Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2023;9(1):68-85

Original Article

Beta-Geometric Regression for Modeling Count Data on First Antenatal Care Visit (ANC) with Application

Zainab M. Al-Balushi,[1], Amadou Sarr,[2] M Mazharul Islam[2*]

[1]Department of Educational Studies and International Cooperation, Ministry of Education, Muscat, Sultanate of Oman.
[2]Department of Statistics, College of Science, Sultan Qaboos University, Muscat, Sultanate of Oman.

ARTICLE INFO

Received 08.11.2022
Revised 05.01.2023
Accepted 09.02.2023
Published 15.03.2023

Key words:
Beta-geometric regression;
Geometric regression;
Count data;
Antenatal care visits (ANC)

ABSTRACT

Introduction: Little attention has been paid to modeling count data with the geometric distribution. There are many real-life phenomena with a constant probability of first success. However, in practice, the probability of the first success may vary, making simple geometric models unsuitable for modeling such data. One can assume one of many continuous distributions for modeling the probability of first success with the parameter space [0, 1]. In this respect Beta distribution defined on the standard unit interval [0,1] is the most useful distribution due to its ability to accommodate a wide range of shapes. Thus, in this paper, by mixing Beta and geometric distribution, we developed a Beta-geometric distribution for modeling the count data through application to real-life count data on time to the first antenatal care (ANC) visit.
Methods: The estimation of the distribution parameters using the method of moments, maximum likelihood estimation (MLE) method, and Bayesian estimation approach are provided. Based on the Beta-geometric distribution, we developed a new Beta-geometric regression model for analyzing count data that follow the geometric distribution. The goodness of fit of the derived model has been tested using real data on time to the first ANC visit.
Results: Beta-geometric distribution has a simple form for its probability mass function (pmf), and is flexible in capturing both underdispersion and overdispersion that may present in count data. It was found that the proposed Beta-geometric regression model fit the count data on the first ANC visit better than simple geometric distribution or Negative Binomial distribution.
Conclusion: Unlike the Poisson or Negative Binomial distribution, Beta-geometric distribution does not need an additional parameter to accommodate underdispersion or overdispersion and thus could be a flexible choice for analyzing any count data. The goodness of fit test of the Beta-geometric model provides better fitting of the model to real data on time to first ANC visit than geometric or Negative binomial models.

*.Corresponding Author: mislam@squ.edu.om

Introduction

Count variable, which takes on only discrete values, intrinsically non-normal, heteroskedastic, right-skewed, and have a variance that increases with the mean,[1,2] and thus cannot be modeled with the classical statistical techniques requiring normality and homoscedasticity assumptions. For modeling count data, the most commonly used model is the Poisson model or its various modified form under Generalized Linear Modeling approach.[3-5] However, the most serious limitation of the Poisson model is the imposed equality of conditional mean and variance of the response variable which is termed as equi-dispersion. In reality, count data with equal mean and variance is very rare. Violation of equi-dispersion condition have similar effect as the violation of heteroskedasticity in linear regression model. Inferences made on the basis of assumption of equi-dispersion for data which actually over-dispersed or under-dispersed could be misleading, despite the fact that the parameter can still be estimated consistently.[6] If over-dispersion or under-dispersion is not taken into consideration while analyzing data, estimates of the standard errors under equi-dispersion will be small, and thus the test statistics for the parameter estimates will be too large, significance will be overestimated, and confidence limits will be too small.[7] To overcome the problem of over-dispersion, Negative Binomial (NB), generalized Poisson, zero-inflated Poisson, or hurdle models are often suggested.[2,8,9]

Although geometric distribution, which is a special case of NB distribution, and also belongs to the family of discrete distributions, little attention has been paid in modeling count data with geometric distribution. The geometric distribution is a probability distribution that is used to model the probability of experiencing a certain amount of failures before experiencing the first success in a series of Bernoulli trials. Since many characterizations of the geometric distribution are analogous to the characterization of the exponential distribution, geometric distribution is considered as the discrete analogue of the continuous exponential distribution.[10] The geometric distribution has been used extensively in the literature in modeling the distribution of the lengths of waiting times.[11-15]

In the area of women's reproductive health care, a variable of primary interest is the time-to-first antenatal care (ANC) visit, which is a random variable that counts the number of months (or trials) to have the first ANC visit (or success) in a series of independent Bernoulli trials, and thus follow the characteristics of a geometric distribution. Antenatal care (ANC) is the routine care of pregnant women starting from the date of conception to onset of delivery to reduce the risk of adverse pregnancy outcomes and improve the maternal and child health and their survival.[10,11] According to the new guidelines of the World Health Organization (WHO), every mother should have at least eight ANC visits during the pregnancy period.[16] It also emphasizes that all pregnant mothers should start ANC visit as soon as possible, preferably within the first trimester of pregnancy. Timing of first ANC visit has been observed to predict the compliance of full coverage of WHO recommended contents of care.[16] Over the period, a good number of studies have been conducted on time to first ANC visits. However, most of the earlier studies considered the timing of first ANC visit as a

categorical outcome variable by dichotomizing it as early initiation or late initiation, and thus applied logistic regression models to find the predictors of timing of first ANC visits.[17-20] Besides, different studies used different cutoff point to dichotomies the timing of first ANC visits as early or late attendance. Since, early or late initiation of ANC visit is not a natural category of ANC visit (e.g. alive or dead is a natural classification of survival status), analysis using such categories for a count variable might deviate from reality and might suffer from the risk of misclassification of the sample objects. This type of classification put all observed values of the exposure variable (i.e. time to first ANC visits) into two exposure levels. This intern modifies a real distribution, mostly a skewed distribution, into a binomial distribution, which might have some consequences in parametric estimation. The loss of information from dichotomizing a count or continuous variable has been documented by many studies.[21-25] In a recent study, Sroka and Nagaraja26 demonstrated that if the count data is analyzed directly using generalized regression model approach, the confidence intervals for the odds ratio could be up to 64% shorter (or 36% as wide) compared to if the data had been dichotomized and analyzed using logistic regression. Since the distributional assumption plays critical role in statistical inference,[27] it is important to choose the appropriate distribution for modeling the variables under consideration. For instance, if a real life data that follow the physical process of a particular distribution, but other distribution may fit it reasonably, even then one should use the distribution that follow the physical process. There is a dearth of literature on the use of geometric regression for analyzing count data.

