Developing A Testing Procedure for the Parameters of the Bivariate Geometric Distribution

A THESIS PAPER

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

MD FITRAT HOSSAIN

DR. MUNNI BEGUM - ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

May 2016
Abstract

THESIS: Developing A Testing Procedure for the Parameters of Bivariate Geometric Distribution

STUDENT: Md. Fitrat Hossain

DEGREE: Master of Science

COLLEGE: Sciences and Humanities

DATE: MAY 2016

PAGES: 33

Bivariate geometric distribution is an extension to the univariate geometric distribution that models the number of trials to get the first success. Thus a bivariate distribution can be considered as a model for number of trials to obtain two different but related events for the first time. Many statisticians have studied different forms of bivariate geometric distribution. In this thesis, we considered the form which is given by Phatak and Sreehari (1981). We estimated the parameters using maximum likelihood estimation. We derived the deviances as the goodness of fit statistics for testing the parameters corresponding to reduced model, generalized linear model (GLM) and the deviance difference to compare two models in order to determine which model fits the data well. To determine the efficiency of our deviances, we simulated data using computer software R. We found that our deviance for the reduced model with pair of parameters of the bivariate geometric distribution worked well.
Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Munni Begum for her continuous support and help on my thesis study. This work would not be possible without her patience, motivation, enthusiasm and immense knowledge. Her guidance helped me a lot all the time through this thesis. Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Yayuan Xiao and Dr. Michael Karls for their encouragement, insightful comments and patience. I am grateful to all my classmates for their kind supports. Last but not least, I would like to thank my family and friends for supporting me throughout my life.

Md Fitrat Hossain

May 2016
Contents

1 Introduction 1

2 Literature Review 1

3 Methodology 2

3.1 Univariate Geometric Distribution 3

3.2 Bivariate Geometric Distribution 4

3.2.1 Marginal distribution 4

3.2.2 Conditional distribution 5

3.2.3 Exponential Form 5

3.3 Maximum Likelihood Estimation 6

3.3.1 Estimation of Parameters of Saturated Model 6

3.3.2 Estimation of Parameters of Reduced Model (Consisting One Pair \((q_1, q_2)\)) 7

3.3.3 Estimation of Parameters of Distribution Based On GLM 9

3.4 Hypothesis Testing 11

3.4.1 Deviance for Reduced Model (Consisting One Pair \((q_1, q_2)\)) 12

3.4.2 Deviance Based On GLM 14

3.4.3 Difference Between Two Deviances 15

4 Data Simulation and Analysis 17

4.1 Deviance Checking for Reduced Model 17

4.2 Deviance Checking for GLM 21

4.3 Difference in Deviance Checking 22

5 Conclusion 23

6 References 23
1 Introduction

In the case of studying systems with several components which have different types of failures such as twin engines of an airplane or the paired organ in a human body, the bivariate geometric distribution can be applied. In the univariate geometric distribution, we focus on the occurrence of failure of one component of the system, ignoring the events corresponding to the other related component of the system, where in the bivariate geometric distribution we take into account the events corresponding to the other related components of the system as well. Bivariate geometric distributions have increasingly important roles in various fields, including reliability and survival analysis. In this thesis, we considered the probability mass function of the bivariate geometric distribution which was proposed by Phatak and Sreehari (1981). We estimated the parameters of bivariate geometric distribution using the maximum likelihood estimation corresponding to a saturated model where we considered that each observed value comes from a distinct bivariate geometric distribution with different parameters, a reduced model where we considered only one pair of parameters, and the generalized linear model (GLM). We derived the deviances as goodness of fit to perform hypothesis testing corresponding to the parameters of reduced model, GLM and deviance difference to compare two related models which fits the data well.

2 Literature Review

The geometric distribution can be applied to survival or reliability analysis or discrete time analysis for follow up, or panel data where occurrence or no-occurrence of an event are reported at different times. In case the of the bivariate geometric distribution, we consider the occurrence of an event taking into ac-
count the occurrence of corresponding other events. Phatak and Sreehari (1981) have provided a form of the bivariate geometric distribution which is employed in this thesis. In their paper they introduce a form of probability mass function which take into consideration three different types of events. There are other forms which can be seen in Nair and Nair (1988), Hawkes (1972), Arnold et al. (1992) and Sreehari and Vasudeva (2012). Basu and Dhar (1995) proposed a bivariate geometric model which is analogous to bivariate exponential model developed by Marshal and Olkin (1967). Characterization results are developed by Sun and Basu (1995); Sreehari (2005) and Sreehari and Vasudeva (2012).

In (2013) Omey and Minkova considered the bivariate geometric distribution with negative correlation coefficient and analyze some properties, probability generating function, probability mass function, moments and tail probabilities. Krishna and Pundir (2009), studied the plausibility of a bivariate geometric distribution as a reliability model. They found the maximum likelihood estimators and Bayes estimators of parameters and various reliability characteristics. They compared these estimators using Monte-Carlo simulation.

In this thesis, we estimate the parameters of bivariate geometric distribution for the saturated model, reduced model and generalized linear model. We also derived the deviance as the goodness of fit statistic to test parameters corresponding to reduced model, generalized linear model and deviance difference to compare two related models in order to determine which model fits the data well.

