Tests of hypotheses for the parameters of a bivariate geometric distribution

Abstract

A bivariate geometric distribution is an extension to a univariate geometric distribution where the occurrence of three different types of events is considered. Many statisticians have studied and given different forms of a bivariate geometric distribution. In this paper, we considered the form given by Phatak & Sreehari. We estimated the parameters of this distribution under three different models using maximum likelihood estimation (MLE) and derived deviances as the goodness of fit statistics for testing the parameters and deviance difference for comparing two models. Using simulated data we found that the deviance measure works well to test a reduced model against a full model.

Keywords: bivariate geometric distribution, deviance, deviance difference.

Introduction

Many situations in real world cannot be described by a single variable. Simultaneous occurrence of multiple events warrants multivariate distributions. For instance, univariate geometric distribution can represent occurrence of failure of one component of a system. However, to study systems with several components that may have different types of failures, such as twin engines of an airplane or the paired organ in a human body, bivariate geometric distributions are suitable. Bivariate geometric distribution has increasingly important roles in various fields, including reliability and survival analysis. There are different forms of a bivariate geometric distribution. Phatak & Sreehari provided a form of the bivariate geometric distribution which is considered here. They introduced a form of probability mass function which take into consideration of three different types of events. There are other forms which can be seen in Nair & Nair, Hawkes, Arnold et al., and Sreehari & Vasudeva. Basu & Dhar proposed a bivariate geometric model which is analogous to bivariate exponential model developed by Marshal & Olkin. Characterization results are developed by Sun & Basu, Sreehari, and Sreehari & Vasudeva.

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

In this paper, the parameters of a saturated model, reduced model and generalized linear model (GLM) for a bivariate geometric distribution are estimated using the maximum likelihood method. We also derived deviances as the goodness of fit statistics for testing parameters corresponding to these models and deviance difference to compare two related models in order to determine which model fits the data well. Rest of the paper is organized as follows: Univariate Geometric Distribution, Bivariate Geometric Distribution, Maximum Likelihood Estimation, Hypothesis Testing, Data Simulation and Analysis and finally Conclusion.

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,

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

The moment generating function can be given by,

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

the mean and the variance of this distribution are

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

An extension to the univariate geometric distribution is the bivariate geometric distribution which is discussed in the next section.

Bivariate geometric distribution

The joint probability mass function of a bivariate geometric distribution can be obtained by the product of a marginal and a
conditional distribution, introduced by Phatak & Sreehari.\(^1\) They considered a process from which the units could be classified as good, marginal and bad with probabilities \(q_1, q_2\) and \(q_4 = (1-q_1-q_2)\) respectively. They proposed that the probability mass function of observing the first bad unit after several good and marginal units is given as follows:

\[
P(Y_i=y_1,Y_2=y_2) = \left(\frac{y_1+y_2}{y_1}\right) q_1^n q_2 y_2 (1-q_1-q_2), \quad y_1, y_2 = 0, 1, 2, ...;
\]

\[0 < q_1 + q_2 < 1.
\]

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

The marginal distribution of \(Y_i\) is a geometric distribution with probability of success \(\left(\frac{1-q_i-q_2}{1-q_2}\right)\), and can be written as follows,

\[
P(Y_i=y_1) = \left(\frac{1-q_i-q_2}{1-q_2}\right) q_1^n q_2^{y_1}, \quad y_1 = 0, 1, 2, ...
\]

(1)

The conditional distribution of \(Y_j\) given \(Y_i\) is

\[
P(Y_j=y_2|Y_i=y_1) = \left(\frac{y_1+y_2}{y_1}\right) q_2^{y_2} (1-q_1-q_2)^{y_1+1}, \quad y_1, y_2 = 0, 1, 2, ...
\]

(2)

The product of the marginal distribution of \(Y_i\) in equation (2) and the conditional distribution of \(Y_j\) given \(Y_i\) in equation (3) gives the mass function of bivariate geometric distribution in equation (1).

