Exponentiated Weibull-geometric regression model

Felix Famoye
Department of Mathematics, Central Michigan University, Mt. Pleasant, MI 48859

Abstract

An exponentiated Weibull-geometric distribution is defined and studied. Some of its properties, such as unimodality and moments are discussed. The method of maximum likelihood estimation is proposed for estimating the model parameters. A count data regression model, based on the exponentiated Weibull-geometric distribution, is also defined. The regression model can be applied to fit an under-dispersed or an over-dispersed count data. Two numerical data sets are used to illustrate the applications of the exponentiated Weibull-geometric regression model.

Key Words and Phrases: Estimation; goodness-of-fit; under- and over-dispersion; zero-inflation.
2010 MSC: 62E15; 62F10; 62J12; 62P10.

1. Introduction

Many techniques for generating families of discrete distributions have been developed in the literature. See for examples the books by Balakrishnan and Nevzorov (2003), Johnson et al. (2005), Consul and Famoye (2006), and the references therein. These discrete distributions are found useful in many different areas of life. Frome et al. (1973) considered the Poisson distribution in the context of non-linear regression analysis for count data where the sample mean and sample variance are about equal. When the sample mean and sample variance are about equal, we have an equi-dispersion situation. When the sample mean is smaller (or greater) than the sample variance, we have over-dispersion (or under-dispersion) situation.

Many researchers obtained discrete distributions by discretizing continuous distributions. Nekoukhou and Bidram (2015) gave a long list of these works. Another method to generalize an existing distribution is by adding parameters to the distribution to form an exponentiated family (Lee et al., 2013 and the references therein). By exponentiating the cumulative distribution function of discrete Weibull distribution (Nakagawa and Osaki, 1975), Nekoukhou and Bidram (2015) defined the exponentiated discrete Weibull distribution.

Mahmoudi and Shiran (2012) defined an exponentiated Weibull-geometric distribution by compounding the exponentiated Weibull and geometric distribution to form a continuous distribution. In this paper, we define an exponentiated Weibull-geometric distribution by using the *T-R* framework proposed by Alzaatreh et al. (2013). This new distribution is a discrete distribution and it is the discrete analogue of the continuous exponentiated Weibull distribution. This is like calling the geometric distribution a discrete analogue of the exponential distribution.

Alzaatreh et al. (2013) introduced a general method for generating a probability distribution. Suppose we have a probability density function (PDF), $f_T(t)$, of a continuous random variable $T \in [a,b]$, $-\infty \le a < b \le \infty$ and a monotonic and absolutely continuous function $W\left(F_R(y)\right)$ of the cumulative distribution function (CDF) $F_R(y)$ for any random variable *R*. The CDF $F_Y(y)$ of a new random variable *Y* is given by

$$F_Y(y) = \int_a^{W[F_R(y)]} f_T(t)dt = F_T\left(W[F_R(y)]\right). \tag{1.1}$$

The distribution in (1.1) belongs to the *T-R* family. Many continuous distributions have been defined and studied by using the result in (1.1). In particular, Alzaatreh et al. (2012) defined the *T*-geometric family. This family consists of the discrete analogue to the distribution of the non-negative continuous random variable *T*. Furthermore, the authors defined and studied the exponentiated-exponential geometric distribution (EEGD).

In this article, an exponentiated Weibull-geometric distribution (EWGD) is defined and studied. The paper is organized as follows: In Section 2, the definition and some properties of EWGD are given. In Section 3, estimation of the parameters is considered along with some test and goodness-of-fit statistics for EWGD. An exponentiated Weibull-geometric regression (EWGR) model to fit a count response variable that follows the EWGD is defined in Section 4. A zero-inflated EWGR is also given in Section 4. In Section 5, the EWGR model is applied to two real life data sets and the results are compared with other count data regression models. Some concluding remarks are provided in Section 6.

2. Definition and some properties of EWGD

The Weibull CDF is given by $1-\exp[-(t/\gamma)^c]$ and the exponentiated Weibull CDF is given as $F_T(t)=\left[1-e^{-(t/\gamma)^c}\right]^a$, for $t > 0$ and $\gamma, c > 0$. The CDF of geometric distribution with probability p of success is $F_R(y)=1-q^{y+1}$ for $y = 0, 1, 2, \ldots$ and $0 < q = 1 - p < 1$. By using equation (1.1), the CDF of the exponentiated Weibull-geometric distribution (EWGD) is given by $F_Y(y)=F_T(W[F_R(y)])$, where $W[F_R(y)]=-\ln[1-F_R(y)]=-\ln[q^{y+1}]=-(y+1)\ln(q)$. Hence,

$$F_Y(y)=F_T\{-\ln(q)[y+1]\}=\left\{1-\exp[-\left((-\ln q)/\gamma\right)^c\,(y+1)^c]\right\}^a=\left\{1-\theta^{(y+1)^c}\right\}^a,$$

where $c > 0$, $a > 0$, and $0<\theta=\exp[-\left((-\ln q)/\gamma\right)^c]<1$. Therefore, the CDF of EWGD is given by

$$F_Y(y)=\left\{1-\theta^{(y+1)^c}\right\}^a, \text{ for } y = 0, 1, 2, 3, \ldots \tag{2.1}$$

The corresponding probability mass function (PMF) for EWGD is given as

$$f_Y(y)=f(y)=F_Y(y)-F_Y(y-1)=\left[1-\theta^{(y+1)^c}\right]^a-\left[1-\theta^{y^c}\right]^a, \text{ for } y = 0, 1, 2, 3, \ldots \tag{2.2}$$

Observe that $f(0)=F_Y(0)=(1-\theta)^a$. The EWGD in (2.1) is the same as the exponentiated discrete Weibull distribution (Nekoukhou and Bidram, 2015). The two distributions are derived through different methods. In this paper, different properties and applications to count data modeling are emphasized.

