Multivariate Zero-Inflated Poisson Regression

A PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yang Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisor: Yang Li

June, 2017

Acknowledgements

First, I would like to thank my family. They have been supporting me throughout my graduate study from my initial application, through the challenges of graduate school, and the final dissertation process. I could not have finished this without their support and love.

This dissertation would not have been possible without my advisor Yang Li. I have had the pleasure of working with him. I have improved my research skill, learned how to do and talk about my research. I appreciate his help more than he knows.

I also would like to thank my committee members, Professor Bruce Peckham and Professor Richard Green. Both of them were willing to spend time discussing various facets of my research, and also offered suggestions that helped me along the way.

Dedication

Abstract

In this report, we develop a procedure to analyze the relationship between the observed multi-dimensional counts and a set of explanatory variables. The counts follow a multivariate Poisson distribution or a multivariate zero-inflated Poisson distribution. Maximum likelihood estimates (MLE) for the model parameters are obtained by the Newton-Raphson (NR) iteration and the expectation-maximization (EM) algorithm, respectively. In Newton-Raphson method, the first and second derivatives of the log-likelihood function are derived to carry out the numerical evaluation. Formulas using EM algorithm are also introduced. A comparison of the estimation performance is made from simulation studies.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables vi

List of Figures vii

1 Introduction 1

2 Multivariate Poisson Regression 5

2.1 Poisson Distribution and Poisson Regression 5

2.2 Multivariate Poisson distribution and Multivariate Poisson regression Model 6

3 Likelihood-based Estimation of Multivariate Poisson Regression 9

3.1 Optimization by Newton-Raphson (NR) Algorithm 10

3.2 Optimization by the Expectation-Maximization (EM) Algorithm 12

3.2.1 Introduction to the EM Algorithm 13

3.2.2 Estimate MPoi Regression by EM Algorithm 14

4 Zero-inflated Multivariate Poisson Distribution 17

5 EM Algorithm for MZIP Regression Models 19

6 Simulation Studies **23**

6.1 Simulations of MPoi Regression 23

6.2 Simulations of MZIP Regression 24

7 Conclusion and Discussion **30**

References **31**

List of Tables

6.1 MLE estimates using NR methods. The first number is the mean estimate from 100 simulations and the number in the parentheses is the bias of that estimate. 24
6.2 MLE estimates using EM methods. 24
6.3 The 3-D zero-inflated Poisson observations. 27
6.4 MLE estimates for β with bias using EM algorithm 27
6.5 MLE estimates for τ and λ_0 with bias using EM algorithm 29

List of Figures

6.1 The box plot for the estimates β and λ_0 obtained by NR algorithm with the true values 25
6.2 The box plot for the estimates β and λ_0 obtained by EM algorithm with the true values 26
6.3 Box plot for estimates τ, β and λ_0 obtained by EM algorithm. The true values are shown with black triangles. 28

Chapter 1

Introduction

In statistics, Poisson regression is known to have a log-linear form which can be used to model count data. Poisson regression assumes the response variable Y has a Poisson distribution, and the logarithm of its expected value can be modeled by a linear combination of some independent variables. Kocherlakota and Kocherlakota considered the extension to the two-dimensional case in which the response vector has the bivariate Poisson distribution [1]. Following this idea, we have studied multivariate Poisson (MPoi) regression and zero-inflated multivariate Poisson (MZIP) regression. The response vector is not limited to two components and the excess of zero counts is also taken into consideration.

Multivariate count data are usually modeled via multinomial models. When we meet small counts with a lot of zeros, the normal approximation is not adequate at all. Then multivariate Poisson models would be an attractive idea, which can be applied to multivariate count data, such as purchase of different products, different types of faults in manufacture process and sports data. Modeling multivariate count data is important in many disciplines. The MPoi distribution provides a generalization of univariate Poisson distribution in high dimensions in which random variables are intercorrelated. A way to construct an MPoi distribution is to use one common term z_0 (Johnson and Kotz 1969),

$$y_1 = z_1 + z_0, y_2 = z_2 + z_0, \ldots, y_m = z_m + z_0. \tag{1.1}$$

where the $z_1, z_2, \ldots, z_m$ and z_0 are $m + 1$ independent Poisson random variables with

respective event rate λ_1,..., λ_m and λ_0. The counts $\{y_1, y_2, \ldots, y_m\}$ are depended on each other and the covariance matrix of it would be

$$\begin{bmatrix} \lambda_1 & \lambda_0 & \ldots & \lambda_0 \\ \lambda_0 & \lambda_2 & \ldots & \lambda_0 \\ \vdots & \vdots & & \vdots \\ \lambda_0 & \lambda_0 & \ldots & \lambda_m \end{bmatrix}$$

The covariance matrix is specified through the rates of the multivariate Poisson distribution.

To illustrate the Multivariate Poisson regression, we can consider an example of market research, in which we study the effect of the location, weather, and incomes $\{x_2, x_3, x_4\}$ on the number of different kinds of customers in a restaurant. The response vector $\{y_1, y_2, y_3, y_4\}$ consists of the numbers of teenagers, women, men and aged people. The logarithm of the marginal expectation of Y_{ij} is related to x_{ir} by $\log(\mathrm{E}(Y_{ij})) = x_{i1} + x_{i2}\beta_{2j} + x_{i3}\beta_{3j} + x_{i4}\beta_{4j}$ for $j = 1, 2, 3, 4$ groups, $i = 1, 2, \ldots, n$ observations. The link functions are the log-link function and the regression coefficient $\vec{\beta_j}$ depends on the $\vec{Y_j}$. We select the MLE method to estimate λ_0 and the coefficients βs for the MPoi regression. Above example would be helpful for the prediction of future outcomes and improve the overall turnover.

In this paper, the first-order and second-order derivatives of the log-likelihood function have been derived, and then the computing could be carried out by the the Newton-Raphson (NR) algorithm. We also use the expectation-maximization (EM) algorithm as an alternative, which is quite useful for a large family of estimation problems with latent variables. We write R functions to fit the MPoi model using both algorithms. These functions will be pooled into a package which is designed to solve the calculation problems related to MPoi and MZIP regression and inference.

In addition, we study the multivariate zero-inflated Poisson (MZIP) regression. When the m-dimensional observation data contain large numbers of zeros, the MZIP model would provide a good fit. MZIP is often used in counting the defects of products in manufacturing processes that usually stay in a perfect state, in which defects are rarely observed. In the MZIP distribution, we assume that zero observations come from two different distributions: the MPoi distribution and a degenerate distribution at zero. Li et al. (1999) constructed a multivariate zero-inflated Poisson distribution,

which is a mixture of a degenerate on the zero point $(0, 0, \cdots, 0)$, m distributions each with univariate Poisson and $m-1$ zeros, and an m-dimensional Poisson. The resulting MZIP distribution is as follows:

$$\begin{aligned}
&(Y_1, Y_2, \cdots, Y_m) \\
&\quad \sim (0, 0, \cdots, 0) && \text{with probability } p_0 \\
&\quad \sim (\text{Poisson}(\lambda_1), 0, \cdots, 0) && \text{with probability } p_1 \\
&\quad \sim (0, \text{Poisson}(\lambda_2), \cdots, 0) && \text{with probability } p_2 \\
&\qquad\qquad\qquad\qquad \vdots \\
&\quad \sim (0, 0, \cdots, \text{Poisson}(\lambda_2)) && \text{with probability } p_m \\
&\quad \sim \text{multivariate Poisson } (\lambda_1, \lambda_2, \cdots, \lambda_m, \lambda_0) \\
&\qquad \text{with probability } p_{11}
\end{aligned} \tag{1.2}$$

where $p_0 + p_1 + p_2 + \cdots + p_m + p_{11} = 1$. The preceding model has $2(m+1)$ parameters. We simplify the model as a mix of two distributions, a point mass on 0 and an m-dimensional Poisson. Other patterns would be studied in the future. Based on the MPoi regression, we add one more link function for the zero-degenerate possibility. It is difficult to maximize this likelihood function, since its derivatives are not a simple function of covariates. In such cases, the EM algorithm gives an efficient method to maximize the likelihood, by repeatedly constructing a lower-bound of the likelihood function and then optimizing it.