### Maximum likelihood estimation

#### Estimation of parameters in the absence of regressors

In order to find the maximum likelihood estimators (mle)s from a saturated model (parameters are different for each pair of observations), it suffices to consider the mass functions based on the marginal and conditional mass functions. Let \(Y_1, ..., Y_n\) be independent random vectors each having bivariate geometric distribution with different pairs of parameters \((q_{1i}, q_{2i})\) for \(i = 1, 2, ..., n\).

The log likelihood function based on the conditional distribution of \(Y_i\) given \(Y_j\) can be written as follows using (3):

\[
l = \sum_{i=1}^{n} \left[ y_1 \ln q_2 + \left( y_1 + 1 \right) \ln (1-q_{2i}) + \ln (y_1 + y_2) + \ln (y_1 - y_1! - y_2! + 1) \right]
\]

(4)

Differentiating (4) with respect to \(q_{2i}\) and setting it equal to zero, we get the mle of \(q_{2i}\) as,

\[
\hat{q}_{2i} = \frac{y_2}{y_1 + y_2 + 1}
\]

(5)

The log likelihood function based on the marginal distribution of \(Y_i\) from (2) is,

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

(6)

Differentiating (6) with respect to \(q_{1i}\) and setting it equal to zero, the mle of \(q_{1i}\) can be derived as,

\[
\hat{q}_{1i} = \frac{y_i}{y_i + y_{2i} + 1}
\]

(7)

Here, \(\hat{q}_{1i}\) and \(\hat{q}_{2i}\) are the maximum likelihood estimators of \(q_{1i}\) and \(q_{2i}\), \(i = 1, ..., n\) respectively under the saturated model.

Similarly the maximum likelihood estimators (mle)s from a reduced model (parameters are the same for each pair of observations) can be obtained as:

\[
\hat{q}_{1} = \frac{y_y}{y_1 + y_{2} + 1}
\]

\[
\hat{q}_{2} = \frac{y_2}{y_1 + y_{2} + 1}
\]

(8)

(9)

Where \(\hat{q}_{1}\) and \(\hat{q}_{2}\) are the maximum likelihood estimators of \(q_{1}\) and \(q_{2}\) respectively under the reduced model.

#### Estimation of parameters in the presence of regressors:

In the presence of regressors, one can employ a generalized linear model and hence estimate the parameters in terms of the estimated model parameters. The conditional distribution of \(Y_j\) given \(Y_i\) in (3) can be set as exponential family representation as follows,

\[
P(Y_2=y_2|Y_1=y_1) = \exp \left[ y_2 \ln q_2 - \left( y_1 + 1 \right) \ln (1-q_{2i}) + \ln \left( \frac{y_1}{y_1+y_{2i}} \right) ! + \ln \left( \frac{y_2}{y_2+y_{2i}} \right) ! \right]
\]

Here the natural parameter and the function of the natural parameter respectively are,

\[
\theta = \ln q_2
\]

\[
b(\theta) = -(y_1 + 1) \ln (1-q_{2i})
\]

Thus the mean of the conditional distribution of \(Y_j\) given \(Y_i\) is

\[
\mu_i = E[Y_2|Y_1=y_1] = b'(\theta) = \frac{y_1 + 1}{1-q_2}
\]
A generalized linear model based on the conditional distribution of \( Y_2 \) given \( Y_1 \) can be written as,

\[
g(\mu_i) = \ln \mu_i = \frac{\beta_j}{n} = l \sum x_{2i} \beta_j; i = 1, \ldots, n; \quad n > p
\]

Since, \( Y_2 \) represents the number of trials before a certain event can occur it is considered as count response, the linear predictor can be written as the logarithm of the mean \( \mu_i \). Thus the conditional link function can be expressed as,

\[
g(\mu_i) = \ln \frac{y_{ii} + 1}{1 - q_{2i}} = \sum x_{2i} \beta_j
\]

\[
\Rightarrow \hat{q}_{2i} = 1 - \frac{y_{ii} + 1}{\exp \left( \sum_{j=1}^{p} x_{2i} \beta_j \right)}
\] (10)

**Hypothesis testing**

In order to test the identical parameter assumption across each pair of observed data, we derived deviance as a goodness of fit statistic. Additional deviance statistics are derived for generalized linear model (glm) to compare two nested glm's.

