Journal of Statistics: Advances in Theory and Applications
Volume 23, Number 1, 2020, Pages 35-57
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122148

GEOMETRIC REGRESSION FOR MODELLING COUNT DATA ON THE TIME-TO-FIRST ANTENATAL CARE VISIT

ZAINAB MOHAMMED DARWISH AL-BALUSHI and M. MAZHARUL ISLAM

Department of Statistics
College of Science
Sultan Qaboos University
Muscat
Oman
e-mail: mislam@squ.edu.om

Abstract

Geometric distribution belongs to the family of discrete distribution that deals with the count of trail needed for first occurrence or success of any event. However, little attention has been paid in applying the GLM for the geometric distribution, which has a very simple form for its probability mass function with a single parameter. In this study, an attempt has been made to introduce geometric regression for modelling the count data. We have illustrated the suitability of the geometric regression model for analyzing the count data on time to first antenatal care visit that displayed under-dispersion, and the results were compared with Poisson and negative binomial regressions. We conclude that the geometric regression model may provide a flexible model for fitting count data sets which may present over-dispersion or under-dispersion, and the model may serve as an alternative model to the very familiar Poisson and negative binomial models for modelling count data.

2010 Mathematics Subject Classification: 60E05, 62H10, 62J12, 62P10.
Keywords and phrases: geometric regression, count data, under-dispersion, GLM.
Received August 7, 2020

1. Introduction

Antenatal care (ANC) is the routine care of pregnant women starting from the date of conception to onset of delivery. Within the continuum of maternity care, visit to health profession for ANC provides a platform in preventing health problems of mothers and the fetus [1]. According to the new guidelines of the World Health Organization (WHO), every mother should have at least eight ANC visits during the pregnancy period to reduce the risk of adverse pregnancy outcomes and improve the maternal and fetus health [2]. It also emphasizes that all pregnant mothers should start ANC visit within the first trimester of pregnancy (i.e., gestational age of < 12 weeks). Timely initiation of ANC is crucial for early detection of pregnancy related problems and adverse pregnancy outcomes and other complications [2, 3]. Timing of first ANC visit has been observed to predict the compliance of full coverage of WHO recommended contents of care [4].

Although, time itself is a continuous variable, but the time-to-first ANC visit occur as a *count variable*. A count variable is a variable that takes on discrete or isolated values representing the number of occurrences of an event in a fixed period of time. Usually, count variables are heteroskedastic, right skewed, and have a variance that increases with the mean of the distribution [5], and thus violate the basic assumption of normality and homoskedasticity of the standard statistical models. Data on time to first ANC visit usually recorded as the first ANC visit that occurred in first, second, or third trimester of pregnancy, or in first, second, third, …., nine month of pregnancy during the pregnancy period. 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, which can be modelled using a geometric distribution.

Until now, the most commonly used model for analyzing any count data is the Poisson model [6-8]. However the most serious weakness of the Poisson model is the imposed equality of conditional mean and variance of the response variable. Violation of equi-dispersion have effects similar to violation of heteroskedasticity in linear regression model. Inference based on equi-dispersion for over-dispersed or under-dispersed data is no longer valid, despite the fact that the parameter can still be estimated consistently [9]. If over-dispersion or under-dispersion is not accounted for, estimates of the standard errors will be too small, test statistics for the parameter estimates will be too large, significance will be overestimated, and confidence limits will be too small [10]. To overcome the problem of over-dispersion, Negative Binomial (NB) models are widely used for analyzing count data, and is well studied with statistical computational ability in many software (e.g., SAS, SPSS, R, etc.). The NB distribution, however, is unable to accommodate the under-dispersion. Recently, Castellares et al. [11] proposed Bell regression as an alternative to Poisson and NB regression. While a good number of studies have been done on over-dispersion and various distributions including NB distributions, generalized Poisson distribution and other distributions [5, 12-14] for modelling over-dispersed count data, under-dispersion in count data is less explored [15].