The regression for jointly counting variables is more complicated, and the literature is rare. Our purpose for this article is to study MPoi and MZIP regression methods, respectively, and give an efficient way to estimate the covariates. Moreover, this article can serve as a rough manual for our MPoi and MZIP regression package. The rest of the paper is organized as follows:

- Chapter 2, illustrates the structure of the MPoi distribution and the regression model.
- Chapter 3, obtains the maximum likelihood estimates for MPoi regression coefficients by NR and EM algorithms, respectively.

- Chapter 4, gives the structure of multivariate Zero-Inflated Poisson (MZIP) distribution and regression model.
- Chapter 5, applies the EM algorithm to MZIP regression.
- Chapter 6, does simulations about MPoi and MZIP regressions. We applied NR and EM algorithms to simulation in R. Then the results are compared with the chosen parameters, and box-plots are shown and analyzed.

Chapter 2

Multivariate Poisson Regression

2.1 Poisson Distribution and Poisson Regression

The Poisson distribution expresses the probability of a given number of events occurring in a set of time, distance, or area. Moreover, these events occur with a known average rate and are independent of time. For instance, the number of customers visiting a restaurant in general follows a Poisson distribution.

Suppose Z is a random variable which follows a Poisson distribution as $Z \sim \text{Poisson}(\lambda)$. Here λ is called the event rate which represents the average number of events within an interval. The probability of observing k events is

$$P(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{2.1}$$

where $k = 0, 1, \ldots$. The Poisson distribution is a discrete probability distribution and (2.1) is its probability mass function. An important property of Poisson distribution is that the rate λ is equal to both the expectation and the variance of Y. That is, $\lambda = E(Y) = \text{Var}(Y)$.

Conversely, if we are interested in knowing the expected number or the rate of events occurring from observed count data, the Poisson regression method can be used by building a relationship between the expectation of the response variable Y and some explanatory variables $x = (x_1, x_2, \cdots, x_k)$. The Poisson regression is a special case of the generalized linear model,

$$g(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2.2}$$

where g is the link function and βs are unknown model parameters. In the Poisson regression model, the link function is the log function such that

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2.3}$$

Generally, Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. In the next step, the maximum likelihood estimators (MLE) for β are obtained by finding the values that maximize the log-likelihood.

2.2 Multivariate Poisson distribution and Multivariate Poisson regression Model

The m-variate Poisson can be expressed as $(Y_1, Y_2, \ldots, Y_m) \sim \text{MPoi}(\lambda_0, \lambda_1, \ldots, \lambda_m)$. Consider a specific case where $m = 2$. We say (Y_1, Y_2) follows a bivariate Poisson distribution with rates λ_0, λ_1, and λ_2 if their joint distribution function satisfies

$$f(y_1, y_2) = \exp(-\lambda_0 - \lambda_1 - \lambda_2) \sum_{j=0}^{\min(y_1, y2)} \frac{\lambda_1^{y_1 - j} \lambda_2^{y_2 - j} \lambda_0^j}{(y_1 - j)!(y_2 - j)! j!} \tag{2.4}$$

Here both y_1 and y_2 can take non-negative integer values. It is convenient to write

$$Y_1 = Z_0 + Z_1 \quad \text{and} \quad Y_2 = Z_0 + Z_2 \tag{2.5}$$

where Z_0, Z_1, and Z_2 follow the Poisson distributions with parameters λ_0, λ_1, and λ_2, respectively, and they are independent from each other. In other words, $Y_1 - Z_0$, $Y_2 - Z_0$, and Z_0 are independent Poisson random variables. Their joint distribution is the product of three Poisson probability functions,

$$f(z_0, y_1 - z_0, y_2 - z_0) = \exp(-\lambda_0 - \lambda_1 - \lambda_2) \frac{\lambda_0^{z_0}}{z_0!} \frac{\lambda_0^{y_1 - z_0}}{(y_1 - z_0)!} \frac{\lambda_0^{y_2 - z_0}}{(y_2 - z_0)!}\,. \tag{2.6}$$

From equation (2.5), we know that z_0 is always less than or equal to the minimum of y_1 and y_2. Therefore the bivariate distribution function of y_1 and y_2 can be obtained by summing over Z_0 from 0 to $\min(y_1, y_2)$ which gives (2.4).

We can derive more properties from (2.5). For example, λ_0 is the covariance between Y_1 and Y_2, $\text{Cov}(Y_1, Y_2)$. If $\lambda_0 \neq 0$, Y_1 and Y_2 are not independent of each other.

Moreover, the expectations of the marginal distribution are $\lambda_1' \equiv E(Y_1) = \lambda_0 + \lambda_1$ and $\lambda_2' \equiv E(Y_2) = \lambda_0 + \lambda_2$.

The bivariate Poisson regression model takes the form of

$$\log \lambda_1^{'} = \sum_{r=1}^{p} x_r \beta_{r1} \,, \tag{2.7}$$

$$\log \lambda_2^{'} = \sum_{r=1}^{p} x_r \beta_{r2} \,. \tag{2.8}$$

We assume that two response variables Y_1 and Y_2 are related to the same set of explanatory variables $(x_1, x_2, \ldots, x_p)$. This assumption will make the formulation easier but is not required.

Generally, for any $m \geq 2$, we can build a MPoi distribution in the following way. Let $Z_j \sim$ Poission(λ_j), where $j = 1, \ldots, m$ and $Z_0 \sim$ Poisson(λ_0) be independent random variables from each other. Consider the random vector $\mathbf{Y}$

$$\begin{aligned} Y_1 &= Z_1 + Z_0 \,, \\ Y_2 &= Z_2 + Z_0 \,, \\ &\cdots \\ Y_m &= Z_m + Z_0 \,. \end{aligned} \tag{2.9}$$

Then $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ is said to follow a MPoi distribution with $m+1$ parameters $(\lambda_0, \lambda_1, \ldots, \lambda_m)$, denoted as $\mathbf{Y} \sim$ MPoi$(\lambda_0, \lambda_1, \ldots, \lambda_m)$. The marginal distributions are also Poisson, that is, $Y_j \sim$ Poisson$(\lambda_j^{'})$, where $\lambda_j^{'} = \lambda_j + \lambda_0$. The covariance between any pair of random variables of Ys is λ_0 such that $\mathrm{Cov}(Y_i, Y_j) = \lambda_0$ if $i \neq j$.

The random vector $\mathbf{Y}$ follows the joint probability mass function

$$f(\mathbf{y}) = \exp\left(-\sum_{j=0}^{m} \lambda_j\right) \phi(y_1, y_2, \ldots, y_m) \,, \tag{2.10}$$

where

$$\phi(y_1, y_2, \ldots, y_m) = \sum_{z_0=0}^{\min(\mathbf{y})} \left(\prod_{j=1}^{m} \frac{\lambda_j^{y_j - z_0}}{(y_j - z_0)!}\right) \frac{\lambda_0^{z_0}}{z_0!} \,. \tag{2.11}$$

We define Poisson regression functions related to each random variable $\vec{Y_j}$ and p-1 predictors as follows. Correspondingly, there are p regression coefficient β for each

variable y_j

$$E(Y_j) = e^{\beta_{1j}+\sum_{r=2}^{p} x_r \beta_{rj}}, \quad j = 1, 2, \ldots, m \tag{2.12}$$

Then we have the following m link functions

$$\begin{aligned} \ln(\lambda_1^{'}) =& \beta_{11} + \sum_{r=2}^{p} x_r \beta_{r1} \\ \ln(\lambda_2^{'}) =& \beta_{12} + \sum_{r=2}^{p} x_r \beta_{r2} \\ & \vdots \\ \ln(\lambda_m^{'}) =& \beta_{1m} + \sum_{r=2}^{p} x_r \beta_{rm} \end{aligned} \tag{2.13}$$

Where $\lambda_j^{'} = \lambda_j + \lambda_0$, $j = (1, 2, \cdots, m)$. The Multivariate regression model (2.13) relates the logarithm of expectations of marginal distribution to the explanatory variables. The covariance λ_0 is a parameter in the model involving the regression. In the following parameter estimation, we use the MLE method to obtain the $m \times p$ unknown regression coefficients βs and an estimator of the covariance λ_0 which is developed along with the estimator of β.