When $c = 1$, EWGD reduces to the exponentiated exponential-geometric distribution (EEGD) defined and studied by Alzaatreh et al. (2012). When $c = a = 1$, the EWGD reduces to the geometric distribution with parameter θ. When $a = 1$, the EWGD reduces to the discrete Weibull distribution defined and studied by Nakagawa and Osaki (1975). When $a = 1$ and $c = 2$, the EWGD reduces to the discrete Rayleigh distribution defined by Roy (2004).

The sum of all probabilities in Equation (2.2) is 1. Thus, we have

$$\sum_{y=0}^{\infty} f(y)=\left[(1-\theta)^a-0\right]+\left[(1-\theta^{2^c})^a-(1-\theta)^a\right]+\left[(1-\theta^{3^c})^a-(1-\theta^{2^c})^a\right]$$
$$+\cdots+\left[(1-\theta^{(k+1)^c})^a-(1-\theta^{k^c})^a\right]+\cdots=1.$$

From the above, the second term in the second square bracket cancels out with the first term of the first square bracket. This action will continue and the last term that remains in the last square bracket will be the term $(1-\theta^{\infty})^{a}=1$, because $0 < \theta < 1$. Hence, the probabilities sum to 1.

Transformations:

The following propositions show the relationships between EWGD and some continuous distributions. These relationships can be used to simulate random variates from the EWGD.

Proposition 1: If U is a uniform (0, 1) random variable, then $Y=\left[\left\{\log_{\theta}(1-u^{1/a})\right\}^{1/c}\right]$, where $[v]$ is the largest integer less than or equal to v, follows an EWGD with parameters a, c, and θ.

Proof: $P(Y=y)=P\left([\{\log_{\theta}(1-U^{1/a})\}^{1/c}]=y\right)=P\left(y\le\{\log_{\theta}(1-U^{1/a})\}^{1/c}<y+1\right)$

$= P\left((1-\theta^{y^c})^{a}\le U<(1-\theta^{(y+1)^c})^{a}\right)=\left(1-\theta^{(y+1)^c}\right)^{a}-\left(1-\theta^{y^c}\right)^{a}$, on simplification.

Hence, Y follows the EWGD in Equation (2.2). □

By using the technique in the proof of Proposition 1, the following propositions can be proved.

Proposition 2: If V follows a standard exponential distribution, then the random variable $Y=\left[\left\{\log_{\theta}\left(1-(1-e^{-v})^{1/a}\right)\right\}^{1/c}\right]$ follows an EWGD with parameters a, c, and θ.

Proposition 3: Let V be an exponentiated exponential random variable with CDF $F(v)=(1-e^{-v})^{d}$, then $Y=\left[\left\{\log_{\theta}\left(1-(1-e^{-v})^{d/a}\right)\right\}^{1/c}\right]$ follows an EWGD with parameters $g(=d/a)$, c, and θ.

Proposition 4: Let V be a standard Pareto random variable with CDF $F(v)=1-v^{-1}$, then $Y=\left[\left\{\log_{\theta}\left(1-(1-v^{-1})^{1/a}\right)\right\}^{1/c}\right]$ follows an EWGD with parameters a, c, and θ.

Proposition 5: Let V be a Gumbel random variable with CDF $F(v)=\exp(-e^{-v})$, then $Y=\left[\left\{\log_{\theta}\left(1-\exp(-a^{-1}e^{-v})\right)\right\}^{1/c}\right]$ follows an EWGD with parameters a, c, and θ.

Proposition 6: Let V be a Fréchet random variable with CDF $F(v)=\exp(-e^{-1/v})$, then $Y=\left[\left\{\log_{\theta}\left(1-\exp(-a^{-1}e^{-1/v})\right)\right\}^{1/c}\right]$ follows an EWGD with parameters a, c, and θ.

Quantile Function:

By using Proposition 1, the quantile function of EWGD is $y=Q_{Y}(u)=\left[\left\{\log_{\theta}(1-u^{1/a})\right\}^{1/c}\right]$, where $[v]$ is the largest integer less than or equal to v. This result can be used to simulate a random sample from EWGD. In order to do this, simulate random variate u from the uniform (0, 1) and compute $Q_{Y}(u)$ to obtain a random variate y from the EWGD.

The exponentiated Weibull distribution with the PDF

$$g(t) = \frac{ac}{\gamma}\left(\frac{t}{\gamma}\right)^{c-1} e^{-(t/\gamma)^c}\left[1-e^{-(t/\gamma)^c}\right]^{a-1}, \qquad (2.3)$$

is monotonically decreasing for all values of γ, $c < 1$ and $a < 1$. Hence, the exponentiated Weibull-geometric distribution is monotonically decreasing for all values of $c < 1$ and $a < 1$. This result is based on Lemma 2 of Alzaatreh et al. (2012) for any *T*-geometric distribution. Note that there are other values of the parameters *c* and *a* for which the EWGD is monotonically decreasing even though the distribution of *T* (i.e., exponentiated Weibull distribution) is not monotonically decreasing.