Although, geometric distribution belongs to the family of discrete distribution, little attention has been given for modelling count data with Geometric model. In this study, we introduce Geometric distribution for modelling the count data. On the basis of the Geometric distribution, we develop geometric regression model where the response variable is a count. In the generalized linear model setup, the mean response of the geometric regression model is related to a linear predictor through a link function, which allows for parameter interpretation in terms of the response variable in the original scale. We illustrate the application of geometric regression with a count data related to time to first visit for antenatal care (ANC) and discuss the interpretations of the coefficient,

testing the model fit and model fitting adequacy. We have also verified that this regression model may be a useful alternative to the usual Poisson and NB regression models for modelling count data.

2. The Geometric Regression

Let the response variable Y_i is a count of the number of trials needed to get the first success, such that the observation $Y_i = 1, 2, \ldots$. If the probability of success is $\boldsymbol{p}$, then the probability model for this count data is the geometric distribution, and its probability mass function (pmf) is given by

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

Geometric distribution also defined as the distribution of the number of trials until the first occurrence of the success. Geometric distribution is a special case of Negative Binomial (NB) distribution where the number of successes (r) is equal to 1. The geometric distribution has the interesting property of being memory less.

The mean and variance of the distribution given in (1) are respectively, $E(Y) = \mu = \frac{1}{p}$ and $\text{Var}(Y) = \frac{1-p}{p^2}$.

In a regression model framework, typically the mean of the response variable is modeled as a linear function of the predictors. So, to obtain a regression model for the mean of the Geometric distribution, we need to re-parameterized the geometric pmf by letting $E(Y) = \mu = \frac{1}{p}$ and hence $p = \frac{1}{\mu}$. It then follows that $E(Y) = \mu$, and $\text{Var}(Y) = \mu(\mu - 1)$, so that $\mu > 1$. The geometric pmf can then be written, in the new parameterization, as

$$P(Y_i = y_i|\mu_i) = \frac{1}{\mu_i}\left(1 - \frac{1}{\mu_i}\right)^{y_i - 1}, \; i = 1, 2, \ldots, n, \; \mu > 1. \tag{2}$$

It can be shown that the pmf (2) belongs to the one-parameter exponential family. Then, the variance function can easily be obtained, which is given by $\mathrm{Var}(Y) = \mu(\mu - 1)$. We have $\mathrm{Var}(Y) = E(Y)$ for $\mu = 2$, $\mathrm{Var}(Y) < E(Y)$ for $\mu < 2$ and $\mathrm{Var}(Y) > E(Y)$ for $\mu > 2$. This implies that the geometric distribution capture both under-dispersion and over-dispersion, and thus can be suitable for modelling count data with under-dispersion as well as over-dispersion. An added advantage of the geometric distribution in relation to the NB distribution is that it involves single parameter and no additional (dispersion) parameter is necessary to accommodate over or under-dispersion.

According to the Generalized Linear Model (GLM) framework [16, 17], 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. 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 $X_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denotes the observations on p known 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 predictor is

$$\mu_i = g^{-1}(X_i'\beta) = e^{X_i'\beta}. \tag{3}$$

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 [18].

The parameters of the geometric regression model is 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 likelihood function of the geometric pmf as given in (2) is

$$L(y, \beta) = \prod_{i=1}^{n} f_i(y_i, \mu_i) = \prod_{i=1}^{n} \frac{1}{\mu_i}\left(1 - \frac{1}{\mu_i}\right)^{y_i - 1}. \tag{4}$$

The log likelihood function is then appears as

$$\ln L(y, \beta) = \sum_{i=1}^{n} (y_i - 1) \ln\left(\frac{\mu_i - 1}{\mu_i}\right) - \sum_{i=1}^{n} \ln(\mu_i), \tag{5}$$

where $\mu_i = g^{-1}(X_i'\beta) = e^{X_i'\beta}$ (considering log link function). Thus, in terms of β, the log likelihood function can be expressed as

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

$$= \sum_{i=1}^{n} \left[(y_i - 1) \ln(e^{X_i'\beta} - 1) - y_i \ln(e^{X'\beta})\right]. \tag{6}$$

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.