Chapter 3

Likelihood-based Estimation of Multivariate Poisson Regression

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations by finding the parameter values that maximize the likelihood function. Under mild regularity conditions, MLE based estimation has some nice properties, such as asymptotic normality. Therefore, MLE is a popular estimation method in statistical analysis. In this chapter, we use the MLE method to estimate the parameters of the MPoi regression model.

Firstly, suppose there are n observations, each with m values. That is, the observed valued can be represented by an $n \times m$ matrix

$$Y = (\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n)^T = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix}$$

in which each row corresponds to one observation. These observations are related to other p predictor variables $(X_1, X_2, \ldots, X_p)$. In the framework of the MPoi regression model, the expectation of the observe matrix Y has a log-linear form

$$E[Y] = \exp(X \cdot \beta)\,, \tag{3.1}$$

where

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

is the predictor matrix where each row corresponds to one observation, and

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pm} \end{bmatrix}$$

is the parameter matrix. Each column of β is related to one of the m variables $(Y_1, Y_2, \dots, Y_m)$.

Secondly, the likelihood function of the data is used to find the value of parameters $\boldsymbol{\beta}$ and λ_0 that maximize the likelihood function (3.2). The following problems would be calculating the roots of first derivatives of (3.2). Since the MPoi model brings a lot of parameters which need to be estimated, we select a numerical method, Newton Raphson, to solve the equation system. The EM method as an alternative can improve the estimation efficiency. The details of that two methods are as follows.

$$\ln L(\vec{\beta}, \lambda_0) = -\sum_{i=1}^{n}\sum_{j=1}^{m} e^{\beta_{0j}+\sum_{r=1}^{p} x_{ir}\beta_{rj}} + n(m-1)\lambda_0 + \sum_{i=1}^{n} \ln \phi(\vec{y_i}) \tag{3.2}$$

3.1 Optimization by Newton-Raphson (NR) Algorithm

In this section, we use the Newton-Raphson (NR) method to estimate the parameters $\boldsymbol{\beta}$ and λ_0. In numerical analysis, the NR method is used for finding successively better approximations to the roots of a real-valued nonlinear function. For illustration, we start from a Taylor series expansion of the function $f(x)$ around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2f(x_0)}(x - x_0)^2 + \cdots \tag{3.3}$$

where x_0 is the initial guess for the root, f is the function, f' is the function's first derivative and so forth. When we have a proper initial guess, $x - x_0$ is small and only

the first two terms in (3.3) are significant to $f(x)$. By truncating away the high order terms, we have

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) \tag{3.4}$$

The NR iteration formula would be generalized as

$$\begin{aligned} x_1 &= x_0 - \frac{f'(x_0)}{f(x_0)} \\ &\vdots \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \end{aligned} \tag{3.5}$$

In order to find the maximum of the likelihood function, we can turn the question into finding the roots of its derivative. Using the NR method to solve equations, we define $V(\boldsymbol{\beta}, \lambda_0) = dl(\boldsymbol{\beta}, \lambda_0)/d\boldsymbol{\beta}$, and $J(\boldsymbol{\beta}, \lambda_0) = d^2 l(\boldsymbol{\beta}, \lambda_0)/d\boldsymbol{\beta}^2$ to be the first and second derivatives of the log-likelihood function, respectively. By plugging V and J into (3.5) we can get the following iteration procedure,

$$\begin{aligned} (\boldsymbol{\beta}^{(1)}, \lambda_0^{(1)}) &= (\boldsymbol{\beta}^{(0)}, \lambda_0^{(0)}) + J^{-1}(\boldsymbol{\beta}^{(0)}, \lambda_0^{(0)}) V(\boldsymbol{\beta}^{(0)}, \lambda_0^{(0)}) \\ &\vdots \\ (\boldsymbol{\beta}^{(n+1)}, \lambda_0^{(n+1)}) &= (\boldsymbol{\beta}^{(n)}, \lambda_0^{(n)}) + J^{-1}(\boldsymbol{\beta}^{(n)}, \lambda_0^{(n)}) V(\boldsymbol{\beta}^{(n)}, \lambda_0^{(n)}) \end{aligned} \tag{3.6}$$

where the vector $V(\boldsymbol{\beta}, \lambda_0)$ has the following $mp + 1$ terms

$$\begin{aligned} \frac{\partial l}{\partial \beta_{rj}} &= \sum_{i=1}^{n} x_{ir} e^{\beta_{0j} + \sum_{r=1}^{p} x_{ir}\beta_{rj}} (\phi_{ij}(1) - 1) \\ \frac{\partial l}{\partial \lambda_0} &= n(m-1) + \sum_{i=1}^{n} \left(\phi_{i111\ldots1}(1) - \sum_{j=1}^{m} \phi_{ij} \right) \end{aligned} \tag{3.7}$$

where $r = (1, 2, \ldots, p)$, $j = (1, 2, \ldots, m)$ and we define

$$\begin{aligned} \phi_{ij}(s) &= \frac{\phi(y_{i1}, \ldots, y_{ij} - s, \ldots, y_{im})}{\phi(y_{i1}, y_{i2}, \ldots, y_{im})} \\ \phi_{ij_1j_2}(s_1, s_2) &= \frac{\phi(y_{i1}, \ldots, y_{ij_1} - s_1, \ldots, y_{ij_2} - s_2, \ldots, y_{im})}{\phi(y_{i1}, y_{i2}, \ldots, y_{im})} \\ \phi_{i111\ldots1}(s) &= \frac{\phi(y_{i1} - s, y_{i2} - s, \ldots, y_{im} - s)}{\phi(y_{i1}, y_{i2}, \ldots, y_{im})} \\ \psi_{ij} &= \frac{\phi(y_{i1} - 1, y_{i2} - 1, \ldots, y_{ij} - 2, \ldots, y_{im} - 1)}{\phi(y_{i1}, y_{i2}, \ldots, y_{im})} \end{aligned} \tag{3.8}$$

On the other hand, $J(\boldsymbol{\beta}, \lambda_0)$ is an $(mp+1) \times (mp+1)$ matrix with elements

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \beta_{rj} \partial \beta_{kj}} &= \sum_{i=1}^{n} \lambda'_{ij} x_{ir} x_{ik} \big[\phi_{ij}(1) - 1 + \lambda'_{ij}[\phi_{ij}(2) - \phi^2_{ij}(1)]\big] \\
\frac{\partial^2 l}{\partial \beta_{rj_1} \partial \beta_{kj_2}} &= \sum_{i=1}^{n} \lambda'_{ij_1} \lambda'_{ij_2} x_{ir} x_{ik} [\phi_{ij_1 j_2}(1,1) - \phi_{ij_1}(1)\phi_{ij_2}(1)] \\
\frac{\partial^2 l}{\partial \beta_{rj} \partial \lambda_0} &= \sum_{i=1}^{n} \lambda'_{ij} x_{ir} \Big[\psi_{ij} - \sum_{k=1}^{m} \phi_{ikj}(1,1) - \phi_{ij}(1)\big[\phi_{111\ldots1}(1) - \sum_{k=1}^{m} \phi_{ik}(1)\big]\Big] \qquad (3.9) \\
\frac{\partial^2 l}{\partial \lambda_0} &= \sum_{i=1}^{n} \Big[\phi_{i111,\ldots,1}(2) - 2\sum_{j=1}^{m} \psi_{ij} + \sum_{j=1}^{m} \phi_{ij}(2) \\
&\qquad + 2\sum_{j_1=1}^{m}\sum_{j_2=1}^{m} \phi_{ij_1 j_2}(1,1) - \big[\phi_{i111,\ldots,1}(1) - \sum_{j=1}^{m} \phi_{ij}(1)\big]^2\Big]
\end{aligned}
$$

where $r, k = (1, 2, \ldots, p)$ and $j = (1, 2, \ldots, m)$. The estimates of $\boldsymbol{\beta}$ and λ_0 can be updated iteratively using (3.6)-(3.9). The procedure will stop when the convergence criterion ($|\max(\beta^{t+1}, \lambda_0^{t+1}) - \max(\beta^t, \lambda_0^t)| \leq \delta$) is satisfied for some pre-specified tolerance value δ.