In a recent study, Al-Balushi and Islam[28] illustrated the suitability of the geometric regression model for analyzing the count data on time to first ANC visit, and concluded that the geometric regression model may provide a flexible model for fitting count data sets which may present over-dispersion or under-dispersion. They further observed that the geometric model could be an alternative model to the widely used Poisson model for modeling the count data in the presence of over or under dispersion. However, the geometric regression model used by Al-Balushi and Islam[28] for modeling the count data on first antenatal care (ANC) visits has a limitation that it was based on assumption that the probability of first success or first ANC visit (p) as constant for all women during pregnancy period. In reality the probability of first ANC visit (p) varies from woman to woman, depending on their demographic, social, economic or behavioral characteristics. In such situation, the results of the geometric model under constant probability of first success could be biased. It is, therefore, important to develop model assuming that the probability of first success vary from woman to woman according to some underlying distribution. Many continuous distributions can be assumed for varying probability of first success that lies in the parameter space [0, 1]. However, the most appropriate and convenient distribution could be the beta distribution, because it is the proper conjugate distribution for the geometric model,[29] and it has flexibility in accommodating wide range of shape. It then produces a mixed distribution, namely the beta-geometric distribution. There are some analogous applications of beta-geometric distribution in studying the human reproduction.[30-32] For example, Weinberg

and Gladen[33] (1986) used beta-geometric distribution for modeling human fecundability – the monthly probability of conception. Because of flexible nature of beta distribution, a good number of beta mixture of other distributions has been developed for some other analogous applications.[34-37]

In recent times, researchers from various academic spheres are increasingly attempting to develop new probability distributions by applying mixture or compound mixture techniques for modeling of various complex real-life phenomena. Some of the new probability distributions are quite flexible in that they result in some other well-defined probability distributions when their parameter(s) are set to certain values. For instance, using mixed Poisson and mixed NB distributions, Zamani, Ismail & Shekari[38] developed new weighted negative binomial-Poisson Lindley distribution for modeling over disperse count data.

Eugene et al.[39] introduced the beta-generated family of univariate continuous distributions. Following Eugene et al.,[39] many other authors have defined a number of the beta-generated distributions, using various distribution function of the outcome variable Y, such as beta-Gumbel distribution by Nadarajah and Kotz,[40] beta-Weibull distribution by Famoye et al.,[41] beta-gamma distribution by Kong et al.,[42] beta-Pareto distribution by Akinsete et al.[43] (2008), beta-Cauchy distribution by Alshawarbeh et al.[44]

In this paper, a beta-geometric regression is proposed for modeling count data on time to first ANC visits to capture the real life situation of variation of the parameter p of the geometric distribution. This is an extension of the simple geometric regression used by Al-Balushi and Islam[28] for analysis the count data on time to first ANC visits by relaxing the assumption of constant probability of first ANC visit by all women. We applied the beta-geometric model on real data. The statistical properties of the beta-geometric distribution are discussed. The estimation of the parameters of the distribution using method of moments, maximum likelihood estimation (MLE) method and Bayesian estimation approach are provided. The goodness of fit of the derived model has been tested. A generalized linear model (GLM) based on the beta-geometric distribution has been developed and applied for identifying the significant predictor of the response variable.

Methods

The Geometric distribution and its characteristics

Geometric distribution is defined as the distribution of the number of trials until the first occurrence of the k-th consecutive success.45 Let the response variable Yi (i=1, 2, ….) is a count of the number of trials needed to get the first success. If all the trails have the constant probability of first success, say p, then the probability distribution of the count variable Yi can be modeled using the geometric distribution, with probability mass function (pmf),

$$P(Y = y \mid p) = p(1-p)^{(y-1)}, y = 1,2,\ldots\ldots; 0 < p < 1. \quad (1)$$

The cumulative distribution function or cdf of the geometric distribution is given by

$$F_Y(y) = P(Y \le y) = 1-(1-p)^y,\ y = 1,\ 2,\ 3,\ \ldots\ldots.$$

Figure 1 depicts the nature and behavior of geometric distribution for varying values of its parameter p. It is evident from the figure 1

that the probability of first success is almost constant for higher mean of the count and negative exponential for smaller mean values.

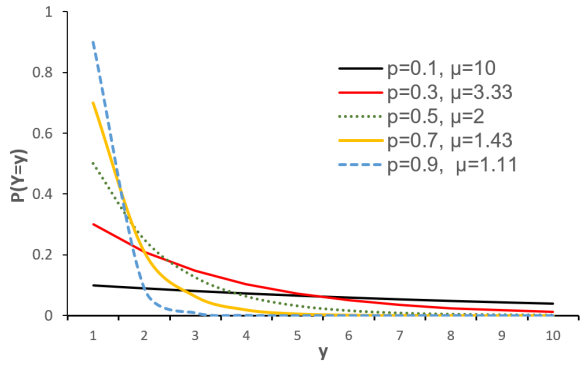

Figure 1. The pmf plot of Geometric distribution for varying values of the parameter *p*

The mean and variance of the distribution given in (1) are respectively,

$$E(Y \mid p) = \mu = \frac{1}{p}$$

and

$$Var(Y \mid p) = \frac{1-p}{p^2} = \mu(\mu - 1)$$

This indicates that the variance of the geometric distribution is a function of its mean.
The dispersion index is given as

$$\frac{Var(Y)}{E(Y)} = \frac{1-p}{p} = \frac{1}{p} - 1 \quad . \qquad (2)$$

It is evident from (2) that the dispersion index of geometric distribution depends on the value of the parameter p. If p=0.5, the distribution is equi-dispersed (i.e. mean = variance); if p <0.5, the distribution is over-dispersed (i.e. variance > mean), and if p > 0.5 then the distribution is under-dispersed (i.e. variance < mean). This implies that the geometric distribution can capture both under-dispersion and overdispersion in data, while NB model generically deals with over dispersion.[2] An added advantage of the geometric distribution in relation to the NB and generalized Poisson distribution is that it involves single parameter and no additional (dispersion) parameter is necessary to accommodate over or under-dispersion.
The Geometric distribution is the only discrete distribution with non-negative integer support that can be characterized as having an interesting property, known as "memory-less" or "*Markovian*" property.[10] For integers s>t, it can be shown that $P(Y > s \mid Y > t) = P(Y > s - t)$.
The geometric distribution is sometimes used to model "lifetimes" or "time until failure" of an object. Thus it is the simplest type of discrete waiting time distribution. Many other characteristics may be seen in Feller.[46] Many works have been done on characterization of the Geometric distribution such as Ferguson,[47] Arnold,[48] Gultekin and Bairamov,[49] while Tripathi et al.[50] provided many generalization of the Geometric distribution.