The ML estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)'$ of β is obtained by solving the equation $U(\beta) = 0$. Unfortunately, there is no closed-form expression and hence its solution has to be performed numerically. One can use the scoring method with Newton-Raphson iterative procedure and Fisher

Information matrix [19]. The estimating equation for the method of scoring using the Newton-Raphson iterative procedure is given by

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [I^{(m-1)}(\beta)]^{-1} U^{(m-1)}(\beta), \tag{7}$$

where $\hat{\beta}^{(m)}$ is the vector of estimates of the parameters $\beta_1, \beta_2, \ldots, \beta_p$ at the m-th iteration. Using Equation (7) and any software with a weighted linear regression routine, the ML estimate $\hat{\beta}$ can be computed iteratively.

Using the log likelihood of the geometric distribution, we can get the deviance statistic as

$$D = 2\left[\sum_{i=1}^{n}\left((y_i - 1)\ln\left(\frac{y_i - 1}{\hat{\mu}_i - 1}\right) + y_i \ln\left(\frac{\hat{\mu}_i}{y_i}\right)\right)\right], \tag{8}$$

which follows a chi-square distribution with (n-p) degrees of freedom. Thus the deviance D may be used as a measure of goodness-of-fit for the Geometric regression model fitted to real data; that is, the smaller the value of D, the better the fit to the real data.

3. Data and Method

To illustrate the application of the geometric regression, data for this study 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 [20].

A nationally representative sample of 2,013 Omani households was selected following a multistage stratified probability sampling design. Ultimately, 2,037 eligible women were successfully interviewed from 2013 selected households. The 2000 ONHS was household based community survey, facilitating details data collect on socio-economic and

demographic characteristics of households and household members as well as reproductive health information of eligible women respondents and their birth histories.

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 that occurred in the five years prior to the survey date, and who has at least one ANC visit. As a result, there were 1299 women who had at least one ANC visits for their last live birth, who constituted our study sample. It is worth mentioning here that delivery in health facilities and at least one ANC visit to health personnel is almost universal in Oman.

Our response variable Y_i is the timing of first ANC visit during the pregnancy period, which follows a geometric distribution. Thus $Yi = i$, where $\boldsymbol{i}$ is the count denoting the number of months required to have first ANC visit. Since in this study we have considered only women with at least one ANC visit, Y_i take only non-zero positive integers, we use Equation (1) as the pmf of the geometric distribution. For further analysis, the corresponding link function (Equation (3)), estimating Equation (7) and deviance (Equation (8)) were used. Since the geometric distribution can be obtained from the negative binomial distribution with heterogeneity or over-dispersion parameter, α, set to 1.0, GLM software that incorporates the negative binomial as a member family can also be used to design geometric models by setting the value of α to a constant value of 1.0 [5]. However, a geometric regression algorithm can be designed 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 geometric regression parameters. To illustrate the application of the proposed geometric regression model, a few selected covariates was considered. The selected covariates include maternal age at the time of last birth, education, marital status, place of residence, region of residence, employment status and parity. Prior to the

multivariate regression analysis using geometric regression, bivariate association between the timing of the first ANC visit and the selected characteristics of the mothers were measured by the cross tabulation of the mean time to first ANC visit by the selected characteristics of the mothers and the significance of association was tested by using the analysis of variance (ANOVA) technique. A P-value of < 0.05 was considered as significant.

4. Results and Discussions

Table 1 presents the distribution of the women according to the count of month of first ANC visit during the pregnancy period. 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 4^{th} month or 16 week of gestation. The mean timing of the first ANC visit was 3.3 months in Oman with standard deviation of 1.64 months. Figure 1 presents the histogram of the distribution of $Y_i = i$ (where $\boldsymbol{i}$ is the count denoting the number of months required to have first ANC visit). The histogram indicates 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 can capture both under-dispersion and over-dispersion in the data set.