3 Methodology

In this section we show the probability mass function (pmf) of univariate and bivariate geometric distribution and how we will get the pmf of bivariate geomet-
ric distribution using geometric distribution and negative binomial distribution. We use the maximum likelihood estimation procedure to estimate parameters of the bivariate geometric distribution under a saturated model, reduced model where we consider one pair of parameters \((q_1, q_2)\), and generalized linear model (GLM) using an appropriate link function. At the end of this section, we derive the form of deviance as the goodness of fit statistic to test hypotheses corresponding to reduced model, and generalized linear model. We also derive the form of the difference in deviances corresponding to two reduced models to test which model fits the data well.

### 3.1 Univariate Geometric Distribution

The probability mass function (pmf) of a random variable \(Y\) which follows a geometric distribution with probability of success \(p\) can be written as,

\[
P(Y = y) = p(1 - p)^y, \quad y = 0, 1, 2, \ldots; \quad 0 < p < 1, q = 1 - p, 0 < q < 1
\]

and the moment generating function can be given by,

\[
M_Y(t) = \frac{p}{1 - qe^t}
\]

The mean and variance of this distribution are

\[
E(Y) = \mu_Y = \frac{1 - p}{p} = \frac{q}{p} \quad \text{and} \quad Var(Y) = \frac{1 - p}{p^2} = \frac{q}{p^2}
\]

An extension of the univariate geometric distribution is the bivariate geometric distribution which is discussed in the next subsection.
3.2 Bivariate Geometric Distribution

For the bivariate geometric distribution, we can employ the joint distribution based on marginal and conditional distributions which was introduced by Phatak and Sreehari (1981) where they considered a process from which the units could be classified as good, marginal and bad with probabilities \( q_1, q_2 \) and \( q_3 = (1 - q_1 - q_2) \) respectively. They proposed the probability mass function of observing the first bad unit after several good and marginal units are passed as follows:

\[
(Y_1 = y_1, Y_2 = y_2) = \binom{y_1 + y_2}{y_1} q_1^{y_1} q_2^{y_2} (1-q_1-q_2), y_1, y_2 = 0, 1, 2, \ldots; 0 < q_1+q_2 < 1
\]  

(1)

Here, \( Y_1 \) and \( Y_2 \) respectively denote the number of good units and the number of marginal units before the first bad unit is observed.

3.2.1 Marginal distribution

The marginal distributions of \( Y_1 \) and \( Y_2 \) are both geometric distributions with probability of success \( \left( \frac{1 - q_1 - q_2}{1 - q_2} \right) \) and \( \left( \frac{1 - q_1 - q_2}{1 - q_1} \right) \).

The marginal distribution of \( Y_1 \) can be written as follows,

\[
P(Y_1 = y_1) = \left( \frac{1 - q_1 - q_2}{1 - q_2} \right) \left( \frac{q_1}{1 - q_2} \right)^{y_1}, y_1 = 0, 1, 2, \ldots  \]  

(2)

Similarly the marginal distribution of \( Y_2 \) can be written as follows,

\[
P(Y_2 = y_2) = \left( \frac{1 - q_1 - q_2}{1 - q_1} \right) \left( \frac{q_2}{1 - q_1} \right)^{y_2}, y_1 = 0, 1, 2, \ldots  \]  

(3)
3.2.2 Conditional distribution

The conditional distribution of $Y_2$ given $Y_1$ is

$$P(Y_2 = y_2|Y_1 = y_1) = \binom{y_1 + y_2}{y_2} q_2^{y_2} (1 - q_1)^{y_1+1}, y_1, y_2 = 0, 1, 2, ...$$  (4)

Again the conditional distribution of $Y_1$ given $Y_2$ can be written as follows,

$$P(Y_1 = y_1|Y_2 = y_2) = \binom{y_1 + y_2}{y_1} q_1^{y_1} (1 - q_2)^{y_2+1}, y_1, y_2 = 0, 1, 2, ...$$  (5)

It can be shown that the multiplication of the marginal distribution of $Y_1$ which is expressed in equation (2) and the conditional the conditional distribution of $Y_2$ given $Y_1$ which is expressed in equation (3) gives the bivariate geometric distribution expressed in equation (1). Similarly, by multiplying the marginal distribution of $Y_1$ which is expressed in equation (3) and the conditional the conditional distribution of $Y_2$ given $Y_1$ which is expressed in equation (4) gives the bivariate geometric distribution expressed in equation (1).

3.2.3 Exponential Form

The exponential form of (1) can be written as:

$$P(Y_1 = y_1, Y_2 = y_2) = \exp \left[ \ln \left( \binom{y_1 + y_2}{y_1} q_1^{y_1} q_2^{y_2} (1 - q_1 - q_2) \right) \right]$$

$$= \exp \left[ \ln \left( \frac{(y_1 + y_2)!}{y_1!y_2!} \right) + y_1 \ln q_1 + y_2 \ln q_2 + \ln(1 - q_1 - q_2) \right]$$

$$= \exp \left[ y_1 \ln q_1 + y_2 \ln q_2 + \ln(1 - q_1 - q_2) + \ln (y_1 + y_2)! - \ln y_1! - \ln y_2! \right]$$  (6)

This leads us to the exponential family representation of the bivariate geometric distribution which can be used to derive sufficient statistics for the
generalized linear model based on this distribution.

### 3.3 Maximum Likelihood Estimation

#### 3.3.1 Estimation of Parameters of Saturated Model

Let \( Y_1, ..., Y_n \) be independent random vectors each having bivariate geometric distribution with different pairs of parameters \((q_{1i}, q_{2i})\) for \( i = 1, 2, ..., n \).

