International Journal of Data Science and Analysis
2020; 6(5): 153-162
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200605.15
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Discrete Weibull and Artificial Neural Network Models in Modelling Over-dispersed Count Data

Kipkorir Collins*, Anthony Waititu, Anthony Wanjoya

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
ckipkorir@gmail.com (K. Collins), agwaititu@gmail.com (A. Waititu), awanjoya@gmail.com (A. Wanjoya)

*Corresponding author

To cite this article:
Kipkorir Collins, Anthony Waititu, Anthony Wanjoya. Discrete Weibull and Artificial Neural Network Models in Modelling Over-dispersed Count Data. *International Journal of Data Science and Analysis*. Vol. 6, No. 5, 2020, pp. 153-162. doi: 10.11648/j.ijdsa.20200605.15

Received: October 2, 2020; **Accepted**: October 20, 2020; **Published**: October 26, 2020

Abstract: In modelling count data, the use of least square regression models suffers several methodological limitations and statistical properties in instances of discrete, non-negative integer count of a dependent variable. Unlike the classical regression model, count data models are non-linear with many properties of the response variable relating to discreteness, non-linearity and deal with non-negative values only. A good starting point for modelling count data is the Poisson regression model since it lends itself well with the nature properties of count data. However, the limitation of equi-dispersion renders it inappropriate for modelling over-dispersed data. Negative Binomial regression model has been widely used and considered as the default regression model for over-dispersed count data. This model is a modification of Poisson regression model and though widely used, it might not be the best model for over-dispersion and other models have been found to perform better. Over-dispersion in this study was defined relative to the Poisson model. This study models over-dispersed count data using discrete Weibull regression model and artificial neural network model with a median neuron in the hidden layer. After fitting the two models on simulated data and real data, the artificial neural network model outperformed the discrete Weibull regression model. Application on data set from German health survey gave RMSE of DW regression model as 69.0668 and 35.5652 for the artificial neural network.

Keywords: Over-dispersion, Count, Discrete Weibull, Artificial Neural Network

1. Introduction

Count data is defined as the number of times an event occurs within a period of time that is fixed. In modelling, the use of least square regression models suffers several methodological limitations and statistical properties in instances of discrete, non-negative integer count of a dependent variable [1]. Unlike the classical regression model, count data models are non-linear with many properties of the response variable relating to discreteness, non-linearity and deal with non-negative values only. A good starting point for modelling count data is the Poisson regression model since it lends itself well with the nature properties of count data. Some examples of such data are the number of road accident deaths, the number of patents awarded to a firm, the number of dengue fever cases which is restricted to a single digit or integer with low number of events and the number of times a doctor visits a patient [2].

Reference [3] indicated that for modelling count data, Poisson regression model has more merits over the conventional linear models. However, Poisson regression model still has one potential problem. This is the property of equi-dispersion, that is the assumption of equality of variance and mean. When this property is violated, for instance, the variance of the observed counts exceeds the mean, an over-dispersion will occur. When the variance of the observed counts is lower than the mean, then under-dispersion occurs. Under-dispersion are experienced on rare occasions and this often happens when the sample mean is low [4]. Failure to control for dispersion will lead to inconsistent estimates, inflated statistics and biased in standard error. Hence with count data modelling, after the development of Poisson regression model, one proceeds with the analysis of correcting for dispersion if it exists.

Negative Binomial (NB) regression has been widely considered as the default choice for data that are over-dispersed relative to the Poisson regression. This is because it has a closed form equation and the mathematical relationship between the mean and variance is easy to manipulate [5]. However, NB regression count cannot deal with data that are under-dispersed relative to the Poisson distribution. These cases of under-dispersion can arise in various applications where the data are preprocessed for confidentiality [6]. There have been attempts to extend the Poisson based models to include under-dispersion such as the generalized Poisson regression models [7], Conway-Maxwell-Poisson regression [8], extended Poisson process models [9] or hyper-Poisson regression models [10]. These models are all modifications of the Poisson model and have been proved to be quite complex and computationally intensive in practice [11]. This study looked at Discrete Weibull (DW) regression and Artificial neural network (ANN) models. Reference [12] presents the DW regression model. The motivation behind this comes from the vital role played by the continuous Weibull distribution in survival analysis of failure time studies. This study evaluates the performance of DW regression with comparison to ANN model

2. Literature Review

Poisson distribution is the classical and first choice in modelling count data [2]. Even though the Poisson model is widely considered as the basic model in analyzing count data, the reliance of this model on a single parameter limits its usage on real data. This arises due to the property of equivalence in variance and mean being violated.