**Deviance for reduced model with identical parameter assumption**

The log likelihood function for the saturated model can be written using (1) and the maximum likelihood estimates of the parameters \( q_{1i} \) and \( q_{2i} \) from equations (5) and (7) respectively as follows,

\[
I(b_{max}; y) = \sum_{i=1}^{n} \left[ y_{ii} \ln y_i - y_{ii} \ln (y_{ii} + y_{1i} + 1) + y_{1i} \ln y_{1i} - y_{1i} \ln (y_{1i} + y_{1i} + 1) \right]
\]

(12)

Similarly, the log likelihood function of the reduced model can be written using (1) and the maximum likelihood estimates of \( q_1 \) and \( q_2 \) from equations (8) and (9) respectively as follows,

\[
I(b_{y}; y) = \sum_{i=1}^{n} \left[ y_{ii} \ln y_i - y_{ii} \ln (y_{ii} + y_{1i} + 1) + y_{1i} \ln y_{1j} - y_{1i} \ln (y_{1i} + y_{1i} + 1) \right]
\]

(13)

Thus the deviance statistic for testing the identical parameter for each observed pair of data can be expressed as follows,

\[
D_j = 2 \left[ I(b_{max}; y) - I(b_{y}; y) \right] = 2 \sum_{i=1}^{n} \left[ y_{ii} \ln y_i - y_{ii} \ln (y_{ii} + y_{1i} + 1) + y_{1i} \ln y_{1i} - y_{1i} \ln (y_{1i} + y_{1i} + 1) \right]
\]

(14)

According to Dobson [12], \( D_j \) follows a \( \chi^2 \) distribution with \( (2n - 2) \) degrees of freedom.

**Deviance for a GLM**

The deviance statistic for the glm of interest can be written using (1) and the maximum likelihood estimates of \( q_{1i} \) and \( q_{2i} \) based on the glm from equations (10) and (11) respectively as follows,

\[
I(b_{y}; y) = \sum_{i=1}^{n} \left[ y_{ii} \ln \frac{y_{ii} + 1}{\exp \left( \sum_{j=1}^{p} x_{2i} \beta_j \right)} + y_{1i} \ln \frac{1 - \hat{q}_{2i}}{\exp \left( \sum_{j=1}^{p} x_{2i} \beta_j \right)} \right] + \ln \frac{1}{\exp \left( \sum_{j=1}^{p} x_{2i} \beta_j \right)} + \ln \left( y_{1i} + y_{2i} \right) - \ln y_{1i} - \ln y_{2i}
\]

(15)
Tests of hypotheses for the parameters of a bivariate geometric distribution

According to Dobson [12], $D_H$ follows $\chi^2$ distribution with $(2n-p)$ degrees of freedom.

Comparison between two GLMs

In order to compare two nested generalized linear models, we consider the following hypotheses. The null hypothesis corresponding to a smaller model ($M_0$) in terms of number of regression parameters is

$$H_0 : \beta_2 = \beta_{20}$$

The alternative hypothesis corresponding to a bigger model ($M_1$ with $q < p < n$) within which the smaller model is nested can be written as,

$$H_1 : \beta_2 = \beta_{21}$$

We can test $H_0$ against $H_1$ using the difference of the deviance statistics. Here, $l(h_{i,y})$ is used to denote the likelihood function corresponding to the model $M_0$ and $l(h_{i,y})$ to denote the likelihood function corresponding to the model $M_1$. Hence the deviance difference can be written as,

$$\Delta D = D_{H_1} - D_{H_0} = 2\left[l(h_{i,y}) - l(h_{i,y})\right] = 2\left[l(h_{i,y}) - l(h_{i,y})\right] = 2\sum_{i=1}^{n-p} x_i y_i \ln \left(y_i + y_{i+1} + 1\right) - y_i \ln \left(\exp \left[\sum_{j=1}^{i-1} x_j \beta_{j1} + (y_{i+1} + y_{i+2}) \right] - y_{i+1}\right) - y_i \ln \left(\exp \left[\sum_{j=1}^{i-1} x_j \beta_{j1} + (y_{i} + y_{i+1}) \right] - y_{i-1}\right)$$