The hazard function of EWGD is given by $h(y) = \frac{f_Y(y)}{1-F_Y(y)} = \frac{(1-\theta^{(y+1)^c})^a - (1-\theta^{y^c})^a}{1-(1-\theta^{(y+1)^c})^a}$.

Nekoukhou and Bidram (2015) illustrated the hazard rate function of EWGD for different values of the parameters *a*, *c* and θ. They noted that the hazard rate function could be decreasing, increasing, bathtub-shaped, and upside-down bathtub. This shows that the EWGD, characterized by two shape parameters, is more flexible than many other discrete distributions.

By using Theorem 2 in Alzaatreh et al. (2012), if the distribution of *T* is unimodal, so also is the distribution of the *T*-geometric distribution. We only need to show that the distribution of the exponentiated Weibull distribution is unimodal.

Proposition 7: The distribution of the exponentiated Weibull distribution is unimodal.
Proof: On differentiating the PDF *g*(*t*) in Equation (2.3) and setting it to zero, we obtain

$(c-1)(1-e^{-u}) - cu = acue^{-u}$, where $u = (t/\gamma)^c$.

If $a < 1$ and $c < 1$, it is obvious that $g'(t) < 0$. Hence, the exponentiated Weibull is monotonically decreasing. We only need to consider other cases. The left hand side, $(c-1)(1-e^{-u}) - cu$, is an increasing function of *u*, since the derivative of the left hand side expression is always positive. Its minimum is when $u \to 0$. The right hand side, $acue^{-u}$, is a concave down function since its derivative is first positive and then becomes negative. Thus, the function first increases to a point of maximum and then decreases. The minimum of the right hand side expression is at 0 when $u \to 0$ and $u \to \infty$. The maximum is at $u = 1$ when the function reaches the maximum of ace^{-1}. Both the left hand side and the right hand side expressions start from zero and since one is concave down and the other is strictly increasing, they can only intersect at only a single point greater than 0. This point of intersection is the mode of the exponentiated Weibull distribution. Hence, the exponentiated Weibull distribution is unimodal, and this ends the proof. □

By using Theorem 2 in Alzaatreh et al. (2012) and Proposition 7, the EWGD is unimodal. Thus, the EWGD is either monotonically decreasing or concave down.

Moments and dispersion:

The moments and the moment generating function cannot be expressed in closed forms. However, the r^{th} central moments can be computed numerically by evaluating $\mu_r = E(X-\mu)^r$, where $\mu = \Sigma_{x=0}^{\infty} xP(X=x)$. The summation is evaluated when the probability $1-P(X \le x)$ is at most 1.0E–10. The mean, variance, skewness and kurtosis are computed for some parameter values. We consider the values $a = 0.2(0.1)10.0$, $c = 0.5(0.1)10.0$, and $\theta = 0.1(0.1)0.9$. From this computation, we observe the following patterns between the mean, variance, skewness, kurtosis and the parameters: When both *a*

and c are fixed, the mean and variance are increasing functions of θ and there is no observed pattern for skewness and kurtosis. For fixed a and θ, the mean, variance, skewness and kurtosis are decreasing functions of c. When c and θ are fixed, the mean is an increasing function of a and there is no observed pattern for the variance, skewness and kurtosis. A small portion of these values are presented in Table 1. A more detailed table is presented in Famoye (2018).

When $a \leq 2$ and $c \leq 1$, the EWGD is over-dispersed. For all other values of a and c, the distribution is either under-dispersed, equi-dispersed or over-dispersed.

Table 1: Moments of EWGD for some parameter values

		$\theta = 0.2$				$\theta = 0.6$			
c	a	μ	σ	sk	ku	μ	σ	sk	ku
0.5	0.5	0.27	1.43	10.09	191.23	3.88	160.29	8.95	156.22
	1.0	0.52	2.70	7.29	102.40	7.26	292.35	6.65	88.37
	1.5	0.75	3.84	6.08	72.94	10.25	404.67	5.68	65.62
	2.0	0.97	4.87	5.38	58.32	12.93	502.59	5.12	54.15
	2.5	1.18	5.80	4.92	49.62	15.37	589.62	4.76	47.17
	3.0	1.38	6.64	4.59	43.87	17.60	668.15	4.49	42.45
1.0	0.5	0.13	0.18	3.97	23.31	0.85	2.47	2.90	14.85
	1.0	0.25	0.31	2.68	12.20	1.50	3.75	2.07	9.27
	1.5	0.36	0.42	2.10	8.65	2.02	4.45	1.74	7.65
	2.0	0.46	0.49	1.75	6.99	2.44	4.86	1.58	6.96
	2.5	0.55	0.55	1.51	6.08	2.79	5.13	1.48	6.59
	3.0	0.63	0.60	1.34	5.56	3.09	5.31	1.43	6.36
1.5	0.5	0.11	0.11	2.94	11.38	0.54	0.71	1.78	6.50
	1.0	0.21	0.19	1.83	5.50	0.93	0.95	1.10	4.28
	1.5	0.30	0.24	1.29	3.68	1.21	1.02	0.84	3.87
	2.0	0.38	0.28	0.95	2.89	1.43	1.02	0.74	3.81
	2.5	0.45	0.30	0.70	2.52	1.61	1.00	0.70	3.82
	3.0	0.52	0.32	0.50	2.38	1.75	0.98	0.69	3.82
1.75	0.5	0.11	0.10	2.74	9.34	0.48	0.51	1.49	5.01
	1.0	0.20	0.17	1.65	4.31	0.81	0.65	0.82	3.42
	1.5	0.29	0.22	1.12	2.76	1.05	0.67	0.57	3.29
	2.0	0.37	0.25	0.76	2.10	1.23	0.65	0.49	3.38
	2.5	0.44	0.27	0.49	1.80	1.37	0.62	0.48	3.45
	3.0	0.50	0.28	0.27	1.70	1.48	0.60	0.50	3.46