Negative Binomial model is suitable for modelling overdispersion. This is because the NB model relaxes the assumption of equi-dispersion of Poisson regression by introducing a dispersion parameter that allows the variance to exceed the mean. The theoretical variance of the NB is always greater than its mean hence the reason for its widespread usage for modelling over-dispersed count data. The NB model has been applied by many researchers for example [1, 13-15] and many more. This model is the most used model in crash frequency modelling. However, the model has its disadvantages, mostly is the inability to model under-dispersed data and the problem of estimating the dispersion parameter when the data has a low sample mean and small sample sizes [16, 17]. This limitation makes it necessary to have models that can cope with the cases of under- dispersed data relative to the Poisson distribution

Reference [18] proposed gamma model for count data to model over- dispersed and under-dispersed count data. Reference [19] used this distribution to analyze crashes collected at RHX in South Korea. They found that the gamma count model provides a good statistical fit for the railway-highway crossing crash data under study. This model performs well statistically but it is a dual-state model. Although the model is able to provide a good fit, its assumptions limits its applicability. Where observations for a time t-1 will affect the observation at time t, the gamma model assumes that observations are not independent.

3. Methodology

3.1. Dispersion for Count Data

Dispersion for any data can be described as the variability or spread of the data. Reference [2, 20] indicates that dispersion in count data should be defined in relation to a specific distribution. In this context, the variance ratio (VR) can be defined as the ratio between the observed variance from the data and the theoretical variance from the model fit. This can be written as;

$$VR = \frac{Observed\ variance}{Theoretical\ variance} \quad (1)$$

Modelling count data exhibit different types of dispersion. Data is over-dispersed when the observed variance is greater than the expected variance specified by the fitted model. Under-dispersion describes a case where the observed variance is less than the theoretical variance. When the observed variance and theoretical variance are equal, the data does not show any dispersion and can be referred to as equi-dispersed.

Furthermore, dispersion of count data can be defined in relation to the Poisson model. Therefore, it is common to refer to these data as being dispersed relative to Poisson. In this case variance of the model is estimated by the sample mean. Thus, dispersion relative to the Poisson refers to cases where the sample variance (observed variance) is greater than sample mean (theoretical variance) for overdispersion, equal for equi-dispersion and smaller for under-dispersion. From this definition, dispersion of a dataset can therefore be identified with regard to a dispersion coefficient (Dip). This is defined as the ratio of the variance to the mean;

$$Dip = \frac{\sigma^2}{\mu} \quad (2)$$

From this, data was considered over-dispersed relative to the Poisson when $Dip > 1$, equi-dispersed when $Dip=1$ and under-dispersed when $Dip < 1$. Dispersion has frequently been defined in literature using the variance-to-mean ratio given above. Specifically, dispersion relative to the Poisson distribution is found when the variance is greater or less than the mean [2]. Failure to account for dispersion in modelling count data may lead to biased parameter estimation and hence lead to false conclusions and decisions. This study considered the *Dip* when checking for dispersion.

3.2. Discrete Weibull (DW) Distribution and Regression Model

3.2.1. DW Distribution

Roy [21] introduces the cumulative distribution of DW distribution. If Y follows a type 1 DW distribution, then the distribution of Y is given by;

$$F(y;q,\beta)=\begin{cases}1-q^{(y+1)^{\beta}}, for\ y=0,1,2,\ldots\\ 0, otherwise\end{cases} \quad (3)$$

and the probability mass function is given by;

$$F(y;q,\beta)=\begin{cases}q^{y^{\beta}}-q^{(y+1)^{\beta}}, for\ y=0,1,2,\ldots\\ 0, otherwise\end{cases} \quad (4)$$

where the parameters $\beta > 0$ and $0 < q < 1$. The parameter q gives the probability of obtaining a non-zero response since $f(0)=1-q$. The DW distribution is connected to other well-known distributions. These include; [22] discrete Releigh distribution with $\beta=2$ and $q=0$. The geometric distribution, is a special case of DW distribution, with $q=1-p$ and $\beta=1$. The variance the geometric distribution is always greater than its mean. Therefore, regardless of the value of q, DW distribution with $\beta=1$ is a case of overdispersion relative to Poisson regression. Actually, when $\beta=1$ and $q=e^{-\lambda}$, the distribution is the discrete exponential distribution introduced by [23]. The parameter β can be considered as controlling the range of values of the variable, that is, controls the skewness of the DW distribution.

