta_{jl}} = \sum_{i=1}^{n} x_{li} (y_{ji} - \pi_{ji}); \; j=1, 2 \text{ and } l=0, 1, \ldots, p \tag{1.10}$$

where

$\pi_{ji} = p(Y = j|\mathbf{x_i})$: is the conditional probability of the outcome variable given the covariate vector.

The maximum likelihood estimates of the unknown parameters can be obtained by setting the previous equations equal to zero and solving for β. However, obtaining the solution is not so straight forward as it requires some iterative computations. Thus, special software is required to solve for β and find the maximum likelihood estimates [13].

1.3-Classification Tables:

A classification table is an appealing way of summarizing the results of a fitted logistic regression model [13]. Classification procedures are useful for predicting an outcome variable given the levels of one or more explanatory variables [6]. For instance, in the occupational choices example, one is interested in predicting the person's occupational choice based on his educational level and father's occupation.

To create a classification table, one must first calculate the estimated logistic probabilities using equation (1.2). Afterwards, each case is classified into the category of *Y* for which it has the highest estimated probability. The classification table could then be created by cross classifying the actual group membership with the predicted group membership derived from the estimated logistic probabilities ([13] and [18]).