3. Statistical Inference

We consider parameter estimation, test of hypothesis and goodness-of-fit tests. In sub-section 3.1, we address the maximum likelihood estimation of the three parameters of EWGD. In sub-section 3.2, we compare the EWGD with its sub-models and briefly describe some goodness-of-fit statistics.

3.1 Maximum likelihood estimation

Suppose a random sample $Y_1, Y_2, \ldots, Y_n$ of size n is taken from the EWGD. The log-likelihood function of the EWGD in Equation (2.2) is given by

$$\ell = \log L(a, c, \theta) = \sum_{i=1}^{n} \log \left\{ [1 - \theta^{(y_i+1)^c}]^a - [1 - \theta^{y_i^c}]^a \right\}. \quad (3.1)$$

The partial derivatives of the log-likelihood function with respect to a, c, and θ are, respectively, given by

$$\frac{\partial \ell}{\partial a} = \sum_{i=1}^{n} \frac{\{[1-\theta^{(y_i+1)^c}]^a \log(1-\theta^{(y_i+1)^c}) - [1-\theta^{y_i^c}]^a \log(1-\theta^{y_i^c})\}}{[1-\theta^{(y_i+1)^c}]^a - [1-\theta^{y_i^c}]^a}, \tag{3.2}$$

$$\frac{\partial \ell}{\partial c} = \sum_{i=1}^{n} \frac{\left\{ y_i^c \theta^{y_i^c} \log y_i \, [1-\theta^{y_i^c}]^{a-1} - (y_i+1)^c \theta^{(y_i+1)^c} \log(y_i+1)[1-\theta^{(y_i+1)^c}]^{a-1} \right\} a \log\theta}{[1-\theta^{(y_i+1)^c}]^a - [1-\theta^{y_i^c}]^a}, \tag{3.3}$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{\{y_i^c \theta^{y_i^c-1}[1-\theta^{y_i^c}]^{a-1} - (y_i+1)^c \theta^{(y_i+1)^c-1}[1-\theta^{(y_i+1)^c}]^{a-1}\} a}{[1-\theta^{(y_i+1)^c}]^a - [1-\theta^{y_i^c}]^a}. \tag{3.4}$$

The maximum likelihood estimates $\hat{c}$, $\hat{a}$, and $\hat{\theta}$ of the parameters can be obtained by setting Equations (3.2)-(3.4) to zero and solving the equations iteratively through an optimization routine. In this paper, we used PROC NLMIXED in SAS to maximize the log-likelihood function in Equation (3.1).

When $a = c = 1$, the EWGD reduces to the geometric distribution. We consider the data to be from geometric distribution and use the moment estimate of the geometric distribution to obtain the initial estimate of θ. Thus, the initial estimate of θ is given by equating the sample mean from the data to the geometric population mean. This is given as $\mu = \theta/(1-\theta) = \overline{y}$. On solving for θ, we obtain $\theta_0 = \overline{y}/(1+\overline{y})$. Hence, one set of initial estimates will be $a = c = 1$ and θ_0. We can use the zero frequency from the EWGD to find the initial estimate for parameter a. We solve equation $f_o = (1-\theta)^a$ for a to obtain $a_0 = \ln f_0 / \ln(1-\theta_0)$, where f_0 is the zero frequency from the sample. In order to find the initial estimate for the parameter c, we equate the first frequency from the sample to the population probability of $Y = 1$. This leads to solving the equation $f_1 = (1-\theta^{2^c})^a - f_0$. On solving the equation, the initial estimate of c is given by $c_0 = \ln\left\{\ln[1-(f_0+f_1)^{1/a_0}]/(\ln\theta_0)\right\}/\ln(2)$. These second initial estimates are based on the assumption that both f_0 and f_1 are non-zero.

3.2 Tests and goodness-of-fit statistics

The EWGD reduces to EEGD when $c = 1$. To compare the EWGD with EEGD, we test the hypothesis $H_0 : c = 1$ against $H_1 : c \neq 1$. The null hypothesis can be tested by using the t-statistic $t = (\hat{c}-1)/se(\hat{c})$, where $se(\hat{c})$ is the standard error of $\hat{c}$. By using the asymptotic normality of the maximum likelihood estimate (MLE), the statistic has an approximate normal distribution. Alternatively, one can use a likelihood ratio statistic with 1 degree of freedom.