Hence the likelihood function based on these random vectors can be written as follows using (6):

\[
l = \sum_{i=1}^{n} \left[ y_{1i} \ln q_{1i} + q_{2i} \ln q_{2i} + \ln(1 - q_{1i} - q_{2i}) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \tag{7}
\]

The likelihood function of the conditional distribution of \( Y_2 \) given \( Y_1 \) can be written as follows using (4):

\[
l = \sum_{i=1}^{n} \left[ y_{2i} \ln q_{2i} + (y_{1i} + 1) \ln (1 - q_{2i}) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \tag{8}
\]

Differentiating (7) with respect to \( q_{2i} \) and setting it equal to zero, we get,

\[
\frac{dl}{dq_{2i}} = \frac{y_{2i}}{q_{2i}} + \frac{y_{1i} + 1}{1 - q_{2i}}(-1) = 0
\]

\[\Rightarrow \quad \frac{y_{2i}}{q_{2i}} = \frac{y_{1i} + 1}{1 - q_{2i}} \]

\[\Rightarrow \quad y_{2i} - q_{2i}y_{2i} = y_{1i}q_{2i} + q_{2i} \]

\[\Rightarrow \quad y_{2i} = q_{2i}(y_{1i} + y_{2i} + 1) \]

\[\Rightarrow \quad q_{2i} = \frac{y_{2i}}{y_{1i} + y_{2i} + 1} \tag{9}
\]

Now let us consider the likelihood function of the marginal distribution of
\[ l = \sum_{i=1}^{n} \left[ \ln(1 - q_{1i} - q_{2i}) - \ln(1 - q_{2i}) + y_{1i} \ln q_{1i} - y_{1i} \ln(1 - q_{2i}) \right] \]  

(10)

Differentiating (8) with respect to \( q_{1i} \) and setting it equal to zero, we get,

\[ \frac{dl}{dq_{1i}} = \frac{1}{1 - q_{1i} - q_{2i}} (-1) + \frac{y_{1i}}{q_{1i}} = 0 \]

\[ \Rightarrow \frac{y_{1i}}{q_{1i}} = \frac{1}{1 - q_{1i} - q_{2i}} \]

\[ \Rightarrow y_{1i} = y_{1i} q_{1i} - y_{1i} q_{2i} = q_{1i} \]

\[ \Rightarrow q_{1i}(y_{1i} + 1) = y_{1i} - y_{1i} q_{1i} \]

\[ \Rightarrow \hat{q}_{1i} = \frac{y_{1i} - y_{1i} q_{2i}}{y_{1i} + 1} = \frac{y_{1i}(1 - q_{2i})}{y_{1i} + 1} \]

\[ \Rightarrow \hat{q}_{1i} = \frac{y_{1i}(1 - q_{2i})}{y_{1i} + 1} \left[ \text{From equation(3)} \right] \]

\[ = \frac{(y_{1i} + 1)(y_{1i} + q_{2i} + 1)}{y_{1i} + q_{2i} + 1} \]

\[ = \frac{y_{1i}}{y_{1i} + q_{2i} + 1} \]

Here, \( \hat{q}_{1i} \) and \( \hat{q}_{2i} \) are the maximum likelihood estimate of \( q_{1i} \) and \( q_{2i} \) respectively.

### 3.3.2 Estimation of Parameters of Reduced Model (Consisting One Pair \( (q_1, q_2) \))

Now, let \( Y_1, Y_2, ..., Y_n \) are random vectors each having bivariate geometric distribution with same pair of parameters \( q_1 \) and \( q_2 \).

Hence, the likelihood function based on these random vectors can be written
as follows using equation (7) with \( q_{1i} = q_1 \) and \( q_{2i} = q_2 \) for all \( i = 1, 2, ..., n \):

\[
l = \sum_{i=1}^{i=n} \left[ y_{1i} \ln q_1 + y_{2i} \ln q_2 + \ln(1 - q_1 - q_2) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \quad (12)
\]

The likelihood function of the conditional distribution of \( Y_2 \) given \( Y_1 \) can be written as follows using equation (8) with \( q_{1i} = q_1 \) and \( q_{2i} = q_2 \) for all \( i = 1, 2, ..., n \):

\[
l = \sum_{i=1}^{i=n} \left[ y_{2i} \ln q_2 + \ln Y_1 + 1)(1 - q_2) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \quad (13)
\]

Differentiating (13) with respect to \( q_2 \) and setting it equal to zero, we get,

\[
\frac{dl}{dq_2} = \sum_{i=1}^{i=n} \frac{y_{2i}}{q_2} + \frac{\sum_{i=1}^{i=n} y_{1i} + n}{1 - q_2} (-1) = 0
\]

\[
\Rightarrow \sum_{i=1}^{i=n} y_{2i} - q_2 \sum_{i=1}^{i=n} y_{2i} = q_2 \sum_{i=1}^{i=n} y_{1i} + nq_2
\]

\[
\Rightarrow q_2 \sum_{i=1}^{i=n} y_{2i} + q_2 \sum_{i=1}^{i=n} y_{1i} + nq_2 = \sum_{i=1}^{i=n} y_{2i}
\]

\[
\Rightarrow \hat{q}_2 = \frac{\sum_{i=1}^{i=n} y_{2i}}{\sum_{i=1}^{i=n} y_{1i} + \sum_{i=1}^{i=n} y_{2i} + n} = \frac{\overline{y}_2}{\overline{y}_1 + \overline{y}_2 + 1}
\]