Beta-geometric distribution

In practice, the probability of first success, *p*, may vary from observation to observation, according to some underlying distribution. As *p* lies between 0 and 1, the most appropriate and convenient distribution for *p* is the beta distribution, because it is the proper conjugate distribution for the geometric model.[29]
Let *p* follow a beta distribution with parameter α and γ, then the probability density function (pdf) of *p* is given by:

$$f(p \mid \alpha, \gamma) = \frac{p^{\alpha-1}(1-p)^{\gamma-1}}{B(\alpha,\gamma)} \qquad 0<p<1. \qquad (3)$$

where,

$$B(\alpha,\gamma)=\frac{\Gamma(\alpha)\Gamma\gamma}{\Gamma(\alpha+\gamma)}=\int_0^1 t^{\alpha-1}(1-t)^{\gamma-1}dt$$

The mean and variance of the beta random variable p are

$$\mu=\frac{\alpha}{\alpha+\gamma}\text{ , and }\quad \sigma^2=\frac{\alpha\gamma}{(\alpha+\gamma)^2(\alpha+\gamma+1)}$$

respectively.

Assuming that the parameter p in equation (1) follow a Beta distribution, we obtain a mixed distribution, namely the beta-geometric distribution and its pmf is given by

$$\mathrm{P}(y,\alpha,\gamma)=\frac{B(\alpha+1,y+\gamma-1)}{B(\alpha,\gamma)}\quad ,y=1,2,3, \qquad (3)$$

The details about derivation of the beta-geometric distribution and other related proofs may be seen in the Appendix.

Since $B(\alpha,\gamma)\frac{\Gamma(\alpha)\Gamma(\gamma)}{\Gamma(\alpha+\gamma)}$,

then the above pmf can be rewritten as:

$$P(y;\alpha,\gamma)=\frac{\alpha\Gamma(\gamma+y-1)\Gamma(\alpha+\gamma)}{\Gamma(\alpha+\gamma+y)\Gamma(\gamma)}\quad ,$$

y=1,2,3……..; α,γ>0 (4)

As a special case, if γ = 1, then the beta-geometric distribution reduces to the Yule-distribution.[33] Moreover, the formula

$$B(\alpha,\gamma)=\frac{(\alpha-1)!(\gamma-1)!}{(\alpha+\gamma-1)!}$$

leads to another form of the pmf:

$$P(y;\alpha,\gamma)=\frac{\alpha(\gamma+y-2)!(\alpha+\gamma-1)!}{(\alpha+\gamma+y-1)!(\gamma-1)!}.$$

(5)

A re-parameterization proposed by Griffiths36 yields the following expression

$$P(Y=y;\pi,\theta)=\frac{\pi\prod_{r=1}^{y-2}[1-\pi+r\theta]}{\prod_{r=1}^{y-1}[1+r\theta]},\ \mathrm{y}=1,2,\ldots$$

(5)

where $\pi=\frac{\alpha}{\alpha+\gamma}$ and $\theta=\frac{1}{\alpha+\gamma}$ are interpreted as the mean parameter and the shape parameter, respectively. Next, we derive the mean and variance of the beta-geometric distribution.

The plots of the pmf of the Beta-geometric distribution for various values of α and γ are given in Figure 2 below.

Mean and variance of the Beta-Geometric distribution

The mean and the variance of beta-geometric distribution are given by:

$$E(Y)=\frac{B(\alpha-1,\gamma)}{B(\alpha,\gamma)}=\frac{\alpha+\gamma-1}{\alpha-1}\quad \text{for}\quad \alpha>1. \qquad (7)$$

$$Var(Y)=\frac{\alpha\gamma(\alpha+\gamma-1)}{(\alpha-1)^2(\alpha-2)}$$

, or under reparameterization,

$$V(Y)=\frac{\pi(1-\pi)(1-\theta)}{(\pi-\theta)^2(\pi-2\theta)},\ for\ \pi>2\theta,\alpha>2$$

(8)

Estimation of the parameters α and γ

For estimating the parameters of the Beta-geometric model, we used three well known methods of estimations. These are: method of moments, maximum likelihood and Bayesian estimation.