$$\Delta D = D_{H_1} - D_{H_0} = 2\left[l(h_{i,y}) - l(h_{i,y})\right] = 2\left[l(h_{i,y}) - l(h_{i,y})\right] = 2\sum_{i=1}^{n-p} x_i y_i \ln \left(y_i + y_{i+1} + 1\right) - y_i \ln \left(\exp \left[\sum_{j=1}^{i-1} x_j \beta_{j1} + (y_{i+1} + y_{i+2}) \right] - y_{i+1}\right) - y_i \ln \left(\exp \left[\sum_{j=1}^{i-1} x_j \beta_{j1} + (y_{i} + y_{i+1}) \right] - y_{i-1}\right)$$

According to Dobson [12] 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.

Data simulation and analysis

To determine the efficiency of our derived deviances we need to have data with known parameters. However, we cannot generate data directly from bivariate geometric distribution using the available computer software packages. Krishna and Pundir suggested an algorithm based on a theorem given by Hogg et al.,13 to generate random numbers from bivariate geometric distribution. According to this, paired values can be generated from a bivariate geometric distribution using the following steps,

1. Generate $k$ random numbers from univariate geometric distribution with probability of success $q_1$ and $q_2$.

Deviance checking for reduced model

In this subsection, we use the following steps to check our derived deviance for the reduced model with identical values of parameters $(q_1, q_2)$ for each observed pair of data.

1. Assume some fixed values of $q_1$ and $q_2$.

2. Generate $k$ random numbers from univariate geometric
distribution with probability of success \(\frac{1-q_1}{1-q_2}\) 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 \(x_1, x_2, \ldots, x_k\).

Step 4: Generate \(k\) random numbers \(Y_{ij}\), \(k\) times each from the negative binomial distribution with parameters \(x_i + 1\) and \(1 - q_i\).

Step 5: The generated pairs are from the bivariate geometric distribution with parameters \(q_1\) and \(q_2\).

Step 6: Estimate deviance which is derived in (14).

We take the values of \(q_1\) and \(q_2\), ranging from 0.10 to 0.90 and satisfying the constraint \(q_1 + q_2 < 1\). We considered 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 values of the pair of parameters and the corresponding deviance values are tabulated as follows.

Table 1 Estimation of deviance for different parameters under consideration

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Deviance</th>
<th>(0.95)</th>
<th>(0.975)</th>
<th>(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>236.0689</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>85.10731</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.20</td>
<td>100.8624</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.20</td>
<td>155.157</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.30</td>
<td>123.3245</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.20, q_2=0.20</td>
<td>113.3245</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.20, q_2=0.20</td>
<td>147.3637</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.20, q_2=0.30</td>
<td>166.6306</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.20, q_2=0.30</td>
<td>157.8232</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.30, q_2=0.10</td>
<td>133.2772</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.30, q_2=0.10</td>
<td>131.2191</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.80</td>
<td>183.8584</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.80</td>
<td>218.8224</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.80, q_2=0.10</td>
<td>203.6515</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.80, q_2=0.10</td>
<td>177.6116</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.40</td>
<td>144.1728</td>
<td>231.8292</td>
<td>238.8612</td>
<td>247.2118</td>
</tr>
<tr>
<td>q_1=0.10, q_2=0.40</td>
<td>168.524</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 works well as we see that all, but four of the values of the deviances are greater than $\chi^2_{0.05}$. However, among these four values of the deviances three are greater than $\chi^2_{0.05}$, but less than $\chi^2_{0.01}$. 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.

**Conclusion**

In this paper, we addressed an important problem of inference regarding bivariate geometric distribution and developed testing procedure for the parameters of this distribution with and without covariate information. Our method depends on deriving the deviance statistics using maximum likelihood estimators (mle) of parameters. Our mles of the parameters of the bivariate geometric distribution are obtained using the conditional and the marginal distributions.

We conducted a numerical analysis based on simulated data for the testing the identical parameter assumption for each pair of observed data. Our numerical example did not consider any covariate information. We found that without covariate information our derived deviance worked well in most cases.

**Acknowledgement**

None.

**Conflict of interest**

None.

**References**


DOI: 10.15406/bbij.2016.04.00112