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

Month of first visit	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

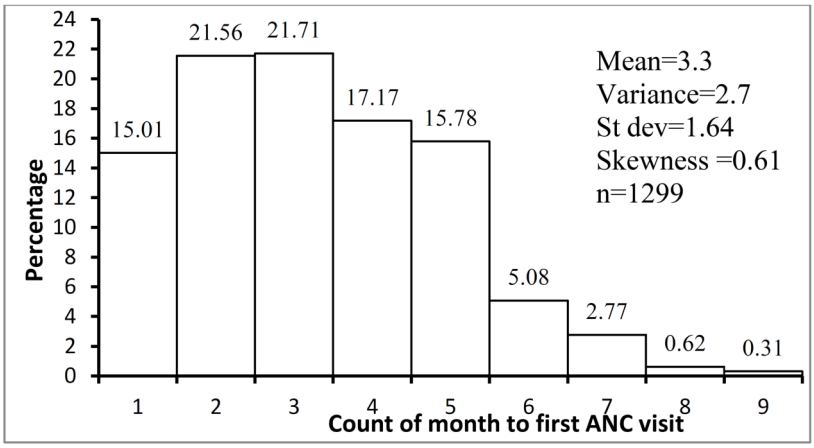

Figure 1. Histogram of the distribution of month of first ANC visit.

Table 2 presents the socio-economic and demographic characteristics of the mothers and the differentials of the mean time to first ANC visits across the characteristics. As can be seen, about 70% of the mothers were in their prime reproductive age between 20 to 34 years. The average age of the mothers was 28.5 (±SD 6.9) years. Majority (73%) of mothers were living in the urban area. Among the 6 region considered in this study, the selected sample of mothers vary from 8.6% in Al-Dhahirah region to 31.9% from Al-Batinah. About 36% of the mothers had no formal education, 18.4% had secondary and higher level of education. Most of the mothers (91%) were multiparous. About 8% mothers were either widowed or divorced or separated and 13.5% were employed.

Table 2 also presents the bivariate association between the timing of the first ANC visit and the selected characteristics of the mothers. Bivariate analysis identified maternal level of education, place of residence, region of residence, employment status and parity as the significant factors affecting the timing of the first ANC visit. As expected, the mean time to first ANC visit shows significant negative association with the mean time to first ANC visits, as the mean time to first ANC visit decreased with the increase of the level of education. For example, mothers with secondary and above level of education had a mean time to first ANC visit of 2.9 months compared to 3.6 months for the mothers with no education. Rural mothers were more likely to have delayed first ANC visit compared to mothers living in the urban area. Mothers with first pregnancy were more likely to have earlier first ANC visit than the mothers with multi-parity. Mean time to first ANC visit vary significantly across the region of residence. Mothers from the Al-Sharqiah region had shorter mean time to first ANC visit (2.9 months), while the mothers from Dhofar and Al-Dhakhlia had longer mean time to first ANC visit (3.6 and 3.5 months, respectively).

Table 2. Distribution of respondents and the mean time to first ANC visit, according to selected characteristics of mothers

Covariates	% (n)	Mean	*P*-value
Total	100.0(1299)	3.28	
Women age at birth of child			0.435
15-19	8.1(105)	3.39	
20-24	25.1(326)	3.18	
25-29	23.2(302)	3.30	
30-34	21.9(284)	3.21	
35+	21.7(282)	3.41	
Mean (±SD)	28.5(± 6.9)		
Education level			< 0.001
No education	36.1(469)	3.61	
Some primary	17.5(227)	3.22	
Primary/preparatory	28.0(364)	3.15	
Secondary+	18.4(239)	2.92	
Marital status			0.266
Currently married	92.3(1199)	3.27	
Divorced/separated/widowed	7.7(100)	3.46	
Place of residence			< 0.001
Urban	73.4(954)	3.18	
Rural	26.6(345)	3.59	
Region			< 0.001
Muscat	22.7(295)	3.15	
Al-Batina	31.9(414)	3.30	
Dhofar	11.6(151)	3.63	
Al-Sharqiah	11.8(153)	2.92	
Al-Dhakhlia	13.4(174)	3.54	
Al-Dhahirah	8.6(112)	3.20	
Employment status			0.049
Employed	13.5(176)	3.05	
Not employed	86.5(1123)	3.32	
Parity			0.028
Primi-parous	9.1(118)	2.97	
Multi-parous	90.9(1181)	3.32	