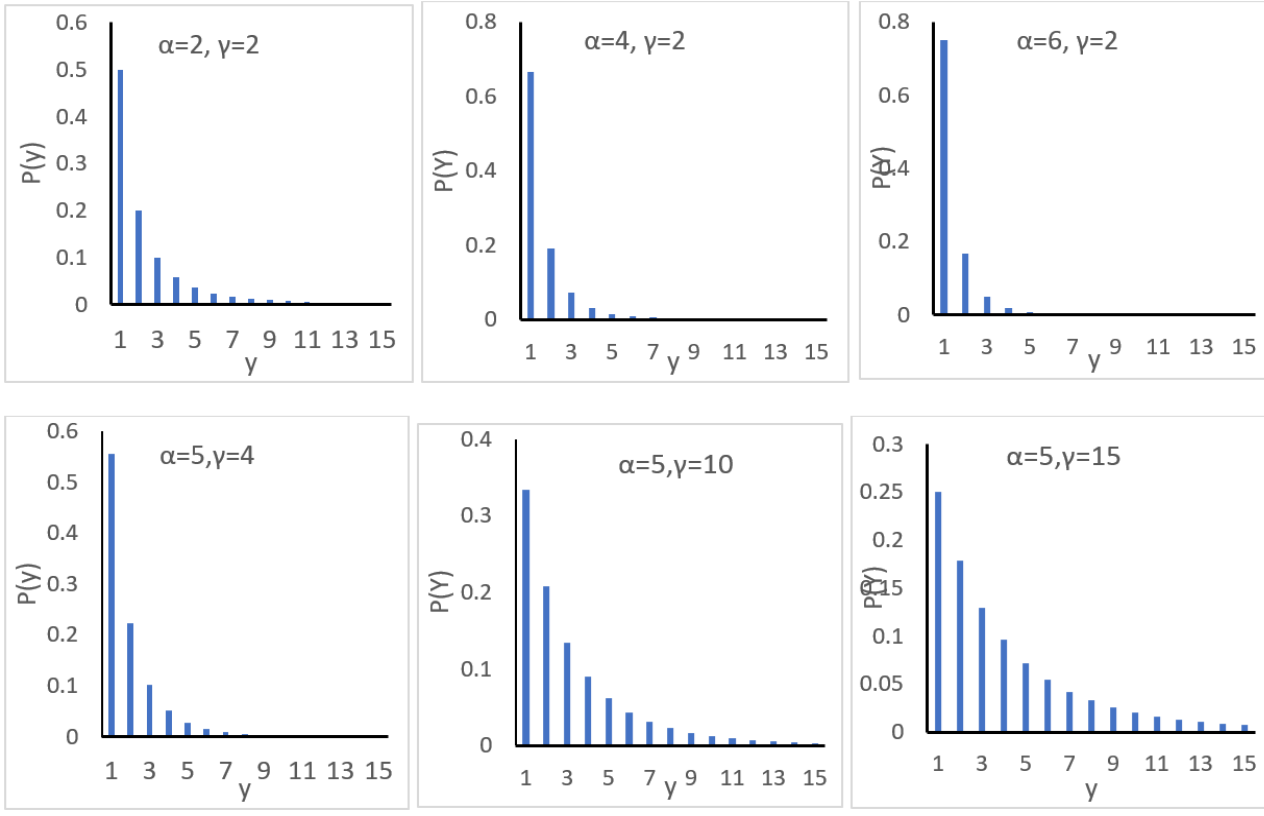

Figure 2. The pmf plot of Beta-geometric distribution for varying values of the parameter α and γ.

Method of moments

The method of moments consists of calculating a few raw moments of the observed sample values and equating them to the corresponding estimates of the population moments, thus getting as many equations as are needed to solve for the unknown parameters.

Let y_1, y_2, ……,y_n be a random sample n observation on time to first ANC visits which follow beta-geometric distribution. Then the rth sample raw moments is defined as

$$m_r = \sum_{i=1}^{n} y_i^r / n$$

The corresponding rth population moments (μ_r) of the distribution is defined as

$$\mu_r = E(Y^r) = \int Y^r f(Y) dy$$

Since we have two unknown parameters, we need to equate, $\mu_1 = m_1$ and $\mu_2 = m_2$. Thus we have

$$\mu_1 = \frac{(\alpha+\gamma-1)}{(\alpha-1)} = m_1$$

$$\mu_2 = \frac{(\alpha+\gamma-1)(\alpha+\gamma-2)}{(\alpha-1)(\alpha-2)} = m_2$$

Now solving these two equations, we derive the method of moment estimate for α and y as follows:

$$\hat{\alpha} = \frac{m_1^{\,2} + m_1 - 2m_2}{m_1^{\,2} - m_2}. \quad (10)$$

$$\hat{\gamma} = \frac{m_1^{\,2} - m_1 m_2 - m_1 + m_2}{m_1^{\,2} - m_2}. \quad (11)$$

Method of moments is the oldest and easiest method of point estimation of the population parameters, yielding almost always some sort of estimate, and therefore considered as a good starting point. However, it lacks some desirable optimal properties of a good estimator. In this respect, most statistician prefer maximum likelihood method of estimation, because it satisfies most of the optimal properties of a good estimator such as, minimum variance, unbiasedness, consistency, etc.

Maximum Likelihood Method

Maximum likelihood estimates (MLE) of the parameters are obtained by finding the parameter values that maximize the likelihood function which is define as

$$L(\theta) = \prod_{i=1}^{n} f(y_i y, \theta) \ .$$

The likelihood function of the beta-geometric distribution is given by:

$$L(y;\alpha,\gamma) = \prod_{i=1}^{n} \frac{B(\alpha+1, y_i+\gamma-1)}{B(\alpha,\gamma)}. \quad (12)$$

To make it simpler for finding the derivatives, we can use the re-parameterized form of the beta-geometric distribution as given in (11). Then the likelihood function is given as:

$$L(y;\theta,\pi) = \pi^n \prod_{i=1}^{n} \frac{\prod_{r=1}^{y_i-1}[1-\pi+(r-1)\theta]}{\prod_{r=1}^{y_i}[1+(r-1)\theta]}$$

and the corresponding log-likelihood can be written as

$$l = n\log\pi + \sum_{i=1}^{n}\left[\sum_{r=1}^{y_i-1}\log\{1-\pi+(r-1)\theta\} - \sum_{r=1}^{y_i}\log\{1+(r-1)\theta\}\right] \quad (13)$$

The maximum likelihood estimates $\hat{\pi}$ and $\hat{\theta}$, and thus $\hat{\alpha}$ and $\hat{\beta}$ are obtained by solving the maximum likelihood estimating equations

$$\frac{\partial l}{\partial \pi} = 0 \text{ and } \frac{\partial l}{\partial \theta} = 0 \text{ simultaneously.}$$

We have

$$\frac{\partial l}{\partial \pi} = \frac{n}{\pi} - \sum_{i=1}^{n}\left[\sum_{r=1}^{y_i-1}\frac{1}{1-\pi+(r-1)\theta}\right] = 0. \quad (14)$$

and

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{n}\left[\sum_{r=1}^{y_i-1}\frac{r-1}{1-\pi+(r-1)\theta} - \sum_{r=1}^{y_i}\frac{r-1}{1+(r-1)\theta}\right] \quad (15)$$

It is to be noted that there is no closed form solution for the above two equations. However, we can use some numerical iterative procedures such as the Newton-Raphson method or a numerical subroutine of the R-Environment to obtain the MLE of parameters.

Although the method of maximum likelihood is often the estimation method that mathematical statisticians prefer because of its good statistical property, sometimes complications arise in its use because the equation(s) obtained from $\frac{\partial l}{\partial \theta}=0$ may be difficult to solve, as we have seen incase of our equations (14) and (15).

In both method of moments and maximum likelihood estimation method, probabilities are interpreted as relative frequencies obtained from a sample. There is another approach in statistical inference, called the Bayesian approach, that combines sample information with another prior information.

Bayesian Estimation

In Bayesian approach, instead of considering the parameter of interest as constant, it is assumed as random variable with a specific distribution, called prior density function. Then a posterior density function is obtained from the joint density function of the sample observations and the prior density function. Finally, the mean of the posterior distribution is taken as the estimate of the parameter.

To obtain the Bayesian estimate of the parameters, we first derived the posterior distribution, assuming a beta prior distribution of p, because beta distribution is the proper conjugate of geometric model within parameter space [0, 1]. Following the Bayes' rule the posterior density function is given by

$$f(p\mid y)=\frac{f(p,y)}{f(y)}=\frac{f(y\mid p)f(p)}{f(y)},$$

where *f(p,y)* is the joint distribution, *f(p)* is the prior which is the beta distribution and *f(y)* is the marginal distribution of the variable which is geometric distribution.