The EWGD reduces to the geometric distribution when $a = c = 1$. Thus, we test $H_0 : c = a = 1$ against $H_1 : H_0$ is false. We use the likelihood ratio test. We define $L_0(\tilde{\theta})$, the likelihood statistic when $a = c = 1$ and $L_1(\hat{c},\hat{a},\hat{\theta})$ is the likelihood statistic when H_0 is false. The test statistic is defined as $\lambda = L_0(\hat{\theta})/L_1(\hat{c},\hat{a},\hat{\theta})$ and $-2\log(\lambda)$ is approximately chi-squared with 2 degrees of freedom.

The goodness-of-fit statistic can be based on the log-likelihood statistic, Akaike Information criterion (AIC), the Bayesian Information criterion (BIC), the Pearson chi-square statistic with its p-value and the ranked probability score (RPS). The AIC and the BIC are respectively defined as $\text{AIC} = -2\log(L) + 2p$ and $\text{BIC} = -2\log(L) + p\log(n)$, where n is the sample size, p is the number of estimated parameters and L is the likelihood

statistic. The smaller the AIC (or BIC), the better the model. The chi-square statistic is given by $\chi^2_{k-p-1} = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$, where O_i is the observed frequency in cell *i*, E_i is the expected frequency in cell *i* and *k* is the total number of cells. The degree of freedom for the chi-square distribution is $k - p - 1$.

The ranked probability score (RPS) is not often used. According to Weigel et al. (2006), "The RPS (Epstein 1969; Murphy 1969, 1971) is a squared measure that compares the cumulative [distribution] function (CDF) of a probabilistic forecast with the CDF of the corresponding observation over a given number of discrete probability categories." Thus, RPS measures the discrepancy between the theoretical CDF and empirical CDF given by

$$\text{RPS} = \sum_{m=1}^{k} \left(\sum_{i=1}^{m} e_i - \sum_{i=1}^{m} o_i \right)^2, \tag{3.5}$$

where e_i is the predicted (or forecasted, or theoretical) probability and o_i is the observed (empirical) proportion in category *i*. The smaller the measure, the better the model.

4. Count data regression

Suppose that *Y* is a count response variable that follows the EWGD in Equation (2.2) and *Y* is associated with a set of predictors. We wish to fit the response variable *Y* by using the predictors. Suppose we have a $k - 1$ row vector of predictors $x_i = (x_{i0} = 1, x_{i1}, x_{i2}, \ldots, x_{i,(k-1)})$. In count data modeling, it is common to model the mean by a log-linear relationship. The mean of EWGD is not in closed form, but it is a function of parameter θ. We assume that the parameter θ of EWGD is a function of x_i given by $\theta(x_i) = \theta_i = f(x_i, \beta)$, where $0 < f(x_i, \beta) < 1$ is a known function of x_i and a *k*-dimensional column vector $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1})$ of regression parameters. Since $0 < \theta < 1$, we take $f(x_i, \beta)$ to be the logit function

$$\theta(x_i) = \theta_i = f(x_i, \beta) = e^{x_i\beta} / (1 + e^{x_i\beta}). \tag{4.1}$$

This leads to the exponentiated Weibull-geometric regression (EWGR) model given by

$$P(Y = y_i \mid x_i) = \omega(y_i) = \left[1 - \theta_i^{(y_i+1)^c}\right]^a - \left[1 - \theta_i^{y_i^c}\right]^a, \; y_i = 0, 1, 2, \ldots, \tag{4.2}$$

where $\theta_i = \theta(x_i)$ is given in Equation (4.1). The estimation of the parameters can be carried out by using the maximum likelihood estimation method. The log-likelihood function is given by

$$\ell_* = \log L(a, c, \beta \mid x_i) = \sum_{i=1}^{n} \log \left\{ [1 - \theta_i^{(y_i+1)^c}]^a - [1 - \theta_i^{y_i^c}]^a \right\}.$$

The derivatives with respect to *a* and *c* are the same as in equations (3.2) and (3.3) respectively for the EWGD. The derivative with respect to the *k* parameters $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1})$ are as follows:

$$\frac{\partial \ell_*}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\{y_i^c \theta_i^{y_i^c - 1}[1 - \theta_i^{y_i^c}]^{a-1} - (y_i + 1)^c \theta_i^{(y_i+1)^c - 1}[1 - \theta_i^{(y_i+1)^c}]^{a-1}\} a}{[1 - \theta_i^{(y_i+1)^c}]^a - [1 - \theta_i^{y_i^c}]^a} \frac{\partial \theta_i}{\partial \beta_j}, \; j = 0, 1, 2, \ldots, k-1,$$

where $\partial \theta_i / \partial \beta_j = \theta_i (1 - \theta_i) x_{ij}$.

A count data may have an inflated number of *k* value in the data. The most common *k* value is the zero which leads to zero-inflated regression model. Similarly, the count data may not have a zero count and this leads to zero-truncated regression model. In this section, we will define a zero-inflated regression model for the EWGR model. A zero-inflated EWGR (ZIEWGR) model is a mixture model with the probability mass function

$$P(Y = y_i \mid x_i, z_i) = \begin{cases} \varphi_i + (1-\varphi_i)\omega(y_i), & y_i = 0 \\ (1-\varphi_i)\omega(y_i), & y_i = 1, 2, 3, \ldots, \end{cases} \tag{4.3}$$

where $\omega(y_i)$ is the EWGR model given in Equation (4.2) and $0 < \varphi_i < 1$. The probability φ_i may be taken as a nuisance parameter when the data set is small or a function of predictors when the sample size is large. If φ_i is a function of predictors $z_i = (z_{i0} = 1, z_{i1}, z_{i2}, \ldots, z_{i,(r-1)})$, then φ_i can be defined as $\varphi_i = \exp(z_i\delta)/[1+\exp(z_i\delta)]$, where δ is an *r*-dimensional column vector $\delta = (\delta_0, \delta_1, \delta_2, \ldots, \delta_{r-1})$ of parameters. In general, z_i may be a subset of x_i or different from x_i.