3.2 Optimization by the Expectation-Maximization (EM) Algorithm

Although the NR algorithm converges fast, it only works well when the starting point is at the neighborhood of the maximum. Furthermore, if the Hessian matrix is almost singular, it will hardly reach a stable solution. To overcome these difficulties, we also apply the expectation-maximization (EM) algorithm as an alternative that is a powerful method for finding the maximum value of a log-likelihood function when missing values are present in the data. In general, the EM algorithm is performed by calculating the pseudo-values based on the current estimates obtained from the n-th iteration, and using that pseudo-values to maximize the lower bound on the log-likelihood to obtain a new setting of estimates, and iterating between the above two steps until some converge criterion is satisfied.

3.2.1 Introduction to the EM Algorithm

Given a data set $(x_1, x_2, \cdots, x_n)$, we are interested to fit the parameters of a model $p(x|\theta)$ to that data, where its log-likelihood function is

$$l(\theta) = \sum_{i=1}^{n} \log p(x; \theta)\,. \tag{3.10}$$

Sometimes it may be hard to find the maximum likelihood estimates of θ explicitly. Moreover, it is often the case that the complete data contain a latent variable z which is not explicitly observed. The likelihood function, however, depends on the observations of both x and z simultaneously. In such a setting, the EM algorithm can be utilized as an efficient method for finding the MLE of θ. Let us rewrite the likelihood function as

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log \left(\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \right) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z^{(i)}} Q_i(z^i) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^i)} \right) \\ &\geq \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^i) \log \left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^i)} \right) \end{aligned} \tag{3.11}$$

where $Q(z)$ is the probability mass function of z such that $\sum_z Q(z) = 1$, and $Q(z) \geq 0$. We use Jensen's inequality to get the relationship between line 2 and line 3 in (3.11). In probability, Jensen's inequality states that, if f is a convex function and Z is a random variable, then

$$E[f(Z)] \geq f(E[Z])\,. \tag{3.12}$$

Define $g(z) = \left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^i)} \right)$. The expectation of $\log(g(z))$ with a $Q(z)$ distribution is

$$E_z \left[\log \left(\frac{p(x, z; \theta)}{Q(z)} \right) \right] = \sum_z Q(z) \log \left(\frac{p(x, z; \theta)}{Q(z)} \right) \tag{3.13}$$

while the log-expectation of $g(z)$ is

$$f \left(E \left[\frac{p(x, z; \theta)}{Q(z)} \right] \right) = \log \left(\sum_z Q(z) \frac{p(x, z; \theta)}{Q(z)} \right) \tag{3.14}$$

Since $\log(\cdot)$ is a convex function, we can immediately obtain the inequality in (3.11). Instead of maximizing $l(\theta)$, we repeatedly construct a lower-bound for it (E-step), and then optimize that lower-bound (M-step). We simply set the $Q(z)$ to be the posterior distribution of z given observation x and the current estimate of θ. The EM algorithm proceeds as follows.

$$\begin{aligned} &\mathrm{E}-\mathrm{step} \\ &\qquad Q(z) := p(z|x;\theta) \\ &\mathrm{M}-\mathrm{step} \\ &\qquad \theta := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} E_{z^{(i)}} \left[\log \left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q(z^{(i)})} \right) \right] \end{aligned} \tag{3.15}$$

In the M-step, we maximize the expectation (3.13) with respect to the parameters θ to obtain a new setting; In the E-step, we set a new lower bound of $l(\theta)$ using the new estimates. This procedure will be repeatly carried out until the pre-specified convergent criterion is satisfied.

3.2.2 Estimate MPoi Regression by EM Algorithm

In the MPoi scheme, we consider the missing value to be $Z_0 \sim \mathrm{Poi}(\lambda_0)$ which is a latent variable related to every variate as shown in section 2. To find the conditional distribution, we illustrate the complete model first. The complete data set for the ith observation is $\mathbf{y}_i = (z_{i0}, y_{i1}, y_{i2}, \ldots, y_{im})$. Its probability mass function is based on the independent assumption of its components,

$$f(z_{i0}, y_{i1}, \ldots, y_{im}, \mid \beta, \lambda_0) = \frac{e^{-\lambda_0} \lambda_0^{z_{i0}}}{z_{i0}!} \prod_{j=1}^{m} \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}-z_{i0}}}{(y_{ij}-z_{i0})!} \tag{3.16}$$

The E-step gives the conditional function of z_{i0} given $(z_{i0}, y_{i1}, y_{i2}, \ldots, y_{im})$ and tth estimates of (β,λ_0). In the M-step, we maximize the conditional expectation of the

log-likelihood function of the above independent MPoi distribution.

$$\begin{aligned} &\text{E-step:} \\ &\qquad f(z_{i0} \mid \vec{y}_i, \vec{\beta}_i^{(t)}, \lambda_{i0}^{(t)}) \\ &\text{M-step:} \\ &\qquad (\lambda_0^{(t+1)}, \vec{\beta}^{(t+1)}) = \operatorname{argmax} \sum_{i=1}^{n} \mathrm{E}_{z_{i0}|\vec{y}_i, \vec{\beta}_i^{(t)}, \lambda_0^{(t)}} [\ln f(\vec{y}_i, z_{i0} \mid \vec{\beta}, \lambda_0)] \end{aligned} \tag{3.17}$$

where the simplified conditional mean of $\ln f(\vec{y}_i, z_{i0} \mid \vec{\beta}, \lambda_0)$ is

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} \big[&- e^{\beta_{0j} + \sum_{r=2}^{p} x_{ir}\beta_{rj}} + (y_{ij} - \mathrm{E}[z_{i0}]) \ln[e^{\beta_{0j} + \sum_{r=2}^{p} x_{ir}\beta_{rj}} - \lambda_0] \\ &+ (m-1)\lambda_0 + \mathrm{E}[z_{i0}] \ln \lambda_0 \big] \end{aligned} \tag{3.18}$$

in which the conditional mean of z_{i0} is

$$\begin{aligned} \mathrm{E}_{z_{i0}|\vec{y}_i, \vec{\beta}_i^{(t)}, \lambda_0^{(t)}} [z_{i0}] &= \sum_{z_{i0}=1}^{\min(\vec{y}_i)} z_{i0} \frac{f(\vec{y}_i, z_{i0} \mid \vec{\lambda}_i)}{f(\vec{y}_i \mid \vec{\lambda}_i)} \\ &= \lambda_0^{(t)} \frac{f(\vec{y}_i - 1 \mid \vec{\beta}^{(t)}, \lambda_0^{(t)})}{f(\vec{y}_i \mid \vec{\beta}^{(t)}, \lambda_0^{(t)})} \end{aligned} \tag{3.19}$$

In (3.19), the conditional mean of z_{i0} given the t-th estimates is the pseudo- value. In (3.18), the conditional expectation of the log-likelihood function of the complete model is the lower bound of the MLE. In the M-step, the maximum estimates of the parameters can be obtained by the NR algorithm. Its first and second order derivatives are less complicated than the initial likelihood function (3.2).

For the t-th estimation in the M-step, the first derivative V has the following elements

$$\begin{aligned} \frac{\partial l(\vec{y}_i, z_{i0})}{\partial \beta_{rj}^{(t)}} &= \sum_{i=1}^{n} \left(y_{ij} - \lambda_{ij} - E_i^{(t-1)} \right) \frac{x_{ir} \lambda_{ij}^{'}}{\lambda_{ij}} \\ \frac{\partial l(\vec{y}_i, z_{i0})}{\partial \lambda_0^{(i)}} &= (m-1)n + \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{m} \frac{E_i^{(t-1)} - y_{ij}}{\lambda_{ij}} \right) + \frac{E_i^{(t-1)}}{\lambda_0} \right] \end{aligned} \tag{3.20}$$

where $j = (1, 2, \cdots, m)$, $r = (1, 2, \cdots, p)$ and vector V has $mp + 1$ elements. The

elements in the second-order derivative matrix J are

$$\begin{aligned}
&\frac{\partial^2 l(\vec{y}_i, z_{i0})}{\partial \beta_{r_1 j_1}^{(t)} \partial \beta_{r_2 j_2}^{(t)}} = 0, \qquad\qquad j_1 \neq j_2 \\
&\frac{\partial^2 l(\vec{y}_i, z_{i0})}{\partial \beta_{r_1 j}^{(t)} \partial \beta_{r_2 j}^{(t)}} = -\sum_{i=1}^{n} x_{ir_1} x_{ir_2} \lambda_{ij}^{'} \left[1 + \frac{(y_{ij} - E_i^{(t-1)})\lambda_0}{\lambda_{ij}^2} \right] \\
&\frac{\partial^2 l(\vec{y}_i, z_{i0})}{\partial \beta_{r_1 j_1}^{(t)} \partial \lambda_0^{(t)}} = \sum_{i=1}^{n} x_{ir} (y_{ij} - E_i^{(t-1)}) \frac{\lambda_{ij}^{'}}{\lambda_{ij}^2} \\
&\frac{\partial^2 l(\vec{y}_i, z_{i0})}{\partial^2 \lambda_0^{(t)}} = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{m} \frac{E_i^{(t-1)} - y_{ij}}{\lambda_{ij}^2} \right) - \frac{E_i^{(t-1)}}{\lambda_0^2} \right]
\end{aligned} \tag{3.21}$$