Bivariate analysis as discussed above, however, presents unadjusted association between outcome variable and the covariates. To obtain the adjusted association of a covariate on the outcome variable (i.e., the count of time to first ANC visit) after controlling the effect of all other covariates, we applied multiple regression analysis using geometric regression. The R programming was used for estimating the parameters of the regression model.

Table 3 lists the ML 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 deviance of the fitted geometric model was observed to be 1034.561 on 1282 degrees of freedom (df), and the deviance/df = $0.807 < 1$ indicate a good fit to data. The results of the geometric regression analysis presented in Table 3 indicate that, after controlling the other factors, women education and their urban/rural place of residence have significant effect on the time to first ANC visits.

Table 3. Geometric regression analysis of the time to first ANC visits

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Intercept	0.8655	0.1921	4.5055	(0.489, 1.242)	< 0.001
Women age at birth of child					
15-19	0[a]	.		.	.
20-24	– 0.0402	0.1300	– 0.3092	(– 0.295, 0.215)	0.756
25-29	– 0.0246	0.1369	– 0.1797	(– 0.293, 0.244)	0.857
30-34	– 0.0787	0.1379	– 0.5707	(– 0.349. 0.192)	0.568
35+	0.0185	0.1410	0.1312	(– 0.258, 0.295)	0.895
Education level					
No education	0[a]	.			.
Some primary	– 0.1557	0.0982	– 1.5855	(– 0.348, 0.037)	0.112
Primary/preparatory	– 0.1957	0.0867	– 2.2572	(– 0.366, – 0.026)	0.023
Secondary+	– 0.2063	0.1028	– 2.0068	(– 0.4077, – 0.005)	0.021
Marital status					
Currently Married	0[a]	.		.	.
Divorced/separated /widowed	0.1068	0.1275	0.8376	(– 0.143, 0.357)	0.402
Place					
Urban	0[a]	.		.	.
Rural	0.1595	0.0766	2.0822	(0.009, 0.306)	0.037
Region					
Muscat	0[a]	.		.	.
Al-Batina	0.0165	0.0944	0.1748	(– 0.169, 0.202)	0.861
Dhofar	0.2327	0.1221	1.9058	(– 0.007, 0.472)	0.056
Al-Sharqiah	– 0.1607	0.1255	– 1.2805	(– 0.407, 0.0852)	0.200
Al-Dhakhlia	0.1112	0.1164	0.9553	(– 0.117, 0.339)	0.339
Al-Dhahirah	0.0219	0.1352	0.1620	(– 0.243, 0.287)	0.870

Table 3. (Continued)

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Employment status					
Employed	0[a]	.		.	.
Not employed	– 0.0265	0.1131	– 0.2343	(– 0.248, 0.195)	0.814
Parity					
Primi-parous	0[a]	.		.	.
Multi-parous	0.0594	0.1192	0.4983	(– 0.174, 0.293)	0.618

To examine the model performance and parameter estimation, we make a comparative analysis of the geometric, Poisson and NB regression analysis of the count of the month required for first ANC visits. 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 4. Based on the model goodness-of-fit criterions, Poisson 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 Poisson model, while NB model showed poor performance with lowest log likelihood and highest AIC and BIC values. With the given under-dispersed data set, geometric model perform better than the NB model. The poor performance of NB model may be related to the fact that it cannot accommodate the under-dispersion of the given data set. NB model can accommodate only the over-dispersion. On the other hand, the one parameter geometric model can accommodate under-dispersion as well as over-dispersion.