5. Applications

In this section, we apply the generalized Poisson regression (GPR) model defined by Famoye (1993), the exponentiated exponential geometric regression (EEGR) model defined by Famoye and Lee (2017) and the EWGR model to two count data sets. These two models are chosen because both can be over- or under-dispersed. Because the data sets have high proportion of zero, the zero-inflated versions of the models were also applied and the results are compared.

5.1 Health Care Data:

Cameron et al. (1988) used the data from 1977-78 Australian Health Survey to analyze various measures of health-care utilization. The data can be obtained from the Journal of Applied Econometrics 1997 Data Archive. Many authors, including Mullahy (1997) and Cameron and Johansson (1997), fitted the data to univariate regression models. A detailed description of the predictor variables can be found in Gurmu and Elder (2000). A summary statistics for the predictor variables were provided in Cameron et al. (1988).

We model the response variable *y*, the total number of non-prescribed medications used in the past two days. The complete data set has six response variables. All the six variables were adequately fitted by the EWGR and ZIEWGR models. The SAS NLMIXED procedure was used to fit the regression models to the response variables. There is an adequate fit when the optimization program converged and the gradient for each of the parameter estimates is less than 1.0E-6. When we considered GPR and EEGR and their inflated models, these two models adequately fitted the response variable *y* and one other response variable (the number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months). The results from this other response variable is similar to the variable *y* reported in Table 2. The response variable *y* ranges from 0 to 8 with a mean of 0.3557 and a standard deviation of 0.507. The variable is over-dispersed and it is highly skewed to the right with skewness of 3.05 and kurtosis of 15.11.

The results of fitting ZIEEGR and ZIEWGR are presented in Table 2. For all models (that of ZIGPR is not provided in the table), the predictors sex, age and illness are positively associated with total number of non-prescribed medications used. However, the predictor freerepa is negatively associated with the response variable. The dispersion parameter *a* in both ZIGPR and ZIEEGR are significantly different from 1. In the ZIEWGR model, the dispersion parameter *c* is significantly different from 1 but the parameter *a* is not

significantly different from 1. The ZIEEGR is nested within the ZIEWGR model. Thus, we can compare ZIEWGR with ZIEEGR by testing if the parameter $c = 1$ under a null hypothesis. Since the null hypothesis is rejected, one should use the ZIEWGR to model the data. The log-likelihood statistics for ZIEEGR and ZIEWGR models in Table 2 support the assertion. The log-likelihood statistic and the RPS for ZIGPR model are respectively -3904.55 and 5.7E-5. By using the RPS and the log-likelihood statistics, we notice that ZIEWGR provided the best fit among all the three models.

The log-likelihood statistics for the GPR, EEGR and EWGR models are respectively -3930.16, -3929.87, and -3918.31. In comparing these values with the corresponding ones for the zero-inflated models, we observe that the zero-inflated models performed better than the non-inflated (ordinary) models. The ordinary models are all nested within the zero-inflated models. The likelihood ratio statistics for testing if all the parameters of the zero-inflation part are all zeros are rejected at 5% level for all models.

Table 2. Parameter estimates (standard errors in parentheses) for health-care data.

Variable	ZIEEGR		ZIEWGR	
x/z	β	δ	β	δ
Constant	-2.3555 (0.226)*	0.0916 (1.495)	-4.4433 (0.740)*	-0.5237 (1.872)
Sex	0.1935 (0.066)*	-0.7508 (0.480)	0.1926 (0.060)*	-0.9426 (0.515)
Age	4.9383 (1.211)*	0.5814 (9.637)	5.2423 (1.126)*	9.2446 (11.42)
Agesq	-6.1885 (1.351)*	-1.9508 (11.07)	-6.6473 (1.267)*	-13.6344 (13.99)
Income	0.1181 (0.100)	0.0878 (0.778)	0.0208 (0.090)	-0.7611 (0.706)
Levyplus	-0.0556 (0.076)	-0.1172 (0.481)	-0.0796 (0.069)	-0.4022 (0.484)
Freepoor	-0.0390 (0.161)	-0.1319 (1.198)	-0.1433 (0.148)	-1.4247 (1.657)
Freerepa	-0.2756 (0.116)*	0.6015 (0.861)	-0.2872 (0.107)*	0.7172 (1.063)
Illness	0.1478 (0.027)*	-3.3287 (1.351)*	0.1622 (0.022)*	-17.0050 (31.47)
Actdays	0.0075 (0.010)	0.2838 (0.156)	-0.0018 (0.009)	1.3707 (2.246)
Hscore	0.0203 (0.013)	-1.1435 (1.313)	0.0212 (0.012)	-29.9175 (260.2)
Chcond1	0.1063 (0.070)	-0.3559 (0.548)	0.0934 (0.066)	-0.5457 (0.591)
Chcond2	-0.0556 (0.100)	-1.1609 (1.502)	-0.0447 (0.095)	-1.9216 (2.342)
$\hat{a}$	1.3611 (0.090)*		9.7881 (6.921)	
$\hat{c}$			0.5708 (0.090)*	
LogL	-3905.91		-3893.49	
AIC	7865.8		7843.0	
BIC	8042.8		8026.5	
RPS	2.95E-5		1.17E-6	