Chapter 4

Zero-inflated Multivariate Poisson Distribution

Observations of counts in which the number of zero events is higher than predicted by a Poisson model may be modeled by a zero-inflated Poission (ZIP) distribution. The ZIP distribution employs two components. The first component is governed by a degenerate distribution which is localized at 0. The second component is governed by a traditional Poisson distribution that generates non-negative counts including zeros. If Y is a random variable following a ZIP distribution, then

$$Y = \begin{cases} \text{Degenerate}(0) \text{ with probability} \varphi \\ \sim \text{Poi}(\lambda) \text{ with probability} 1 - \varphi \end{cases} \tag{4.1}$$

If the observed counts are not univariate, we need to consider the multivariate distributions. If an excess of zero events occurs, the multivariate zero-inflated Poisson (MZIP) model should be considered instead. There are many ways to construct a MZIP model as Kim (1999) demonstrated. Here we selected the one which follows a mixture of a multidimensional degenerate distribution at point zero and a traditional MPoi distribution as follows. Let Y be a random variable following a MZIP distribution,

$$Y = \begin{cases} (0, 0, \ldots, 0) \text{ with probability } \varphi \\ \sim \text{MPoi}(y_1, y_2, \ldots, y_m) \text{ with probability } 1 - \varphi \end{cases}, \tag{4.2}$$

where the observation follows a degenerate distribution at $(0, 0, \ldots, 0)$ with probability φ and a MPoi distribution with probability $1 - \varphi$ where $\varphi \in [0, 1]$. It is a reasonable model, since the MZIP distribution is mainly used for situations in which most defect counts are 0. The probability distribution function of Y is

$$\begin{aligned} f(y_1, y_2, \ldots, y_m \mid \varphi, \lambda) = &\left(\varphi + (1-\varphi)e^{-\sum_{j=0}^{m}\lambda_j}\right)^{I(\vec{y}=0)} \\ &\times \left((1-\varphi)e^{-\sum_{j=0}^{m}\lambda_j} \sum_{z_0=0}^{min(\vec{y})} \left(\frac{\lambda_j^{y_j - z_0}}{(y_j - z_0)!}\frac{\lambda_0^{z_0}}{z_0!}\right)\right)^{1-I(\vec{y}=0)} \end{aligned} \tag{4.3}$$

If $\vec{y} = (0, 0, \ldots, 0)$, the indicator function $I(\vec{y}) = 1$ and the probability mass function is the first line of (4.3). Otherwise, it is the second line. Moreover, the expectation of MZIP is

$$\begin{aligned} E(y_1) =& (1-\varphi)(\lambda_1 + \lambda_0) \\ &\vdots \\ E(y_m) =& (1-\varphi)(\lambda_m + \lambda_0) \end{aligned} \tag{4.4}$$

From the previous chapter, we know that there are m log-link functions in the MPoi regression. For the MZIP model, there is an extra link function $\text{logit}(\varphi)$ for the Bernoulli probability φ,

$$\begin{aligned} \ln \lambda_j' &= \sum_{r=1}^{p_1} x_r \beta_{rj}, \quad j = (1, 2, \cdots, m) \\ \text{logit}(\varphi) &= \ln\left(\frac{\varphi}{1-\varphi}\right) = \sum_{r=1}^{p_2} g_r \tau_r \end{aligned} \tag{4.5}$$

where λ and φ are not necessarily dependent on the same covariates. We have p_1 parameter β's for each λ and p_2 parameter τ's for φ. The MLE method can be carried out to find the estimation of all parameters in a regression model, which will be introduced in the next chapter.

In summary, in the MZIP model, there are two possible states, the perfect-state and the non-perfect state. If the observed counts are all zeros, it is possible the system is in either the perfect or the non-perfect state. However, if the observed counts are not all zeros, we know the system must be in the non-perfect state.

Chapter 5

EM Algorithm for MZIP Regression Models

The estimations for the MZIP regression coefficient matrix β and vector τ are also obtained by the MLE method. For $n_1 + n_2$ observations following MZIP distribution, the log-likelihood function would be

$$\begin{aligned}\ln L(\tau, \beta, \lambda_0) = &\sum_{i=1}^{n_1} \ln \left[\frac{e^{\sum g\tau}}{1 + e^{\sum g\tau}} + \frac{1}{1 + e^{\sum g\tau}} \exp\left((m-1)\lambda_0 - \sum_{j=0}^{m} e^{\sum_{r=1}^{P_1} x_{ir}\beta_{rj}} \right) \right] \\ &+ \sum_{i=1}^{n_2} \left[\ln \frac{1}{1 + e^{\sum g\tau}} + (m-1)\lambda_0 - \sum_{j=0}^{m} \sum_{r=1}^{P_1} x_{ir}\beta_{rj} + \ln \phi(\vec{y}_i) \right]\end{aligned} \tag{5.1}$$

where we assume n_1 is the number of observations $\vec{y} = (0, 0, \ldots, 0)$ and n_2 is the number of observations $\vec{y} \neq (0, 0, \ldots, 0)$. For the first term of function (5.1), we cannot get a simple first-order derivative. The computational difficulties make it impossible to solve the maximum of this likelihood function by the NR method. Hence, we utilize the EM algorithm to maximize the likelihood.

Following the ideas of the EM algorithm in section 3.2, we choose latent a variable w which is a two-point distribution: $w = 1$ the system is in perfect state(the observation follows degenerate distribution at $(0, 0, \cdots, 0)$); $w = 0$ the system is in non-perfect state (the observation follows MPoi distribution). Then the joint distribution function

of $(w, \vec{y})$ would be

$$f(w, \vec{y} \mid \varphi, \vec{\lambda}) = \begin{cases} \varphi^w \left[(1-\varphi)e^{-\sum \lambda_j}\right]^{1-w}, & \vec{y} = (0, 0, \ldots, 0) \\ (1-\varphi)e^{-\sum \lambda_j}\phi(\vec{y}), & \vec{y} \neq (0, 0, \ldots, 0) \end{cases} \tag{5.2}$$

where we assume the w and $\vec{y}$ are not independent. The EM method can be organized as follows. Given arbitrary initial values for $(\varphi^{(0)}, \vec{\beta}^{(0)}, \lambda_0^{(0)})$, we can find the conditional function and the mean of the variable w in the E step, and get the MLE for the joint distribution to update the estimates. We repeat the EM-method until the convergence conditions are satisfied.