Table 4. Comparison of goodness of fit of Geometric, Poisson and Negative binomial regression model

Criterion	Geometric	Negative binomial	Poisson
Deviance	1034.561	1058.872	1027.637
Log likelihood	– 2592.152	– 3016.991	– 2430.571
AIC	5252.304	6067.982	4895.142
BIC	5340.182	6155.861	4983.021
df	1282	1282	1282
Deviance/df	0.8069	0.8244	0.8016

A comparison of the ML estimates of the parameters (βs) by the Geometric, Poisson and NB regression models that are presented in Tables 3, 5 and 6, respectively, indicates that the estimates of the parameters under three models are very close, but they produced different standard errors (SEs). Because of equi-dispersion characteristics of the Poisson model, it produced lower SEs of the estimated parameters than those of geometric model and NB model. As a result, Poisson model is likely to fails in capturing the real value of the parameters when the data involved with over-dispersion or under-dispersion. It appear that the Poisson model for the mean may be correct but the true distribution is mis-specified; the ML estimates of model parameters can still be consistent but standard errors are incorrect. In such situation, the inference about the significance of the parameters would be misleading. In this study, we observed that the Poisson model identified more predictors as significant than the other two models. For example, Poisson regression identified women education, urban/rural place of residence, region of residence, and parity as the significant predictors of the time to first ANC visits, while geometric regression identified women education and urban/rural place of residence.

Table 5. Poisson regression analysis of the time to first ANC visits

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Intercept	1.202	0.0996	12.0683	(1.0453,1.392)	0.000
Women age at birth of child					
15-20	0[a]	.		.	.
20-24	– 0.045	0.0621	– 0.7246	(– 0.167, 0.077)	0.467
25-29	– 0.008	0.0632	– 0.1266	(– 0.132, 0.116)	0.905
30-34	– 0.065	0.0636	– 1.0220	(– 0.189, 0.061)	0.309
35+	– 0.003	0.0636	– 0.0472	(– 0.128, 0.122)	0.962
Education level					
No education	0[a]	.		.	.
Some primary	– 0.111	0.0448	– 2.4777	(– 0.199, – 0.024)	0.013
Primary/preparatory	– 0.134	0.0399	– 3.3584	(– 0.212, – 0.056)	< 0.001
Secondary+	– 0.187	0.0537	– 3.4823	(– 0.292, – 0.082)	< 0.001
Marital status					
Currently Married	0[a]	.		.	.
Divorced/separated/widowed	0.079	0.0571	1.3835	(– 0.033, 0.191)	0.169
Place					
Urban	0[a]	.		.	.
Rural	0.108	0.0348	3.1034	(0.040, 0.176)	0.002

Table 5. (Continued)

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Region					
Muscat	0[a]	.		.	.
Al-Batina	0.016	0.0436	0.3670	(– 0.069, 0.102)	0.706
Dhofar	0.161	0.0554	2.9061	(0.052, 0.269)	0.002
Al-Sharqiah	– 0.106	0.0591	– 1.7936	(– 0.222, 0.010)	0.067
Al-Dhakhlia	0.079	0.0530	1.4906	(– 0.025, 0.183)	0.135
Al-Dhahirah	0.020	0.0625	0.3200	(– 0.103, 0.142)	0.750
Employment status					
Employed	0[a]	.		.	.
Not employed	– 0.010	0.0526	– 0.1901	(– 0.114, 0.093)	0.796
Parity					
Primi-parous	0[a]	.		.	.
Multi-parous	0.060	0.0301	1.9931	(0.001, 0.119)	0.023