On fitting the ZIGPR to the data, we obtain the log-likelihood as -3904.55 and the RPS as 5.70E-6. The observed proportion of zeros in the data is 73.49%. After fitting the ZIGPR, ZIEEGR and ZIEWGR models, the predicted proportion of zeros are respectively given by 73.83%, 73.56% and 73.48%. The ZIEWGR provided the best predicted probability of zero. We also calculated the chi-square values by combining the last three classes in the frequency table. The chi-square values for ZIGPR, ZIEEGR and ZIEWGR are respectively given by 16.41, 16.68 and 2.20. Note that we have a total of 7 classes after the last three classes were combined. The goal for computing the chi-square values is not to check if these values are significant, but to see which of these models provides the closest expected frequencies. In this analysis, the ZIEWGR model provided the best fit by using the goodness of fit statistics.

5.2 Violence Data:

The National Violence Against Women (NVAW) Survey of 1995-1996 was conducted to obtain a public-use data set. Interviews were completed from men and women, but the data used in this sub-section is a subset of the 8000 interviews completed by women who were at least 18 years old living in US households. Respondents were asked questions on various topics including physical assault they had experienced as adults by any type of perpetrator. The response variable used in the data analysis is physical assault or violence. This is the total number of twelve possible violent physical actions directed toward a woman by her current and/or past partners. A high score on this variable indicates a woman experienced severe violence.

In the analysis, seven predictor variables were used. The variables are age in years; level of education is one of the seven school levels (0 = no schooling to 6 = postgraduate); race (1 = white, 0 = others); number of children under 18 years of age (Nchid); respondent's income level is one of 10 levels (1 = below $5,000 to 10 = over $1,000,000); health level is one of 5 levels (0 = poor to 4 = excellent); and drug is a binary variable that indicates illicit drug use with 1 = yes and 0 = no. The variable drug indicates if a woman has used marijuana, cocaine, heroin, angel dust, etc. in the past month. After excluding the cases having missing information on any of the predictor variables and the response variable, we have 6110 observations.

The descriptive statistics for the response and predictor variables are given in Table 3. The response variable, violence, is positively skewed (skewness = 2.24, kurtosis = 4.70). Tjaden and Thoennes (1999) provided detailed description of the variables and the most recent publications on the data.

Table 3. Descriptive statistics for the response and predictor variables (n = 6110)

Variable	Description	Mean ± SD	Proportion of 1's
Age	Age in years	42.54 ± 15.37	
Educ	Education level	3.79 ± 1.16	
Race	Race		0.8146
Nchild	Number under 18 years	0.97 ± 1.21	
Income	1995 family income level	3.95 ± 2.44	
Health	Health condition	2.74 ± 1.08	
Drug	Illicit drug use		0.0172
violence	Response variable	1.23 ± 2.34	

SD = standard deviation

The results of fitting the ZIEEGR and ZIEWGR models are presented in Table 4. For the ZIGPR (not included in Table 4) and ZIEEGR models, the variables education and health are significantly associated with the response variable violence. The higher the level of education (or the better the health condition), the lower the number of violence a respondent experienced. The other five predictor variables are not significantly related to violence. In the ZIEWGR model, the predictor variables education and health are negatively associated with the number of violence. In addition to these two predictor variables, drug is positively related to the number of violence under the ZIEWGR model. The respondents who used illicit drug in the past month of the survey tend to have higher number of violence.

Table 4. Parameter estimates (standard errors in parentheses) for violence data.

Variable	ZIEEGR		ZIEWGR	
x/z	β	δ	β	δ
Constant	1.4382 (0.147)*	-0.8980 (0.191)*	8.3098 (0.985)*	-8.6298 (1.294)*

Age	-0.0029 (0.002)	0.0200 (0.002)*	0.0009 (0.006)	0.1027 (0.016)*
Educ	-0.1050(0.023)*	0.0075 (0.031)	-0.1765 (0.054)*	-0.0566 (0.108)
Race	0.0566 (0.054)	0.1872 (0.081)*	0.0614 (0.127)	1.1172 (0.416)*
Nchild	0.0291 (0.020)	-0.0467 (0.028)	0.0819 (0.048)	0.1955 (0.133)
Income	-0.0050 (0.011)	-0.0404 (0.014)*	0.0154 (0.025)	0.0217 (0.063)
Health	-0.0753 (0.021)*	0.2123 (0.031)*	-0.2277 (0.053)*	0.4199 (0.094)*
Drug	0.1609 (0.125)	-1.0985 (0.268)*	0.7492 (0.360)*	-11.4066 (163.1)
$\hat{a}$	1.6930 (0.135)*		0.0714 (0.009)*	
$\hat{c}$			3.3861 (0.338)*	
LogL	-8092.68		-8054.58	
AIC	16219.0		16145.0	
BIC	16334.0		16266.0	
RPS	5.446E-4		1.461E-4	

*Significant at 5% level.