E-step:

$$f(w \mid \vec{y}, \varphi^{(t)}, \beta^{(n)}, \lambda_0^{(t)})$$

M-step: (5.3)

$$(\varphi^{(t+1)}, \vec{\beta}^{(t+1)}, \lambda_0^{(t+1)}) = \operatorname{argmax} \sum_{i=1}^{n} \mathrm{E}_{w|\vec{y},\varphi^{(t)},\beta^{(n)},\lambda_0^{(t)}} [\ln f(\varphi, \beta, \lambda_0 \mid w, \vec{y})]$$

In the $(t+1)$-th E-step, we want to find the conditional distribution of w given the data and the t-th estimates for β, λ_0, and τ. The condition distribution of w is

$$f(w \mid \vec{y}, \varphi, \phi) = \frac{f(w, \vec{y} \mid \varphi, \phi)}{f(\vec{y} \mid \varphi, \phi)} \tag{5.4}$$

For convenience, we can organize the conditional distribution of w as two kinds of distributions given different observations: when the random variables of $\vec{y}$ are all zero, w is a Bernoulli distribution; when the random variables are not all zero, the system can only be in the non-perfect state, where the w is a degenerate distribution at 0

$$w \mid (\vec{y}, \varphi, \phi) \sim \begin{cases} \text{Bernoulli}(p), & \vec{y} = (0, 0, \ldots, 0) \\ \text{Degenerate}(0), & \vec{y} \neq (0, 0, \ldots, 0) \end{cases} \tag{5.5}$$

where the Bernoulli random variable has the possibility of $p = \frac{\varphi}{\varphi + (1-\varphi)\exp(-\sum \lambda_j)}$ with the system in the perfect state. The conditional expectation can be derived from the distribution

$$\mathrm{E}_{w_i|(\vec{y_i},\varphi_i,\beta_i,\lambda_{i0})}(w_i) = \begin{cases} \frac{\varphi_i}{\varphi_i + (1-\varphi_i)\exp(-\sum \lambda_{ij})}, & \vec{y_i} = (0, 0, \ldots, 0) \\ 0, & \vec{y_i} \neq (0, 0, \ldots, 0) \end{cases} \tag{5.6}$$

In the $(t+1)$-th M-step, we find the maximum value of expectation of the log-likelihood function for the above joint distribution (5.2). The log-likelihood function is

$$\begin{aligned} lnL(\tau,\beta,\lambda_0) &= \sum_{i=1}^{n_1} w_i \ln\varphi_i + (1-w_i)\big(\ln(1-\varphi_i) - \sum_{j=0}^{m}\lambda_{ij}\big) + \sum_{i=1}^{n_2} ln(1-\varphi_i) + \ln f(\vec{y}_i \mid \lambda_i) \\ &= \sum_{i=1}^{n_1}\left[w_i \sum_{r=1}^{p_2} g_{ir}\tau_r - \ln(1+e^{\sum_{r=1}^{p_2} g_{ir}\tau_r}) + (1-w_i)\left[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_1} x_{ir}\beta_{rj}} \right] \right] \\ &+ \sum_{i=1}^{n_2}\Big[-ln(1+e^{\sum_{r=1}^{p_2} g_{ir}\tau_r}) + lnf\left(\vec{y}_i \mid \lambda_0,\beta\right) \Big] \end{aligned} \tag{5.7}$$

where the $f(\vec{y}_i \mid \lambda_0,\beta)$ is the MPoi distribution function (2.10). The NR method is applied to find the maximum $(t+1)$-th estimates. For convenience, We rewrite the above log-likelihood as the sum of two functions L_1 and L_2

$$\begin{aligned} L_1(\tau) &= \sum_{i=1}^{n_1}\left(w_i \sum_{r=1}^{p_2} g_{ir}\tau_r \right) - \sum_{i=1}^{n_1+n_2} ln(1+e^{\sum_{r=1}^{p_2} g_{ir}\tau_r}) \\ L_2(\beta,\lambda_0) &= \sum_{i=1}^{n_1}(1-w_i)\left[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_1} x_{ir}\beta_{rj}} \right] + \sum_{i=1}^{n_2} lnf(\vec{y}_i \mid \lambda_0,\beta) \end{aligned} \tag{5.8}$$

L_1 is a function of τ and L_2 is a function of β and λ_0. By the NR method (3.6), we should derive the first derivatives V and second derivatives J of the likelihood function (5.7), where

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \qquad J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}$$

Vectors V_1 and V_2 are the first derivatives of L_1 and L_2, respectively. They have following

items

$$
\begin{aligned}
\frac{\partial \ln \mathrm{E}(L_1)}{\partial \tau_{r_2}} &= \sum_{i=1}^{n_1} \mathrm{E}[w_i] g_{ir_2} - \sum_{i=1}^{n_1+n_2} g_{ir_2}\varphi_i \\
\frac{\partial \ln \mathrm{E}(L_2)}{\partial \beta_{r_1 j}} &= \sum_{i=1}^{n_1} (\mathrm{E}[w_i]-1) x_{ir} \lambda_{ij}^{'} + \sum_{i=1}^{n_2} x_{ir}\lambda_{ij}^{'}(\phi_{ij}(1)-1) \\
\frac{\partial \ln \mathrm{E}(L_2)}{\partial \lambda_0} &= \sum_{i=1}^{n_1} (1-\mathrm{E}[w_i])(m-1) + n_2(m-1) + \sum_{i=1}^{n_2}\big(\phi_{i111\ldots1}(1) - \sum_{j=1}^{m}\phi_{ij}\big)
\end{aligned}
\tag{5.9}
$$

$$r_1 = (1,2,\ldots,p_1), r_2 = (1,2,\ldots,p_2), j = (1,2,\ldots,m)$$

The Hessian matrices J_1 and J_2 are the second derivatives of L_1 and L_2, respectively, who have following items

$$
\begin{aligned}
\frac{\partial^2 \ln \mathrm{E}(L_1)}{\partial \tau_{r_2}\partial \tau_{k_2}} &= \sum_{i=1}^{n_1+n_2} g_{ir_2} g_{ik_2} \varphi_i(\varphi_i - 1) \\
\frac{\partial^2 \log L}{\partial \beta_{r_1 j}\partial \beta_{k_1 j}} &= \sum_{i=1}^{n_1} (\mathrm{E}[w_i]-1) x_{r_1 j} x_{k_1 j} \lambda_{ij}^{'} + \sum_{i=1}^{n_2} \lambda_{ij}^{'} x_{ir_1} x_{ik_1} \big[\phi_{ij}(1) - 1 + \lambda_{ij}^{'}[\phi_{ij}(2) - \phi_{ij}^2(1)]\big] \\
\frac{\partial^2 \log L}{\partial \beta_{r_1 j_1}\partial \beta_{k_1 j_2}} &= \sum_{i=1}^{n_2} \lambda_{ij_1}^{'}\lambda_{ij_2}^{'} x_{ir_1} x_{ik_1}[\phi_{ij_1 j_2}(1,1) - \phi_{ij_1}(1)\phi_{ij_2}(1)] \\
\frac{\partial^2 \log L}{\partial \beta_{r_1 j}\partial \lambda_0} &= \sum_{i=1}^{n_2} \lambda_{ij}^{'} x_{ir_1}\Big[\psi_{ij} - \sum_{k=1}^{m}\phi_{ikj}(1,1) - \phi_{ij}(1)\big[\phi_{111\ldots1}(1) - \sum_{k=1}^{m}\phi_{ik}(1)\big]\Big] \\
\frac{\partial^2 \log L}{\partial \lambda_0^2} &= \sum_{i=1}^{n_2}\Big[\phi_{i111,\ldots,1}(2) - 2\sum_{j=1}^{m}\psi_{ij} + \sum_{j=1}^{m}\phi_{ij}(2) \\
&\quad + 2\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\phi_{ij_1 j_2}(1,1) - \big[\phi_{i111,\ldots,1}(1) - \sum_{j=1}^{m}\phi_{ij}(1)\big]^2\Big] \\
&\qquad r_1, k_1 = (1,2,\ldots,p_1), r_2, k_2 = (1,2,\ldots,p_2), j = (1,2,\ldots,m)
\end{aligned}
\tag{5.10}
$$

Where the ϕ and ψ are functions in (3.8). With the NR method, we repeatedly update the estimates until they satisfy certain criteria. After obtaining the $(t+1)$-th estimates, we check the convergence criterion first. If it is satisfied, we can stop; if not, we continue to do the EM algorithm.

Chapter 6

Simulation Studies

6.1 Simulations of MPoi Regression

The simulation studies are conducted in R to study the performance of the NR and EM methods when optimizing the likelihood function. We consider the bias and the interquartile interval. The parameters used for simulating the data set $Y_{100\times 4}$ (100 observations and 4 correlated variables for each observation) are arbitrarily chosen as follows.

$$\beta = \begin{bmatrix} 0.1 & 0.2 & 0.30 & 0.12 \\ 0.6 & 0.5 & 0.60 & 0.72 \\ 0.6 & 0.4 & 0.33 & 0.45 \\ 0.4 & 0.82 & 0.78 & 0.25 \end{bmatrix}, \qquad \lambda_0 = 0.89$$

where there are one common covariate λ_0 and 16 β parameters. We repeat the simulations 100 times in R to get 100 different count samples.