According to the de Finetti's theorem, we can rewrite the posterior density function as:

$$f(p\mid y)=\frac{L(y\mid p)f(p)}{\int_0^1 L(y\mid p)f(p)\,dp}, \quad (16)$$

where *L(y|p)* is the likelihood function of the data distribution, and in our case it is the likelihood function of the geometric distribution. After applying the above theorem, we have

$$f(p,y)=p^n(1-p)^{\sum_{i=1}^n y_i-1}\times\frac{p^{\alpha-1}(1-p)^{\gamma-1}}{B(\alpha,\gamma)}$$

$$=\frac{p^{\alpha+n-1}(1-p)^{\gamma+\sum_{i=1}^n y_i-1-1}}{B(\alpha,\gamma)}.$$

The posterior function is evaluated to be as follows:

$$f(p\mid y)=\frac{p^{\alpha+n-1}(1-p)^{\gamma+\sum_{i=1}^n y_i-1-1}}{B(\alpha,\gamma)}$$

$$\div\frac{B\left(\alpha+n,\gamma+\sum_{i=1}^n y_i-1\right)}{B(\alpha,\gamma)}$$

$$=\frac{p^{\alpha+n-1}(1-p)^{\gamma+\sum_{i=1}^n y_i-1-1}}{B\left(\alpha+n,\gamma+\sum_{i=1}^n y_i-1\right)},\quad 0<p<1. \quad (17)$$

Therefore, our posterior model is a beta distribution with $(\alpha+1,\gamma+\sum y_i-1)$ parameters. This can be written in another form as follows:

$$f(p\mid \boldsymbol{y})=\frac{\tilde{\Gamma}\left(\alpha+\gamma+\sum_{i=1}^n y_i\right)}{\tilde{\Gamma}(\alpha+n)\tilde{\Gamma}\left(\gamma+\sum_{i=1}^n y_i-1\right)}$$

$$p^{\alpha+n-1}(1-p)^{\gamma+\sum_{i=1}^n y_i-1-1}$$

Bayesians believe that everything which needs to be known about the parameter is summarized

in the posterior pdf *f(y|p)*. There are many ways to find a point estimate of the parameter using the posterior distribution. However, the mean of the posterior distribution is taken as the best estimator of the parameter. We have

$$E(p \mid y) = \frac{B\left(\alpha + n + 1, \gamma + \sum_{i=1}^{n} y_i - 1\right)}{B\left(\alpha + n + 1, \gamma + \sum_{i=1}^{n} y_i - 1\right)},$$

which can further be simplified as

$$E\left(p \mid y\right) = \frac{\alpha + n}{\alpha + \gamma + \sum_{i=1}^{n} y_i + n - 1}. \quad (19)$$

Thus the Bayes estimator of the parameter *p* is:

$$\hat{p} = \frac{\alpha + n}{\alpha + \gamma + \sum_{i=1}^{n} y_i + n - 1}.$$

Without loss of generality, one can assume the value of and obtained by method of moment and maximum likelihood. If the sample results are inconsistent with the prior assumptions, the Bayes estimate may differ considerably from the maximum likelihood estimate. In these situations, the maximum likelihood estimate would be the better estimate to use.

Beta-geometric Regression

In a regression model framework, typically the mean of the response variable Y is modeled as a linear function of the predictor vector X. According to the Generalized Linear Model (GLM) framework,[51,52] we need a link function to obtain a functional relationship between the mean of the response variable and the linear predictors. There are several link functions available. One of these is the identity link, given by $g(\mu_i) = \mu_i = X_i'\beta$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients $(p < n)$, and $(xi_1, xi_2, \ldots, xi_p)$ denotes the observations on p predictors or covariates. When identity link is used, $E(y_i) = \mu_i = X_i'\beta$ since $\mu_i = g^{-1}(X_i'\beta)$ However, the most suitable link function is the log link function, given by $g(\mu_i) = \ln(\mu_i) = X_i'\beta$ For the log link function, the relationship between the mean of the response variable and the linear predictors is $\mu_i = g^{-1}(X_i'\beta) = e^{X_i'\beta}$ (20)

The log link function is particularly attractive for count data because it ensures that all of the predicted values of the response variable will be nonnegative.[53]

The parameters of the Beta-geometric regression model can be obtained by the method of maximum likelihood (ML). If we have a random sample of n observations on the response y and the predictors X, then the log likelihood function of the Beta-geometric pmf is given by

$$\ln L(y, \beta) = \sum_{i=1}^{n} (y_i - 1) \ln\left(\frac{e^{X_i'\beta_i} - 1}{e^{X_i'\beta}}\right) - \sum_{i=1}^{n} \ln(e^{X_i'\beta})$$

Differentiating (6) with respect to β provides the score function and the information matrix as

$$U(\beta) = \sum_{i=1}^{n} \frac{(y_i - e^{X_i'\beta})X_i'}{e^{X_i'\beta} - 1} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)X_i'}{(\mu_i - 1)}$$

and $I(\beta) = \sum_{i=1}^{n} \frac{\mu_i}{(\mu_i - 1)} X_i'X_i$, respectively. (21)

The ML estimator of β is obtained by solving the equation Unfortunately, there is no closed-form expression for the solution of the ML estimate of using above equation, hence its solution has to be performed numerically. However, a Beta-geometric regression algorithm can be developed with the any programming language, e.g. SAS's IML, STATA's ML capabilities, or by programming in R. In this study, we have used programming

in R for estimating the parameters.

Results

Application to real data

To illustrate the application of the proposed Beta-geometric regression, data was obtained from the 2000 Oman National Health Survey (ONHS). The survey was conducted by the Ministry of Health of Oman in collaboration with the UN Organizations such as UNFPA, UNICEF, WHO and the UN Statistics Division. Ever-married women aged 15-49 years from Omani nationals only were considered as respondents in the survey. The details of the survey may be seen elsewhere.[54]

The survey covered a nationally representative sample of 2,037 Omani married women selected following a multistage stratified probability sampling design. Among the respondents, 1,299 women had at least one ANC visit for their last live birth that occurred in the five years prior to the survey date. This study considered individual women's record of timing of first antenatal care (ANC) visit to health personnel during the pregnancy period of their last birth. Although, time is a continuous variable, however, survey data on time to first ANC visit occur as a count variable, such as the first ANC visit occur in month 1, 2, 3 …,and so on, during pregnancy. Thus the time-to-first ANC visit is a random variable that count the number of trails to obtain the first success in a series of independent and identical Bernoulli trails. In this application our response variable $Yi = i$, where i is the count denoting the number of months required to have first ANC visit.