The likelihood function of the marginal distribution of \( Y_1 \) is

\[
l = \sum_{i=1}^{i=n} \left[ \ln(1 - q_1 - q_2) - \ln(1 - q_2) + y_{1i} \ln q_1 - y_{1i} \ln(1 - q_2) \right] \quad (15)
\]
Differentiating (15) with respect to $q_1$ and setting it equal to zero we get,

\[
\frac{dl}{dq_1} = \frac{n}{1 - q_1 - q_2} (-1) + \sum_{i=1}^{i=n} y_{1i} = 0
\]

\[
\Rightarrow \sum_{i=1}^{i=n} y_{1i} = \frac{n}{1 - q_1 - q_2}
\]

\[
\Rightarrow \sum_{i=1}^{i=n} y_{1i} - q_1 \sum_{i=1}^{i=n} y_{1i} - q_2 \sum_{i=1}^{i=n} y_{1i} = nq_1
\]

\[
\Rightarrow q_1 \left( \sum_{i=1}^{i=n} y_{1i} + n \right) = \sum_{i=1}^{i=n} y_{1i} (1 - q_2)
\]

\[
\Rightarrow \hat{q}_1 = \frac{\sum_{i=1}^{i=n} y_{1i} (1 - q_2)}{\sum_{i=1}^{i=n} y_{1i} + n}
\]

\[
= \frac{\sum_{i=1}^{i=n} y_{1i} \left( 1 - \frac{\sum_{i=1}^{i=n} y_{2i}}{\sum_{i=1}^{i=n} y_{1i} + n} \right)}{\sum_{i=1}^{i=n} y_{1i} + n}
\]

\[
= \frac{\sum_{i=1}^{i=n} y_{1i} \left( 1 - \frac{\sum_{i=1}^{i=n} y_{2i}}{\sum_{i=1}^{i=n} y_{1i} + n} \right)}{\left( \sum_{i=1}^{i=n} y_{1i} + n \right) \left( \sum_{i=1}^{i=n} y_{1i} + n \right) \left( \sum_{i=1}^{i=n} y_{2i} + n \right)}
\]

\[
= \frac{\sum_{i=1}^{i=n} y_{1i}}{\sum_{i=1}^{i=n} y_{1i} + \sum_{i=1}^{i=n} y_{2i} + n}
\]

\[
= \frac{\bar{y}_1}{\bar{y}_1 + \bar{y}_2 + 1}
\]

Here, $\hat{q}_1$ and $\hat{q}_2$ are the maximum likelihood estimate of $q_1$ and $q_2$ respectively.

3.3.3 Estimation of Parameters of Distribution Based On GLM

Let’s again consider the conditional distribution of $Y_2$ given $Y_1$,

\[
P(Y_2 = y_2 | Y_1 = y_1) = \begin{pmatrix} y_1 + y_2 \end{pmatrix} q_2^{y_2} (1 - q_2)^{y_1 + 1}
\]

Now, it is known that the expected number of trials in case of negative binomial distribution with parameters $(r,p)$ is $r/p$, where $r$ represents the desired
number of trials and represents probability of success in each trial. Hence, from
the above conditional distribution we get the conditional expectation of the
random variable $Y_2$ given $Y_1 = y_1$ can be written as follows where $y_2$ represents
the desired number of successes,

$$E[Y_2 = y_2|Y_1 = y_1] = \frac{y_2}{q_2}$$

Now, let us consider a matrix $X$ with order $n \times p$, which is a matrix of the
explanatory variables. It is also called covariate. Let us consider another matrix
$\beta$ with order $1 \times p$ which represents the coefficient of $X$. In the case of GLM,
the conditional mean of $Y_2$ given $Y_1 = y_1$ can be expressed as a linear function
of regressors $\eta_i = x_{2i}^T \beta$ which is the linear predictor via a link function $g$ of
the conditional mean. Here, $x_{2i}$ represents an element corresponding to second
column and $i$ th row of the matrix $X$. Hence, the linear predictor corresponding
to the $i$th observation is

$$g(\mu_i) = \eta_i; i = 1, 2, ..., n.$$  

where $\mu_i$ represents the conditional mean corresponding to the $i$th observation.

Since, the variable $Y_2$ can be considered as count data, so considering the log
link, we get,

$$g(\mu_i) = \ln \eta_i = x_{2i}^T \beta$$

$$\Rightarrow \ln \frac{y_{2i}}{q_{2i}} = x_{2i}^T \beta$$

$$\Rightarrow \frac{y_{2i}}{q_{2i}} = \exp \left\{ x_{2i}^T \beta \right\}$$

$$\Rightarrow q_{2i} = y_{2i} \exp \left\{ x_{2i}^T \beta \right\}$$

(17)
Here, $\beta$ is an element of the matrix $\beta$ corresponding to the covariate $x_{2i}$ which represents the effect of covariate to the mean responses through the link function $g(\mu_i) = \eta_i$.