Table 6. Negative binomial regression analysis of the time to first ANC visits

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Intercept	1.208	0.2024	5.9684	(0.811, 1.605)	< 0.001
Women age at birth of child					
15-20	0[a]	.		.	.
20-24	– 0.052	0.1297	– 0.4009	(– 0.307, 0.202)	0.687
25-29	– 0.011	0.1333	– 0.0825	(– 0.272, 0.251)	0.936
30-34	– 0.064	0.1334	– 0.4798	(– 0.326, 0.197)	0.629
35+	– 0.011	0.1338	– 0.0822	(– 0.274, 0.251)	0.933
Education level					
No education	0[a]	.		.	.
Some primary	– 0.110	0.0935	– 1.1765	(– 0.199, – 0.023)	0.240
Primary/preparatory	– 0.135	0.0668	– 2.0209	(– 0.266, – 0.004)	0.022
Secondary+	– 0.189	0.0960	– 1.9687	(– 0.377, – 0.001)	0.024
Marital status					
Currently married	0[a]	.		.	.
Divorced/separated/widowed	0.076	0.1213	0.6265	(– 0.161, 0.314)	0.529
Place					
Urban	0[a]	.		.	.
Rural	0.111	0.0733	1.5143	(– 0.033, 0.255)	0.129

Table 6. (Continued)

Variables	B	SE	Test statistics (Z)	95% CI	*P*-value
Region					
Muscat	0[a]	.		.	.
Al-Batina	0.014	0.0897	0.1561	(– 0.162, 0.190)	0.876
Dhofar	0.162	0.1166	1.3894	(– 0.066, 0.391)	0.164
Al-Sharqiah	– 0.108	0.1182	– 0.9137	(– 0.339, 0.124)	0.361
Al-Dhakhlia	0.077	0.1112	0.6924	(– 0.141, 0.295)	0.490
Al-Dhahirah	0.018	0.1281	0.1405	(– 0.233, 0.269)	0.888
Employment status					
Employed	0[a]	.		.	.
Not employed	– 0.015	0.1071	– 0.1401	(– 0.225, 0.195)	0.890
Parity					
Primi-parous	0[a]	.		.	.
Multi-parous	0.058	0.0984	0.5894	(– 0.135, 0.251)	0.278

5. Conclusion

Geometric distribution belongs to the family of discrete distribution that deals with the count of trail needed for first occurrence or success of any event. The GLM of the geometric distribution can be used for modelling the factors associated with the count of trails need for the first success of any event. Considering the limitation of Poisson model and the NB model, geometric model can be used as an alternative modelling approach for both under-dispersed and over-dispersed count data. However, until now, little attention has been paid in applying the GLM for the geometric distribution. In this study, an attempt has been made to introduce geometric regression for modelling the count data. We have examined the suitability of the geometric regression for analyzing count of month needed for receiving the first ANC visit. The fitting of the geometric regression model was found to be good. A comparison of the performance of the geometric regression with Poisson regression and NB regression indicates that the estimates of the parameters under three models are consistent, but they produced different standard errors (SEs), which might have ramification on the inference about the significance of the model parameters. Based on the model goodness of fit criterions, Poisson model appeared to outperform the other two models, and geometric model perform better than NB model. It is worth mentioning here that, sometimes a model may appear well fitted to a particular data set, yet it would be incorrectly specified [5]. Although, Poisson regression appeared to better fit our real data, but there is a mis-specification of the true distribution, as the true distribution of the real data follow geometric distribution and the data exhibit under-dispersion. Thus, any inference based on the Poisson regression for the given data set would be misleading. In conclusion, the geometric regression model may provide a flexible model for fitting a wide spectrum of discrete real world data sets which may present over-dispersion or under-dispersion, and we expect that the geometric model may serve as an alternative model to the very familiar Poisson and NB models for modelling count data.

References

[1] World Health Organization (WHO), Antenatal Care Randemization Trial: Manual for Implementation of the New Model, WHO: Geneva, 2002.