Table 1 presents the distribution of the women according to the month of first ANC visit during the pregnancy period. Since in this study we have considered only women with at least one ANC visit, Yi take only non-zero positive integers starting from 1. The data indicate that most of the mothers (58%) received the first ANC visit within the first trimester of pregnancy and three-fourth (75%) received first ANC visit in 4th month or16 week of gestation. It is worth mentioning that at least one ANC visit during pregnancy is almost universal in Oman, and most of the women had ANC visit during 2nd or 3rd month of pregnancy.

Table 1. Distribution of women according to the month of first ANC visit

Month of first visit (Y_i)	Frequency	Percentage
1	195	15.01
2	280	21.56
3	282	21.71
4	223	17.17
5	205	15.78
6	66	5.08
7	36	2.77
8	8	0.62
9	4	0.31
Total	1299	100.0

Figure 3 depicts the graphical representation of the distribution of month of first ANC visit. It is evident from the graph that the distribution of the time to first ANC visit is skewed to the right. The data also indicate that the mean of the distribution (3.3 months) is higher than its variance (2.7 months), indicating that the distribution is under-dispersed. It, therefore, violate the principle of equi-dispersion (mean = variance) of Poisson distribution and over-dispersion (variance > mean) of Negative Binomial distribution, and thus may not be suitable for modelling with Poisson or

Negative Binomial Distribution. However, it can be modelled with geometric distribution as the geometric distribution capture both under-dispersion and over-dispersion in the data set.

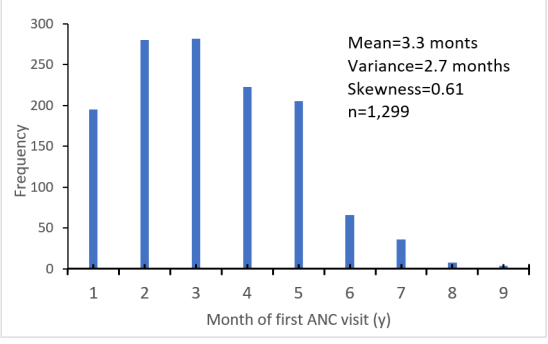

Figure 3. Distribution of time to first ANC visit

Table 2 presents the estimates of the parameters of Beta-geometric model for the given data using method of moments, maximum likelihood method and method of Bayesian estimation. The results in Table 2 revealed that all the three methods provide very close estimates of the parameter α with lowest ($\hat{\alpha}$ =4.747) for method of moments and highest for Bayesian method ($\hat{\alpha}$ =5.053). However, the estimates of γ vary with methods, having lowest ($\hat{\gamma}$ =5.355) for ML method and highest for method of moments ($\hat{\gamma}$ =8.544). It is possible that different methods of finding estimates of the parameters produce the same results, which makes the evaluation of the methods of estimation a bit easier, however, in many cases, different methods may lead to different estimates. However, the mean squared errors indicate that all the three methods considered here are more or less equally efficient.

To illustrate the application of the proposed Beta-geometric regression model, we applied the model to the count data on time to first ANC visit, obtained from the 2000 Oman National Health Survey (ONHS). To examine the model performance, we made a comparative analysis of the fitting of the geometric, Negative Binomial and Beta-geometric regression model with the 2000 ONHS data. Results of the three regression models are compared based on their respective deviance, log likelihood and the AIC and BIC values as presented in Table 3. Based on the model goodness of fit criterions, Beta-geometric model appeared to outperform the other two models, as it has highest log likelihood value and the lowest AIC and BIC values. The geometric model closely follow the Beta-geometric model, while NB model showed poor performance with lowest log likelihood and highest AIC and BIC values. Thus we applied Beta-geometric regression model for analyzing the predictors of first ANC visit considering few selected predictors such as maternal age at the time of last birth, education, marital status, place of residence, region of residence, employment status and parity of mothers The results are presented in Table 4.

Table 4 lists the estimates of regression coefficients, standard errors (SEs) of the estimated coefficients, value of the test statistics, P-value and the 95% confidence interval (CI). The results

Table 2. Estimates of parameters and the Mean squared Errors (MSEs)

Method of estimation	Parameters		MSE
	$\hat{\alpha}$	$\hat{\gamma}$	
Method of moments	4.747	8.544	12.811
Method of maximum likelihood	4.949	5.355	12.875
Bayesian Method	5.053	7.874	12.822

of the beta regression analysis presented in Table 6. The results indicate that, after controlling the other factors, women education and their urban/rural place of residence and region of residence appeared as significant predictors of the time to first ANC visits.

Table 3. Comparison of goodness of fit of Geometric, Negative Binomial and Beta-geometric regression model

Criterion	Geometric	Negative binomial	Beta-geometric
Deviance	1034.561	1058.872	1021.637
Log likelihood	-2592.152	-3016.991	-2160.571
AIC	5252.304	6067.982	4355.142
BIC	5340.182	6155.861	4443.0209
df	1282	1282	1282
Deviance/df	0.8069	0.8244	0.7969

Table 4. Beta-geometric regression analysis of the time to first ANC visits

Variables	B	SE	Test statistics (Z)	95% CI	P-value
Intercept	1.259	0.106	11.905	(1.020,1.392)	<0.001
Women age at birth of child					
15-20 (ref.)	0[a]	.		.	.
20-24	-0.009	0.072	-0.123	(-0.150,0.097)	0.902
25-29	-0.021	0.076	-0.277	(-0.149,0.109)	0.782
30-34	-0.059	0.077	-0.78	(-0.203,0.059)	0.435
35+	-0.002	0.079	-0.031	(-0.099,0.166)	0.975
Education Level					
No education	0[a]	.		.	.
Some primary	-0.108	0.055	-1.974	(-0.243,-0.056)	0.048
Primary/preparatory	-0.142	0.049	-2.935	(-0.268,-0.103)	<0.001
Secondary+	-0.213	0.061	-3.473	(-0.367,-0.146)	<0.001
Marital Status					
Currently Married	0[a]	.		.	.
Divorced/Separated/Widowed	0.109	0.071	1.529	(-0.017,0.218)	0.126
Place					
Urban	0[a]	.		.	.
Rural	0.139	0.043	3.172	(0.074,0.217)	0.002
Region					
Muscat	0[a]	.		.	.
Al-Batina	0.045	0.051	0.879	(-0.077,0.106)	0.379
Dhofar	0.253	0.069	3.617	(0.100,0.329)	<0.001
Al-Sharqiah	-0.160	0.066	-2.396	(-0.263,-0.009)	0.016
Al-Dhakhlia	0.111	0.065	1.696	(-0.006,0.213)	0.089
Al-Dhahirah	0.092	0.074	1.239	(-0.111,0.153)	0.215
Employment status					
Employed	0[a]	.		.	.
Not employed	-0.012	0.061	-0.201	(-0.130,0.092)	0.840
Parity					
Primi-parous	0[a]	.		.	.
Multi-parous	0.103	0.064	1.605	(-0.077,0.155)	0.108