Differentiation of (7) with respect to $q_{1i}$ and setting it equal to zero, we get,

$$\frac{dl}{dq_{1i}} = \frac{1}{1 - q_{1i} - q_{2i}}(-1) + \frac{y_{1i}}{q_{1i}} = 0$$

$$\Rightarrow \frac{y_{1i}}{q_{1i}} = \frac{1}{1 - q_{1i} - q_{2i}}$$

$$\Rightarrow \hat{q}_{1i} = \frac{y_{1i} - y_{1i}\hat{q}_{2i}}{y_{1i} + 1} = \frac{y_{1i}(1 - \hat{q}_{2i})}{y_{1i} + 1} = \frac{y_{1i}(1 - y_{2i}\exp\{-x_{2i}^T\beta\})}{y_{1i} + 1}$$

(18)

3.4 Hypothesis Testing

In general, hypothesis test is a statistical test which is used to determine whether there is enough evidence in the sample data to infer that a condition is true for the entire population. Here, we are performing hypothesis tests to compare how well two related models fit the data. Different kinds of goodness of fit statistics are used to make these comparisons. In this case, we assign the reduced model to the null hypothesis $H_0$ and the saturated model to the alternative hypothesis $H_1$. We derive the deviance as the goodness of fit statistic for the reduced model, GLM, and the difference of deviances between two reduced models to make comparison between them. The deviance makes comparisons based on the ratio of likelihood function of the saturated model and the model of interest.
3.4.1 Deviance for Reduced Model (Consisting One Pair \((q_1, q_2)\))

The likelihood function for the saturated model can be written as follows using (7) and the maximum likelihood estimation of the parameters \(q_{1i}\) and \(q_{2i}\) from equation (9) and (11) respectively,

\[
l(b_{\text{max}}; y) = \sum_{i=1}^{i=n} \left[ y_{1i} \ln \frac{y_{1i}}{y_{1i} + y_{2i} + 1} + y_{2i} \ln \frac{y_{2i}}{y_{1i} + y_{2i} + 1} \\
+ \ln \left( 1 - \frac{y_{1i}}{y_{1i} + y_{2i} + 1} - \frac{y_{2i}}{y_{1i} + y_{2i} + 1} \right) \\
+ \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \\
= \sum_{i=1}^{i=n} \left[ y_{1i} \left\{ \ln y_{1i} - \ln (y_{1i} + y_{2i} + 1) \right\} + y_{2i} \left\{ \ln y_{2i} - \ln (y_{1i} + y_{2i} + 1) \right\} \\
+ \ln \left( \frac{1}{y_{1i} + y_{2i} + 1} \right) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \\
= \sum_{i=1}^{i=n} \left[ y_{1i} \ln y_{1i} - y_{1i} \ln (y_{1i} + y_{2i} + 1) + y_{2i} \ln y_{2i} - y_{2i} \ln (y_{1i} + y_{2i} + 1) \\
- \ln (y_{1i} + y_{2i} + 1) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right]
\]

(19)

Now, the likelihood function of the reduced model can be written as follows using (7) and the maximum likelihood estimation of \(q_1\) and \(q_2\) from equation (14) and (16),
\[ l(b; y) = \sum_{i=1}^{n} \left[ y_{1i} \ln \frac{\bar{y}_1}{y_1 + \bar{y}_2 + 1} + y_{2i} \ln \frac{\bar{y}_2}{y_1 + \bar{y}_2 + 1} \\
+ \ln \left( 1 - \frac{\bar{y}_1}{y_1 + \bar{y}_2 + 1} - \frac{\bar{y}_2}{y_1 + \bar{y}_2 + 1} \right) \\
+ \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \]

\[ = \sum_{i=1}^{n} \left[ y_{1i} \ln \bar{y}_1 - \ln (\bar{y}_1 + \bar{y}_2 + 1) \right] + y_{2i} \left[ \ln \bar{y}_2 - \ln (\bar{y}_1 + \bar{y}_2 + 1) \right] \\
- \ln (\bar{y}_1 + \bar{y}_2 + 1) + \ln (y_{1i} + y_{2i})! - \ln y_{1i}! - \ln y_{2i}! \right] \]

\[ = \sum_{i=1}^{n} \left[ y_{1i} \ln \bar{y}_1 - y_{1i} \ln (\bar{y}_1 + \bar{y}_2 + 1) + y_{2i} \ln \bar{y}_2 - y_{2i} \ln (\bar{y}_1 + \bar{y}_2 + 1) \\
- \ln (y_{1i} + y_{2i} + 1) + \ln (\bar{y}_1 + \bar{y}_2 + 1) \right] \]

(20)

So, the deviance can be written as follows from (19) – (20),

\[ D = 2 \left[ l(b_{max}; y) - l(b; y) \right] \]

\[ = 2 \sum_{i=1}^{n} \left[ y_{1i} \ln \frac{y_{1i}}{\bar{y}_1} - y_{1i} \ln \bar{y}_1 - y_{1i} \ln (y_{1i} + y_{2i} + 1) \\
+ y_{1i} \ln (\bar{y}_1 + \bar{y}_2 + 1) + y_{2i} \ln y_{2i} - y_{2i} \ln \bar{y}_2 \\
- y_{2i} \ln (y_{1i} + y_{2i} + 1) + y_{2i} \ln (\bar{y}_1 + \bar{y}_2 + 1) \\
- \ln (y_{1i} + y_{2i} + 1) + \ln (\bar{y}_1 + \bar{y}_2 + 1) \right] \]

(21)
According to Dobson (2001), this $D$ follows a $\chi^2$ distribution with $(2n - 2)$ degrees of freedom.