The numerical results are shown in Table 6.1 and Table 6.2 for NR and EM methods, respectively. In these two tables, the mean values of the MLE estimates are shown along with their biases. We find that the NR method works well when the starting point is around the neighborhood of the real parameter values. We first get a proper starting point from several trials and do the iteration starting from that point using both algorithms. Figures 6.1 and Figure 6.2 show similar results using box plots.

We can see that the biases of estimates obtained by the EM method are in general smaller than those from the NR method. Furthermore, since the NR method highly

j	β_{1j}(b)	β_{2j}(b)	β_{3j}(b)	β_{4j}(b)	λ_0
1	-0.162(0.262)	0.079(0.121)	0.160(0.140)	-0.232(0.352)	0.838(0.052)
2	0.676(0.077)	0.489(0.011)	0.616(0.016)	1.022(0.302)	
3	0.800(0.200)	0.513(0.113)	0.412(0.082)	0.545(0.095)	
4	0.480(0.079)	0.855(0.035)	0.890(0.110)	0.306(0.056)	

Table 6.1: MLE estimates using NR methods. The first number is the mean estimate from 100 simulations and the number in the parentheses is the bias of that estimate.

j	β_{1j}(b)	β_{2j}(b)	β_{3j}(b)	β_{4j}(b)	λ_0
1	0.087(0.013)	0.179(0.021)	0.282(0.018)	0.083(0.037)	0.915(0.025)
2	0.614(0.014)	0.510(0.010)	0.616(0.016)	0.747(0.027)	
3	0.600(0.000)	0.426(0.026)	0.327(0.003)	0.437(0.013)	
4	0.404(0.004)	0.814(0.006)	0.809(0.029)	0.289(0.039)	

Table 6.2: MLE estimates using EM methods.

depend on the choice of the starting point with a longer computing time, we are in favor of the EM method when fitting the MPoi regression model.

6.2 Simulations of MZIP Regression

The simulation studies about MZIP regression are also conducted in R, where we aim to validate the method described in Chapter 5. In each simulation, the first thing is to generate multivariate zero-inflated observations. Here we simulate 100 3-dimensional y's, and the date set look like the one shown in Table 6.3.

Based on the link functions (4.5), we need the Poisson parameters τ, β, and the covariance λ_0 to generate the desired data sets. The parameters we used are as following

$$\beta = \begin{bmatrix} 0.15 & 0.25 & -0.31 \\ 0.36 & 0.20 & 0.36 \\ 0.20 & 0.24 & -0.30 \\ -0.25 & -0.72 & 0.48 \end{bmatrix}, \quad \tau = [1.75, -0.98, 0.83, -1.21], \quad \lambda_0 = 0.089 \qquad (6.1)$$

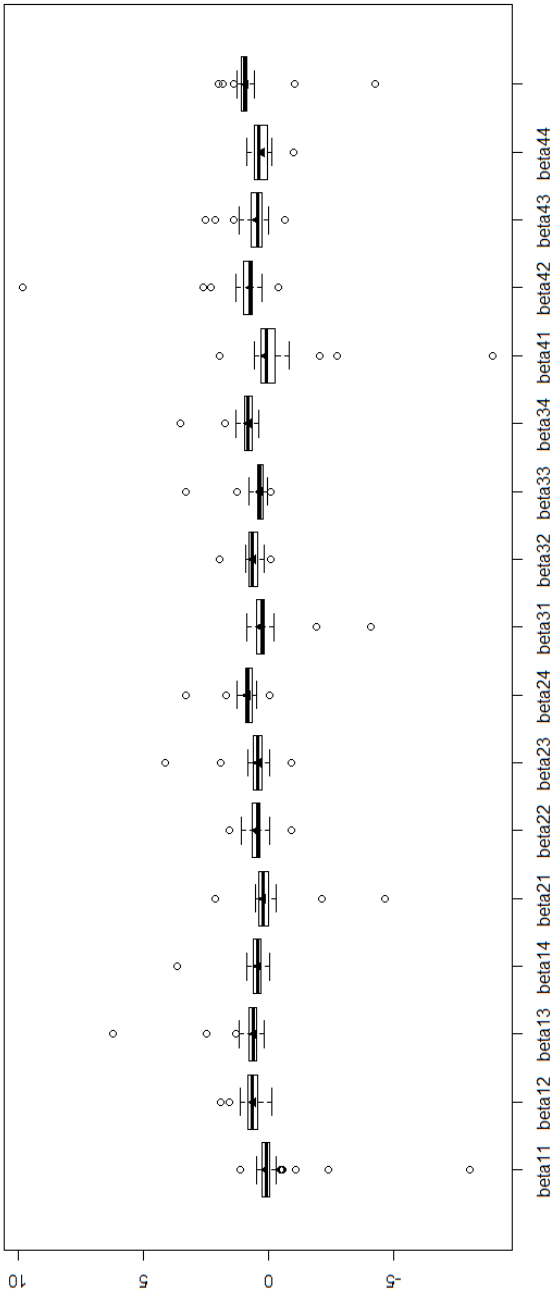

Figure 6.1: The box plot for the estimates β and λ_0 obtained by NR algorithm with the true values

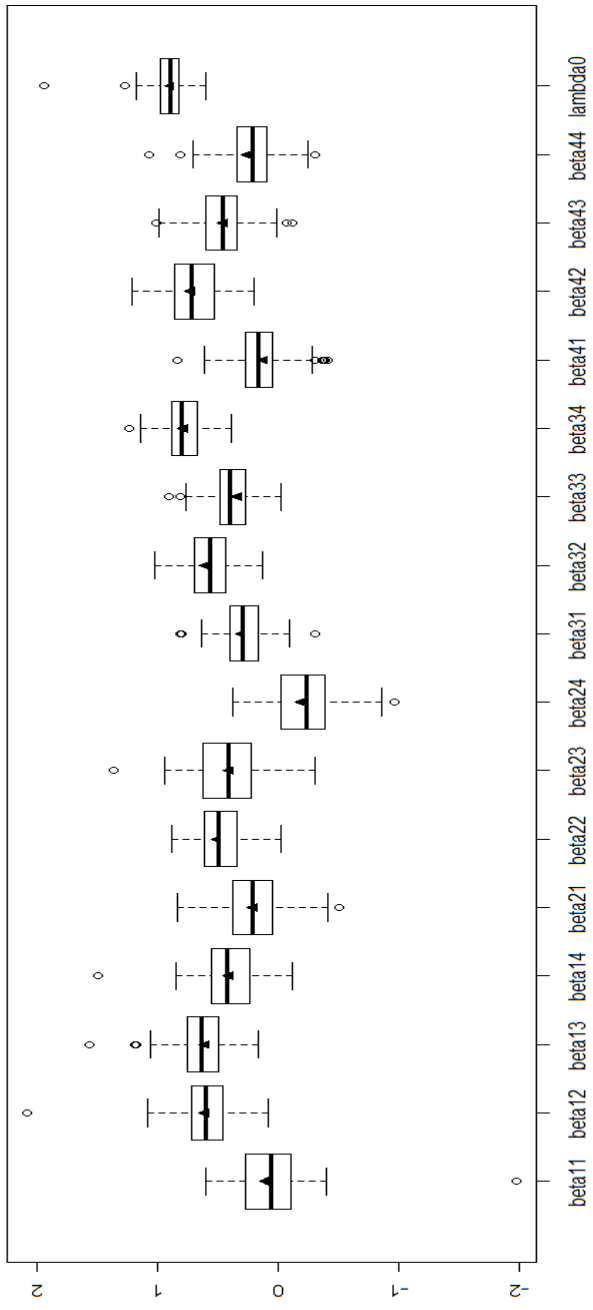

Figure 6.2: The box plot for the estimates β and λ_0 obtained by EM algorithm with the true values

(y_1, y_2, y_3)	(y_1, y_2, y_3)	(y_1, y_2, y_3)	(y_1, y_2, y_3)
(0,0,0)	(0,0,0)	(1,0,0)	(1,0,3)
(2,1,2)	(0,0,0)	(0,0,0)	(0,0,0)
(0,0,0)	(1,2,1)	(0,0,0)	(0,0,0)
(0,0,0)	(1,0,0)	(1,1,0)	(0,0,0)
(0,1,2)	(0,2,0)	(0,0,0)	(0,0,0)
(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)
⋮	⋮	⋮	⋮
(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)
(0,0,0)	(0,0,0)	(0,1,0)	(1,0,1)
(0,0,1)	(0,0,0)	(0,0,0)	(2,1,1)