Discussions

Modeling count data using the standard negative binomial (NB) model has recently become a foremost method of analyzing count response models, yet relatively few researchers or applied statisticians are familiar with the varieties of available NB models, or how best to incorporate them into a research plan.[2] Hilbe[2] in his book on "Negative Binomial Regression" indicated geometric regression as an special case of the Negative Binomial regression, and suggested that it could be used as an alternative model to address specific facts in the data that give rise to overdispersion and thus altering the distributional assumptions of the Poisson distribution.

Although geometric distribution, is a special case of NB distribution, and also belongs to the discrete family of distributions, little attention has been paid in modeling count data with the geometric distribution. There are many real-life phenomena that follow the geometric distribution. In the context of women reproductive health, one important variable is the time to first ANC visit, which occur as a count variable in survey data such as first ANC visit occur in 1st moth, 2nd month and so on. This type of data follow a geometric distribution. Since the distributional assumption plays critical role in statistical inference, it is important to choose the appropriate model for data analysis.

Al-Balushi and Islam[28] used geometric regression for analyzing data on time to first ANC visit. The limitation of the geometric regression model is that it assume constant probability of fist visit (success) for all women, which is unrealistic, and thus may affect the statistical inference. In this study, we extended the geometric regression model by relaxing the assumption of constant probability of success, and assumed that it follows a beta distribution. The resulting model is defined as beta-geometric model. The Beta-geometric distribution can be thought of being composed of two pieces: the probability that success will occur, and the success follow a beta distribution with two shape parameter α and γ which must always be positive numbers. This combination creates a powerful tool in mathematical modelling. The use of the Bayesian approach in the construction of the Beta-geometric model seems to be the key that explains its performance. Indeed, the idea of treating the probability of first success, p, as random, is very reasonable.

For estimating the two parameters α and γ of beta-geometric distribution, three methods of estimation were applied, including method of moment, MLE and Bayesian method. All the three methods provided very close estimate of α, but slight variation in γ. However, the MSEs were observed to be almost same.

The suitability of the Beta-geometric regression for analyzing the count data on time to first ANC visit was illustrated using real life data. The fitting of two other similar type of regression model, such as simple geometric and Negative Binomial (NB) models, are compared with the beta-geometric model. Based on the model goodness of fit criterions, Beta-geometric model appeared to outperform the simple geometric and the NB models. It is worth mentioning here that unlike NB regression which can generically accommodates only over dispersion, beta-geometric regression can be used for both over-dispersed and under-dispersed count data. In addition beta-geometric regression model does not need an additional parameter to accommodate underdispersion

or overdispersion. For the purpose of further research, an alternative options could be the Kumaraswamy distribution instead of the beta distribution and then fitting the count data with a new "Kumaraswamy-geometric" model, since the beta and Kumaraswamy[55] distributions share similar properties and support [0, 1]. The limitation of the beta-geometric model is that it is mainly suitable for modeling the count data that follow negative exponential distribution and less flexible for unimodal distribution. The beta-geometric model would be less ideal in a situation where the probability of success is not a random variable.

Conclusion

Beta-geometric regression model could be a flexible choice for analyzing real world count data set, and we expect that the Beta-geometric model may serve as an alternative model to the widely used Poisson or NB models for modeling count data with overdispersion or underdispersion. Further research is needed to test the adequacy of Beta-geometric model for analyzing the real-life phenomenon that follow the geometric distribution with varying probability of success such as the number of drills in an area before observing the first productive well by an oil prospector, the number of tosses of a fair coin before the first head (success) and so on.

References

1. Lewis-Beck, M. Applied regression: An introduction. Newbury Park, CA: Sage, 1989.

2. Hilbe JM. Negative Binomial Regression. Cambridge University Press New York, NY, 2012, pp. 180-185.

3. Coxe S, West SG & Aiken LS. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. Journal of personality assessment. 2009;91(2):121-136.

4. Sellers KF & Shmueli G. A flexible regression model for count data. The Annals of Applied Statistics.2010; 4(2): 943-961.

5. Cameron A & Trivedi PK. Econometric models based on count data: Comparisons and applications of some estimators and tests. Journal of Applied Econometrics.1986; 1: 29-53.

6. Winkelmann R & Zimmermann KF. Recent developments in count data modelling: theory and application. Journal of economic surveys.1995; 9(1): 1-24.

7. Wang W & Famoye F. Modeling household fertility decisions with generalized Poisson regression. Journal of Population Economics.1997;10(3): 273-283.

8. Mullahy J. Specification and testing in some modified count data models. Journal of Econometrics. 1986; 33: 341-365.

9. Cameron AC & Trivedi PK. Regression Analysis of Count Data. Cambridge University Press, Cambridge, 1998.

10. Johnson NL, Kotz S & Kemp AW. Univeriate discrete distribution, 3rd edition, John Wiely &Sons Inc., Hoboken, New Jersy, 2005.

11. Majumdar M, & Sheps M. Estimators

of a Type I Geometric Distribution From Observations on Conception Times. Demography. 1970; 7:349-360.

12. Sheps M & Menken J. Mathematical Models of Conception and Births, Chicago: University of Chicago Press, 1973.

13 Paul SR. Testing goodness of fit of the geometric distribution: an application to human fecundability data. Journal of Modern Applied Statistical Methods. 2005; 4(2): 8.