### 3.4.2 Deviance Based On GLM

The likelihood function based on the GLM interest can be written as follows using equation (7) and the maximum likelihood estimate of $q_{1i}$ and $q_{2i}$ based on GLM from equation (17) and (18) respectively,

\[
\begin{align*}
\ell(b; y) = \sum_{i=1}^{i=n} \left[ y_{1i} \ln \frac{y_{1i}(1 - y_{2i} \exp \{-x_{2i}^T \beta\})}{y_{1i} + 1} + y_{2i} \ln(y_{2i} \exp \{-x_{2i}^T \beta\}) \\
+ \ln \left( 1 - \frac{y_{1i}(1 - y_{2i} \exp \{-x_{2i}^T \beta\})}{y_{1i} + 1} - y_{2i} \exp \{-x_{2i}^T \beta\} \right) \right] \\
+ \ln \left( y_{1i} + y_{2i} \right) - \ln y_{1i} - \ln y_{2i} \right] \quad (22)
\end{align*}
\]

So, the deviance can be written as from (19) – (22),

\[
D = 2 \left[ \ell(b_{\text{max}}; y) - \ell(b; y) \right] \quad (23)
\]

\[
\begin{align*}
&= 2 \sum_{i=1}^{i=n} \left[ y_{1i} \ln y_{1i} - y_{1i} \ln(y_{1i} + y_{2i} + 1) + y_{2i} \ln y_{2i} \\
&- y_{2i} \ln(y_{1i} + y_{2i} + 1) - \ln(y_{1i} + y_{2i} + 1) \\
&- y_{1i} \ln \frac{y_{1i}(1 - y_{2i} \exp \{-x_{2i}^T \beta\})}{y_{1i} + 1} \\
&- y_{2i} \ln(y_{2i} \exp \{-x_{2i}^T \beta\}) \\
&- \ln \frac{1 - y_{2i} \exp \{-x_{2i}^T \beta\}}{y_{1i} + 1} \right] \\
\end{align*}
\]

According to Dobson (2001), this $D$ follows $\chi^2$ distribution with $(2n - p)$
3.4.3 Difference Between Two Deviances

In this case, we are going to compare two models based on the GLM.

The null hypothesis corresponding to a more reduced model is

\[ H_0 : \beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix} \]

Suppose that it is corresponding to the model \( M_0 \). The alternative hypothesis corresponding to a more general model is

\[ H_1 : \beta = \beta_1 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \]

Suppose that it is corresponding to the model \( M_1 \) with \( q < p < N \).

We can test \( H_0 \) against \( H_1 \) using the difference of the deviance statistics. Here, \( l(b_0; y) \) is used to denote the likelihood function corresponding to the model \( M_0 \) and \( l(b_1; y) \) to denote the likelihood function corresponding to the model \( M_1 \). Hence the deviance difference can be written as,
\[ \Delta D = D_0 - D_1 = 2 \left[ l(b_{\text{max}}; y) - l(b_0; y) \right] - 2 \left[ l(b_{\text{max}}; y) - l(b_1; y) \right] \]

\[ = 2 \left[ l(b_1; y) - l(b_0; y) \right] \]

\[ = 2 \sum_{i=1}^{i=n} \left[ y_{1i} \ln \left( \frac{y_{1i}(1 - y_{2i} \exp \{ -x_{2i}^T \beta_0 \})}{y_{1i} + 1} \right) \right. \]

\[ - \ln \left( \frac{y_{1i}(1 - y_{2i} \exp \{ -x_{2i}^T \beta_1 \})}{y_{1i} + 1} \right) \]

\[ + y_{2i} \left\{ \ln(y_{2i} \exp \{ -x_{2i}^T \beta_0 \}) - \ln(y_{2i} \exp \{ -x_{2i}^T \beta_1 \}) \right\} \]

\[ + \ln \frac{1 - y_{2i} \exp \{ -x_{2i}^T \beta_0 \}}{y_{1i} + 1} \]

\[ - \ln \frac{1 - y_{2i} \exp \{ -x_{2i}^T \beta_1 \}}{y_{1i} + 1} \right] \]

\[ = 2 \sum_{i=1}^{i=n} \left[ y_{1i} \ln \frac{1 - y_{2i} \exp \{ -x_{2i}^T \beta_0 \}}{1 - y_{2i} \exp \{ -x_{2i}^T \beta_1 \}} \right. \]

\[ + y_{2i} \ln \frac{\exp \{ -x_{2i}^T \beta_0 \}}{\exp \{ -x_{2i}^T \beta_1 \}} \]

\[ + \ln \frac{1 - y_{2i} \exp \{ -x_{2i}^T \beta_0 \}}{1 - y_{2i} \exp \{ -x_{2i}^T \beta_1 \}} \right] \]

\[ (24) \]

Now according to Dobson (2001) this \( \Delta D \) follows \( \chi^2 \) distribution with \( p - q \) degrees of freedom.

If the value of \( \Delta D \) is consistent with the \( \chi^2_{(p-q)} \) distribution we would generally choose the \( M_0 \) corresponding to \( H_0 \) because it is simpler. On the other hand, if the value of \( \Delta D \) is in the critical region i.e., greater than the upper tail \( 100 \times \alpha\% \) point of the \( \chi^2_{(p-q)} \) distribution then we would reject \( H_0 \) in favor of \( H_1 \) on the grounds that model \( M_1 \) provides a significantly better description of the data.
4 Data Simulation and Analysis

In order to determine whether or not our derived deviances work to test the hypothesis regarding the parameters of the bivariate geometric distribution and model adequacy we need appropriate data to analyze. Due to time constraints, we could not collect appropriate data to check our deviances. On the other hand, there is no computer software to generate data directly from bivariate geometric distribution.