Table 6.3: The 3-D zero-inflated Poisson observations.

where there are 4 τs and 4 βs in each link function. We conduct the simulations 25 times based on the same β, λ_0 and τ. The results are shown in Table 6.4, Table 6.5, and Figure 6.3. Table 6.4 and 6.5 give the mean of 25 MLE estimates with bias obtained by the EM algorithm. Figure 6.3 displays a box plot for 25 estimates. It can be seen that most of the 17 true parameters fall within the interquartile range, which verifies the validity of the algorithm introduced in Chapter 5. The biases, however, are not as small as the ones in Table 6.1 and Table 6.2, most likely due to the increased complexity of the model.

j	1	2	3
β_{j1}	0.116(0.034)	-0.051(0.300)	-0.564(0.254)
β_{j2}	0.512(0.152)	0.474(0.274)	0.391(0.030)
β_{j3}	0.260(0.06)	0.219(0.05)	-0.162(0.138)
β_{j4}	-0.25(0.18)	-0.580(0.140)	0.520(0.040)

Table 6.4: MLE estimates for β with bias using EM algorithm

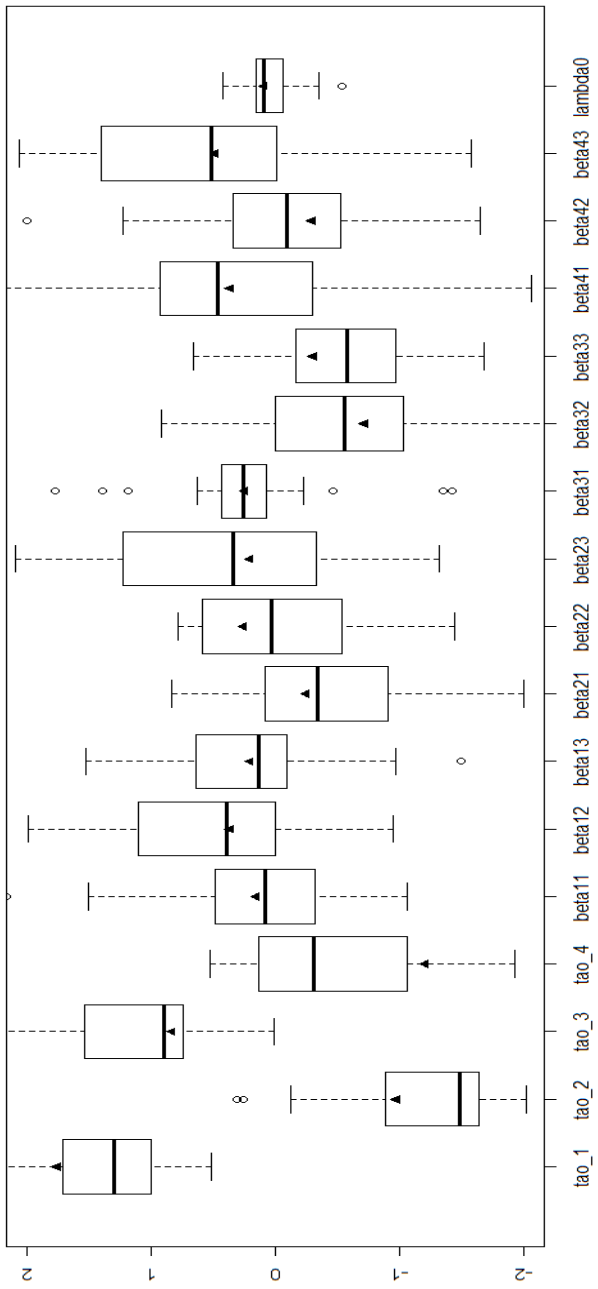

Figure 6.3: Box plot for estimates τ, β and λ_0 obtained by EM algorithm. The true values are shown with black triangles.

τ_1	τ_2	τ_3	τ_4	λ_0
1.365(0.135)	-1.164(0.184)	1.165(0.335)	-0.530(0.67)	0.01435371(0.075)

Table 6.5: MLE estimates for τ and λ_0 with bias using EM algorithm

Chapter 7

Conclusion and Discussion

In this paper, we introduce the MPoi and MZIP distributions and the corresponding regression models. Both NR and EM algorithms are used to optimize the MLE equations for MPoi regression. Detailed formulas for carrying out these two estimating methods are given. The EM method in general gives a better result than the NR method. It can accurately estimate the regression coefficients, while the NR method could fail to work due to the dependency on the starting point. We also provide the EM procedure for the MZIP regression which can efficiently estimate the regression coefficients. In this project, I have pooled all written R functions related to MPoi and MZIP regressions into a R package, which could help people fit multivariate counts to the MPoi or MZIP model conveniently.

On the other hand, there are still some problems in this research project. First, we assume that the multivariate Poisson model has a single common covariance term. In real applications, this assumption may be too restrictive since it assumes all pairs of variables have the same covariance. It may not be true. Second, our multivariate zero-inflated Poisson model is not applicable to every zero-inflated Poisson case. We only studied the mixture distribution of multivariate Poisson and a multivariate degenerate zero distribution. Therefore, this model is only suitable for the multivariate count data with a lot of zeros counts like $(0, 0, \ldots 0)$. If only part of the variables happen to be zero-inflated, but others are not, then our MZIP model may not be adequate. Overall, the increased number of variables brings a lot of problems in calculation and analysis. This may be a research topic in the future.

References

[1] Subrahmaniam Kocherlakota and Kathleen Kocherlakota. Regression in the bivariate poisson distribution. 2001.

[2] Dimitris Karlis, Ioannis Ntzoufras, et al. Bivariate poisson and diagonal inflated bivariate poisson regression models in r. *Journal of Statistical Software*, 14(10):1–36, 2005.

[3] Henry Teicher. On the multivariate poisson distribution. *Scandinavian Actuarial Journal*, 1954(1):1–9, 1954.

[4] Sotirios Loukas and H Papageorgiou. On a trivariate poisson distribution. *Applications of Mathematics*, 36(6):432–439, 1991.

[5] K Adamids and S Loukas. Ml estimation in the bivariate poisson distribution in the presence of missing values via the em algorithm. *Journal of statistical computation and simulation*, 50(3-4):163–172, 1994.

[6] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[7] Chin-Shang Li, Jye-Chyi Lu, Jinho Park, Kyungmoo Kim, Paul A Brinkley, and John P Peterson. Multivariate zero-inflated poisson models and their applications. *Technometrics*, 41(1):29–38, 1999.

[8] Yin Liu and Guo-Liang Tian. Type i multivariate zero-inflated poisson distribution with applications. *Computational Statistics & Data Analysis*, 83:200–222, 2015.

[9] ME Ghitany, D Karlis, DK Al-Mutairi, and FA Al-Awadhi. An em algorithm for multivariate mixed poisson regression models and its application. *Applied Mathematical Sciences*, 6(137):6843–6856, 2012.

[10] Dimitris Karlis. An em algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.

[11] Rajibul Mian and Sudhir Paul. Estimation for zero-inflated over-dispersed count data model with missing response. *Statistics in Medicine*, 35(30):5603–5624, 2016.

[12] Yiwen Zhang, Hua Zhou, Jin Zhou, and Wei Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, (just-accepted):1–37, 2016.

[13] Felix Famoye and Karan P Singh. Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4(1):117–130, 2006.

[14] Jean-Paul Fox. Multivariate zero-inflated modeling with latent predictors: Modeling feedback behavior. *Computational statistics & data analysis*, 68:361–374, 2013.

[15] Chunjiao Dong, Stephen H Richards, David B Clarke, Xuemei Zhou, and Zhuanglin Ma. Examining signalized intersection crash frequency using multivariate zero-inflated poisson regression. *Safety science*, 70:63–69, 2014.

[16] Kazutomo Kawamura. The structure of multivariate poisson distribution. *Kodai Mathematical Journal*, 2(3):337–345, 1979.