The chi-square values from the ZIGPR, ZIEEGR and ZIEWGR models are respectively given by 58.73, 46.84 and 31.23. The ZIEWGR model provided the closest expected frequencies. The observed proportion of zero for the response variable violence is 67.05%. The predicted proportion of zero from ZIGPR, ZIEEGR and ZIEWGR models are respectively 67.07%, 67.08% and 66.93%. The ZIGPR provided the best expected zero frequency.

The log-likelihood statistic and the RPS for ZIGPR model are respectively -8094.93 and 6.221E-4. In comparing these values with the corresponding values for ZIEEGR and ZIEWGR models in Table 4, we observe that the ZIEWGR model provided the best fit followed by the ZIEEGR model. The log-likelihood statistics for the GPR, EEGR and EWGR models are respectively -8416.13, -8224.73 and -8151.13. In comparing the ordinary regression models with their corresponding zero-inflated regression models, we observe that the zero-inflated models performed better. The results from the data analysis show that the ZIEWGR provided the best fit by using the goodness of fit statistics.

6. Summary and conclusions

The exponentiated Weibull-geometric distribution can be applied to fit count data with over-dispersion, equi-dispersion or under-dispersion. The distribution has closed form probability mass function and a cumulative distribution function. One limitation of the distribution is that its moments cannot be expressed in closed forms. However, the moments can easily be computed numerically.

A count data regression, the exponentiated Weibull-geometric regression model, is defined. A modified version, the ZIEWGR model is defined and illustrated with two numerical data sets. The goodness-of-fit of ZIEWGR model is compared with ZIEEGR and ZIGPR by using the AIC and the ranked probability scores among other statistics. In the two numerical examples, the ZIEWGR performed better than the other two count data regression models.

Acknowledgement:
The author acknowledges the financial support received from the U.S. Department of State, Bureau of Education and Cultural Affairs under the Fulbright Grant # PS00230565.

References

Alzaatreh, A., Lee, C. and Famoye, F. (2013) A new method for generating families of continuous distributions. *Metron*, 71(1), 63-79.

Alzaatreh, A, Lee, C. and Famoye, F. (2012) On the discrete analogues of continuous distributions. *Statistical Methodology*, 9, 589-603.

Balakrishnan, N. and Nevzorov, V.B. (2003). *A Primer on Statistical Distributions*, John Wiley & Sons, Inc., Hoboken, NJ.

Cameron, A.C. and Johansson, P. (1997) Count data regression using series expansion: With applications. *Journal of Applied Econometrics*, 12, 203-223.

Cameron, A.C., Trivedi, P.K., Milne, F. and Piggott, J. (1988) A microeconomic model of the demand for health care and health insurance in Australia. *Review of Econometric Studies*, LV, 85-106.

Consul, P.C. and Famoye, F. (2006). *Lagrangian Probability Distributions*, Birkhäuser, Boston, Massachusetts.

Epstein, E.S. (1969) A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 985-987.

Famoye, F. (2018) Exponentiated Weibull-geometric distribution and its application to count data. Accepted by Journal of Data Science.

Famoye, F. (1993) Restricted generalized Poisson regression model. *Communications in Statistics - Theory & Methods*, 22(5), 1335-1354.

Famoye, F. and Lee, C. (2017) Exponentiated exponential-geometric regression model. *Journal of Applied Statistics*, 44(16), 2963-2977.

Frome, E.L., Kurtner, M.H. and Beauchamp, J.J. (1973) Regression analysis of Poisson-distributed data. *Journal of the American Statistical Association*, 68, 288-298.

Gurmu, S. and Elder, J. (2000) Generalized bivariate count data regression models. *Economics Letters*, 68, 31-36.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions, 3rd edition*, John Wiley & Sons, Inc., Hoboken, NJ.

Lee, C., Famoye, F. & Alzaatreh, A. (2013) Methods for generating families of univariate continuous distributions in recent decades, *WIREs Computational Statistics*, 5, 219-238.

Mahmoudi, E. and Shiran, M. (2012) Exponentiated Weibull-geometric distribution and its applications. https://www.researchgate.net/publication/227173418_Exponentiated_Weibull-Geometric_Distribution_and_its_Applications.

Mullahy, J. (1997) Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12, 337-350.

Murphy, A.H. (1969) On the ranked probability skill score. *Journal of Applied Meteorology*, 8, 988-989.

Murphy, A.H. (1971) A note on the ranked probability skill score. *Journal of Applied Meteorology*, 10, 155-156.

Nakagawa, T. & Osaki, S. (1975). The discrete Weibull distribution, *IEEE Transactions on Reliability*, 24(5), 300-301.

Nekoukhou, V & Bidram, H. (2015) The exponentiated discrete Weibull distribution, *Statistics and Operations Research Transactions*, 39(1), 127-146.

Roy, D. (2004). Discrete Rayleigh distribution, *IEEE Transactions on Reliability*, 53(2), 255-260.

Tjaden, P. and Thoennes, N. (1999) Violence and threats of violence against women and men in the United States, 1994-1996 [Computer file]. ICPSR version. Denver, CO: Center for Policy Research [producer], 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1999. https://doi.org/10.3886/ICPSR02566.v1

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2006) The discrete Brier and ranked probability skill scores. *Monthly Weather Review,* 135, 118-124.