In 2009, Krishna and Pundir had suggested an algorithm which is based on Theorem 11.4.1, p. 615 of Hogg et al. (2005) to generate random numbers from bivariate geometric distribution. According to this, the numbers from bivariate geometric distribution can be generated using the following steps,

- Step 1: Generate ten random numbers from univariate geometric distribution with probability of success \( \frac{1-q_1-q_2}{1-q_2} \).
- Step 2: Suppose that our generated random numbers from the geometric distribution are \( x_1, x_2, ..., x_n \).
- Step 3: Generate ten random numbers \( y_{ij} \) ten times each from negative binomial distribution with parameters \( x_i + 1 \) and \( (1 - q_2) \).
- Step 4: These generated pairs are from the bivariate geometric distribution with parameters \( q_1 \) and \( q_2 \).

In the following subsection we discuss how we will use simulation to check our deviances under different models.

4.1 Deviance Checking for Reduced Model

In this subsection, we use the following steps to check our derived deviance for the reduced model with a pair of parameters \((q_1, q_2)\).
• Step 1: Assume some fixed values of $q_1$ and $q_2$.

• Step 2: Generate ten random numbers from univariate geometric distribution with probability of success \( \frac{1-q_1-q_2}{1-q_2} \) using the assumed values of $q_1$ and $q_2$ from Step 1.

• Step 3: Suppose that our generated random numbers from the geometric distribution are $x_1, x_2, ..., x_n$.

• Step 4: Generate ten random numbers $y_{ij}$ ten times each from the negative binomial distribution with parameters $x_i + 1$ and $(1 - q_2)$.

• Step 5: The generated pairs are from the bivariate geometric distribution with parameters $q_1$ and $q_2$. item

Step 6: Estimate deviance which is derived in section 3.4.1 in equation (21).

We take the value of $q_1$ and $q_2$ ranging from 0.10 to 0.90. In this case, we have to be cautious about choosing the value of the parameters $q_1$ and $q_2$ such that $q_1 + q_2 < 1$, otherwise we will not be able to generate random numbers from the geometric distribution which is necessary to generate random numbers from the bivariate geometric distribution at the initial stage. We have taken several values for the pair $(q_1, q_2)$ and generate random pairs to observe the efficiency of our derived deviance under different parametric values. For each specified pairs of parameters $(q_1, q_2)$, we ran this experiment twice to see whether there is a change in our decision due to randomness. The value of the pair of parameters and the corresponding deviance value are given below. The R code to generate data from the bivariate geometric distribution and to find the deviance is given in the Appendix for the pair of parameters $(q_1 = 0.30, q_2 = 0.40)$. This code is used to deviances for the reduced model using other parametric values as well.
Table 1: Estimation of deviance for different parameters under consideration:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Deviance</th>
<th>$\chi_{198}(0.95)$</th>
<th>$\chi_{198}(0.975)$</th>
<th>$\chi_{198}(0.99)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1=0.30, q_2=0.30$</td>
<td>177.4164</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.30$</td>
<td>172.3071</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.40$</td>
<td>185.3107</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.40$</td>
<td>159.5293</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.50$</td>
<td>193.8942</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.50$</td>
<td>158.266</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.60$</td>
<td>223.1697</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.60$</td>
<td>193.667</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.40, q_2=0.30$</td>
<td>216.3456</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.40, q_2=0.30$</td>
<td>211.828</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.50, q_2=0.30$</td>
<td>148.1757</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.50, q_2=0.30$</td>
<td>254.3887</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.60, q_2=0.30$</td>
<td>239.3245</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.60, q_2=0.30$</td>
<td>215.4915</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.50$</td>
<td>232.1984</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.50$</td>
<td>191.7516</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.60$</td>
<td>184.1803</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.30, q_2=0.60$</td>
<td>236.0869</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.10, q_2=0.10$</td>
<td>97.9206</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>$q_1=0.10, q_2=0.10$</td>
<td>85.1073</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
</tbody>
</table>
The deviance we derived to test the parameters of the reduced model model works well as we see that all, but four of the values of the deviances are smaller than $\chi^2_{198}(0.95)$. However, among these four values of the deviances three are greater than $\chi^2_{198}(0.95)$, but less than $\chi^2_{198}(0.99)$. So, it can be concluded that
our derived deviance works well. On the other hand, if most of the values of the
deviances had a larger value than our desired $\chi^2$ value, then we had to conclude
that our derived deviance does not work in testing hypothesis regarding the
parameters of the reduced model.

4.2 Deviance Checking for GLM

In this subsection, we discussed how we will check the deviance which we have
derived to test parameters of generalized linear model (GLM). However, due to
time limitation this computation with covariate presence was not carried out.
Here we provided two algorithms. The first algorithm is to check the efficiency
of our derived deviance and the second is to determine whether a random vari-
able has a significant coefficient or not. The first algorithm is described in the
following steps:

- Step 1: Generate a random variable $X$ from a distribution say, uniform
  $(1, 2)$.
- Step 2: Assume a value of the coefficient $\beta$ to estimate parameters of bi-
  variate geometric distribution.
- Step 3: Estimate parameters $q_1$ and $q_2$ of bivariate geometric distribution.
- Step 4: Generate random pairs from bivariate geometric distribution using
  estimated parameter values from Step 3.
- Step 5: Estimate the deviance which is mentioned in section 3.4.2 in equa-
  tion (23).