[2] World Health Organization, Recommendations on Antenatal Care for a Positive Pregnancy Experience, World Health Organization: Geneva, 2016.

http://apps.who.int/iris/bitstream/10665/250796/1/9789241549912-eng.pdf?ua=1

[3] C. L. Abou-Zahr and T. M. Wardlaw, World Health Organization, Antenatal Care in Developing Countries: Promises, Achievements and Missed Opportunities: An Analysis of Trends, Levels and Differentials, 1990-2001, WHO: Geneva, 2003.

[4] J. E. Lawn, H. Blencowe, P. Waiswa, A. Amouzou, C. Mathers, D. Hogan et al., Stillbirths: Rates, risk factors, and acceleration towards 2030, Lancet 387(10018) (2016), 587-603.

DOI: https://doi.org/10.1016/S0140-6736(15)00837-5

[5] J. M. Hilbe, Negative Binomial Regression, Cambridge University Press, New York, 2011.

DOI: https://doi.org/10.1017/CBO9780511973420

[6] S. Coxe, S. G. West and L. S. Aiken, The analysis of count data: A gentle introduction to Poisson regression and its alternatives, Journal of Personality Assessment 91(2) (2009), 121-136.

DOI: https://doi.org/10.1080/00223890802634175

[7] K. F. Sellers and G. Shmueli, A flexible regression model for count data, Annals of Applied Statistics 4(2) (2010), 943-961.

DOI: https://doi.org/10.2307/29765537

[8] A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data, 2nd Edition, Econometric Society Monograph No. 53, Cambridge University Press, Cambridge, 2013.

[9] R. Winkelmann and K. F. Zimmermann, Recent developments in count data modelling: Theory and application, Journal of Economic Surveys 9(1) (1995), 1-24.

DOI: https://doi.org/10.1111/j.1467-6419.1995.tb00108.x

[10] W. Wang and F. Famoye, Modeling household fertility decisions with generalized Poisson regression, Journal of Population Economics 10(3) (1997), 273-283.

DOI: https://doi.org/10.1007/s001480050043

[11] F. Castellares, S. L. P. Ferrari and A. J. Lemonte, On the Bell distribution and its associated regression model for count data, Applied Mathematical Modelling 56 (2018), 172-185

DOI: https://doi.org/10.1016/j.apm.2017.12.014

[12] Z. Yang, J. W. Hardin, C. L. Addy and Q. H. Vuong, Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model, Biometrical Journal 49(4) (2007), 565-584.

DOI: https://doi.org/10.1002/bimj.200610340

[13] P. C. Consul and G. C. Jain, A generalization of the Poisson distribution, Technometrics 15(4) (1973), 791-799.

DOI: https://doi.org/10.2307/1267389

[14] R. Winkelmann and K. F. Zimmermann, Count data models for demographic data, Mathematical Population Studies 4(3) (1994), 205-221.

DOI: https://doi.org/10.1080/08898489409525374

[15] K. F. Sellers and D. S. Morris, Underdispersion models: Models that are "under the radar", Communications in Statistics - Theory and Methods 46(24) (2017), 12075-12086.

DOI: https://doi.org/10.1080/03610926.2017.1291976

[16] J. A. Nelder and R. W. A. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society: Series A 135(3) (1972), 370-384.

DOI: https://doi.org/10.2307/2344614

[17] P. McCullagh and J. A. Nelder, Generalized Linear Models, Second Edition, Chapman and Hall, London, 1989.

[18] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis, Fifth Edition, John Wiley and Sons, Inc., New Jersey, USA, 2012.

[19] A. J. Dobson and A. G. Barnett, An Introduction to Generalized Linear Models, Fourth Edition, Taylor and Francis Group, CRC Press, London, 2018.

DOI: https://doi.org/10.1201/9781315182780

[20] A. Riyami, M. Afifi, H. Al-Kharusi and M. Morsi, National Health Survey, Volume 2, Reproductive Health Survey, Ministry of Health, Muscat, Oman, 2000.

■