14. Sozou PD, Hartshorne GM. Time to Pregnancy: A Computational Method for Using the Duration of Non-Conception for Predicting Conception. PLoS ONE. 2012; 7(10): e46544. doi:10.1371/journal.pone.0046544

15. Banik AD, Chaudhry ML & Kim JJ. A Note on the Waiting-Time Distribution in an Infinite-Buffer GI[X]/C-MSP/1 Queueing System. Journal of Probability and Statistics, 2018; 2018: 1-10. https://doi.org/10.1155/2018/7462439

16. World Health Organization(WHO). WHO recommendations on antenatal care for a positive pregnancy experience. Geneva, 2016.
http://apps.who.int/iris/bitstream/10665/250796/1/9789241549912-eng.pdf?ua=1.

17. Pell C, Meñaca A, Were F, Afrah NA, Chatio S, Manda-aylor, L, et al. Factors Affecting Antenatal Care Attendance: Results from Qualitative Studies in Ghana, Kenya and Malawi. PLoS ONE. 2013; 8(1): e53747.

18. Feleke1 SA, Mulatu MA, & Yesmaw YS. Medication administration error: magnitude and associated factors among nurses in Ethiopia. BMC Nursing.2015; 14: 53. s12912-015-0099-1.pdf

19. Guleman H & Berhane Y. Timing of First Antenatal Care Visit and its Associated Factors among Pregnant Women Attending Public Health Facilities in Addis Ababa, Ethiopia. Ethiop J Health Sci. 2017; 27(2), 139–146.

20. Gebresilassie B, Belete T, Tilahun W, Berhane B & Gebresilassie S. Timing of first antenatal care attendance and associated factors among pregnant women in public health institutions of Axum town, Tigray, Ethiopia, 2017: a mixed design study. BMC Pregnancy and Childbirth. 2019; 19: 340.

21. Suissa S & Blais L. Binary regression with continuous outcomes. Stat Med. 1995;14: 247–55.

22. Moser BK & Coombs LP. Odds ratios for a continuous outcome variable without dichotomizing. Stat Med.2004; 23:1843–60.

23. Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol. 1991; 44: 241–8.

24. Peacock JL, Sauzet O, Ewings SM & Kerry SM. Dichotomising continuous data while retaining statistical power using a distributional approach. Stat Med. 2012; 21:3089–103.

25. Preisser JS, Das K, Benecha H & Stamm JW. Logistic Regression for Dichotomized Counts. Stat Methods Med Res.2016; 25:3038–56.

26. Sroka CJ & Nagaraja HN. Odds ratios from logistic, geometric, Poisson, and negative binomial regression models. BMC Medical Research Methodology. 2018; 18: 112

27. Pradhan B & Kundu D. Bayes estimation and prediction of the two-parameter gamma distribution. Journal of Statistical Computation and Simulation, 2011; 81(9): 1187-1198.

28. Al-Balushi ZMD & Islam MM. Geometric regression for modeling count data on the time-to-first antenatal care visit. Journal of Statistics: Advance in Theory and Applications. 2020; 23(1): 35-57.

29. Dobson AJ & Barnett AG. An introduction to generalized linear models. CRC press, 2018.

30. Weinberg CR & Gladen BC. The beta-geometric distribution applied to comparative fecundability studies. Biometrics. 1986; 42(3):547-560.

31. Kemp AW. The q-beta-geometric distribution as a model for fecundability. Communications in Statistics-Theory and Methods. 2001; 30(11), 2373-2384.

33. Weinberg CR & Gladen BC. The beta-geometric distribution applied to comparative fecundability studies. Biometrics. 1986; 547-560.

34. Miller AJ. A queueing model for road traffic flow, Journal of the Royal Statistical Society, Series B. 1961; 23: 64–90.

35. Pielou EC. Runs of one species with respect to another in transects through plant populations. Biometrics. 1962; 18(4):579-593.

36. Griffiths DA. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics. 1973; 29:637 – 648.

37. Hughes G & Madden LV. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. Phytopathology. 1993 83:759-763.

38. Zamani H, Ismail N and Shekari M. Weighted negative binomial-Poisson Lindley with application to genetic data. J Biostat Epidemiol. 2018; 4(3): 136-141

39. Eugene N, Lee C, Famoye F. The beta-normal distribution and its applications. Comm. Stat. Theor. Meth. 2002; 31(4): 497–512.

40. Nadarajah, S. & Kotz, S. The beta Gumbel distribution. Math. Probl. Eng. 2004; 4 :323–332.

41. Famoye F, Lee C & Olumolade O. The beta-Weibull distribution. J. Stat. Theory Appl. 2005; 4(2): 121–136.

42. Kong L, Lee C, Sepanski JH. On the properties of of beta-gamma distribution. J. Mod. Appl. Stat. Meth. 2007; 6(1), 187–211.

43. Akinsete A, Famoye F, Lee C. The beta-Pareto distributions. Statistics. 2008; 42(6):547–563

44. Alshawarbeh E, Famoye F, Lee C. Beta-

Cauchy distribution: some properties and its applications. J. Stat. Theory Appl. 2013; 12: 378–391.

45. Aki S. & Hirano K. Estimation of parameters in the discrete distributions of order k. Annals of the Institute of Statistical Mathematics, 1989; 41(1):47-61.

46. Feller W. An Introduction to Probability Theory and Its Applications, second edition, Vol. I. Wiley, New York, 1968.

47. Ferguson TS. A Characterization of the Geometric Distribution. The American Mathematical Monthly. 1965; 72 (3): 256-260.

48. Arnold BC.Two characterizations of the geometric distribution. J. Appl. Prob. 1980;17: 570–573.

49. Gultekin O. & Bairamov İ. A trivariate geometric distribution, characterization and asymtotic distribution. Ege University Journal of the Faculty of Science. 2013; 37(1): 1-18.

50. Tripathi RC, Gupta RC & White TJ.(1987). Some generalizations of the geometric distribution. Sankhya, Series B. 1987; 49: 218–223.

51. Nelder JA & Wedderburn, RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972; 135(3): 370-384.

52. McCullagh P & Nelder JA. Generalized Linear Models 2nd Edition Chapman and Hall. London, UK, 1989.

53. Montgomery DC, Peck EA & Vining GG. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons, Inc., New Jersey, USA., 2012.

54. Riyami A, Afifi M, Al-Kharusi H & Morsi M, National Health Survey, Volume 2, Reproductive Health Survey, Ministry of Health, Muscat, Oman, 2000.

55. Kumaraswamy P. A generalized probability density function for double-bounded random processes. Hydrology. 1980; 46, 79–88.