The other algorithm which is used to check the significant of the coefficient
of the covariate are discussed in the following steps:

- Step 1: Generate random pairs from bivariate geometric distribution.
4.3 Difference in Deviance Checking

In this subsection, we consider the following hypothesis to test:

\[ H_0 : \beta = \beta_1 \]

\[ H_1 : \beta = \beta_0 = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \]

We use the following steps to test the efficiency of our derived deviance difference:

- Step 1: Generate random variable X from uniform (1, 2).
- Step 2: Assume a value of the coefficient \( \beta_1 \) of X to estimate parameters of \( q_1 \) and \( q_2 \) of bivariate geometric distribution.
- Step 3: Estimate parameters \( q_1 \) and \( q_2 \) of bivariate geometric distribution.
- Step 4: Generate random pairs from bivariate geometric distribution using estimated values of parameters from Step 3.
- Step 5: Assume an integer value of \( \beta_0 \) which can be used as the intercept of the GLM.
- Step 6: Estimate the deviance which is mentioned in section 3.4.3 in equation (24).
• Step 7: Compare this deviance with the tabulated value of \( \chi^2 \). If the value of deviance is smaller than the tabulated value of \( \chi^2 \), then we can say that our derived deviance difference works well.

5 Conclusion

In this thesis, we addressed an important problem of inference in bivariate geometric distribution. We derived testing procedure to test parameters of this distribution both with and without covariate information. The test has been derived based on deviance statistic. In the absence of covariate the computation simplifies and the test provides desired results. The derived deviance works well as we can see that in most of the cases it could lead us to the right decision of not to reject the parametric values of the bivariate geometric distribution when the data are originally generated from that distribution.

We derived a similar approach based on deviance in the presence of covariate. Due to time limitation, the computation for the deviance approach was not carried out in the presence of covariates. However, the algorithm has been described to present the steps involved in the computation. Another limitation is that the absence of real life data to test the efficiency of our deviance approach.

6 References


7 Appendix

The R code to test the deviance for the reduced model:
q1<-0.30
q2<-0.40
y1<-rgeom(10,((1-q1-q2)/(1-q2)))
y1
y21<-rnbinom(10,y1[1]+1,1-q2)
y21
y22<-rnbinom(10,y1[2]+1,1-q2)
y22
y23<-rnbinom(10,y1[3]+1,1-q2)
y23
y24<-rnbinom(10,y1[4]+1,1-q2)
y24
y25<-rnbinom(10,y1[5]+1,1-q2)
y25
y26<-rnbinom(10,y1[6]+1,1-q2)
y26
y27<-rnbinom(10,y1[7]+1,1-q2)
y27
y28<-rnbinom(10,y1[8]+1,1-q2)
y28
y29<-rnbinom(10,y1[9]+1,1-q2)
y29
y210<-rnbinom(10,y1[10]+1,1-q2)
y210
y1c<-rep(y1[1],10)
y1c
df1<-data.frame(y1c,y21)
df1
y2c<-rep(y1[2],10)
y2c
df2<-data.frame(y2c,y22)
df2
y3c<-rep(y1[3],10)
y3c
df3<-data.frame(y3c,y23)
df3
y4c<-rep(y1[4],10)
y4c
df4<-data.frame(y4c,y24)
df4
y5c<-rep(y1[5],10)
y5c
df5<-data.frame(y5c,y25)
df5
y6c<-rep(y1[6],10)
y6c
df6<-data.frame(y6c,y26)
df6
y7c<-rep(y1[7],10)
y7c
df7<-data.frame(y7c,y27)
df7
y8c<-rep(y1[8],10)
y8c
df8<-data.frame(y8c,y28)
df8
y9c<-rep(y1[9],10)
y9c
df9<-data.frame(y9c,y29)
df9
y10c<-rep(y1[10],10)
y10c
df10<-data.frame(y10c,y210)
df10
dm1<-data.matrix(df1,rownames.force=NA)
dm1
dm2<-data.matrix(df2,rownames.force=NA)
dm2
dm3<-data.matrix(df3,rownames.force=NA)
dm3
dm4<-data.matrix(df4,rownames.force=NA)
dm4
dm5<-data.matrix(df5,rownames.force=NA)
dm5
dm6<-data.matrix(df6,rownames.force=NA)
dm6
dm7<-data.matrix(df7,rownames.force=NA)
dm7
dm8<-data.matrix(df8,rownames.force=NA)
dm8
dm9<-data.matrix(df9,rownames.force=NA)
```r
dm9
dm10<-data.matrix(df10,rownames.force=NA)
dm10
final<-rbind(dm1,dm2,dm3,dm4,dm5,dm6,dm7,dm8,dm9,dm10)
final
y1_-mean(final[,1])
y1_
y2_-mean(final[,2])
y2_
f1<-sum(ifelse(final[,1]>0,final[,1]*log(final[,1]/y1_,base=exp(1)),0.00001*log(0.00001/y1_,base=exp(1))))
f1
f2<-sum(final[,1]*log((final[,1]+final[,2]+1)/(y1_+y2_+1),base=exp(1)))
f2
f3<-sum(ifelse(final[,2]>0,final[,2]*log(final[,2]/y2_,base=exp(1)),0.00001*log(0.00001/y1_,base=exp(1))))
f3
f4<-sum(final[,2]*log((final[,1]+final[,2]+1)/(y1_+y2_+1),base=exp(1)))
f4
f5<-sum(log((final[,1]+final[,2]+1)/(y1_+y2_+1),base=exp(1)))
f5
Deviance<-2*(f1-f2+f3-f4-f5)
Deviance
qchisq(0.95,df=198)
qchisq(0.975,df=198)
qchisq(0.99,df=198)
```

28