Moments and Quantiles

For the DW distribution, the first two moments are given as;

$$E(Y)=\mu=\sum_{y=1}^{\infty}q^{y^{\beta}} \quad (5)$$

$$E(Y^2)=2\sum_{y=1}^{\infty}yq^{y^{\beta}}-E(Y) \quad (6)$$

The moments have no closed form expressions but numerical approximations can be obtained on a truncated support [24].

Equations (5) and (6) show that the cases of the mean are greater than the variance and the mean being lower than the variance are possible. This makes the DW distribution suitable for overdispersion and under-dispersion.

DW distribution has a nice property in that its τ ($0 < \tau < 1$) quantile, that is the smallest value of y for which $F(y) \geq \tau$, has a closed form expression which is given by;

$$Q(\tau)=\left\lceil\left(\frac{\log(1-\tau)}{\log(q)}\right)^{\frac{1}{\beta}}-1\right\rceil \quad (7)$$

with $\lceil . \rceil$ representing the ceiling function. Given that Y is non-negative and the cumulative density function is $1-q$ at 0, the quantile is defined only for $\tau \geq 1-q$.

This is in contrast to the Poisson and Negative Binomial regression, which do not have a closed form expression for quantiles.

The median of a DW distribution is considered a special case and is given by;

$$Q(0.5)=\left\lceil\left(\frac{\log(2)}{\log(q)}\right)^{\frac{1}{\beta}}-1\right\rceil \quad (8)$$

Hence the quantiles of a DW distribution are given by simple analytical formulae.

Parameter Estimation

Given a sample $y_1, y_2, \ldots, y_n$ from a DW distribution, the log-likelihood can be written as;

$$logL=\sum_{i-1}^{m}log\left(q^{y_i^{\beta}}-q^{(y_i+1)^{\beta}}\right) \quad (9)$$

From this it is easy to obtain the maximum likelihood estimators (MLE) of q and β by maximizing the log-likelihood directly using any standard non-linear optimization tool.

3.2.2. DW Regression Model

The advantageous property of the DW distribution is exploited within a regression context, where interest is to model the relationship between a count response variable and a set of covariates.

Model Formulation

DW regression model for count data is introduced in analogy with the continuous Weibull regression, which is mostly used in life-time modelling and survival analysis. The distribution function of a continuous Weibull distribution function is given by;

$$F(y;\lambda,\beta)=1-e^{-\lambda y^{\beta}}, y\geq 0 \quad (10)$$

with scale parameter q and shape parameter β.

The parameter q of a DW distribution is equivalent to $e-\lambda$ in the continuous case. Reference [25], Weibull regression imposes a log link between the parameter λ and the predictors. DW regression can be introduced through the parameter q.

From (9) with $\tau=\frac{1}{2}$, the median of Y denoted by M satisfies;

$$\log(M+1)=\frac{1}{\beta}\log(\log(2))-\frac{1}{\beta}\log(-\log(q)) \quad (11)$$

To introduce a DW regression model, assume for $i=1, 2,\ldots, n$, the response Y_i has a DW conditional distribution $f(y_i, q(x_i), \beta|x_i)$, where $q(x_i)$ is the DW parameter and is related to the independent variables x_i through a link function given as;

$$\log(-\log(q(x_i))-x_i'\alpha \quad (12)$$

$$x_i\alpha=\alpha_0+x_{i1}\alpha_1+\cdots+x_{ip}\alpha_p \quad (13)$$

This link function transforms q from the probability scale to interval $[-\infty, +\infty]$ and also ensures that this parameter remains in the interval [0, 1]. The log(-log) link function in q is motivated by the analytical formula for the quantile which facilitates the interpretation of the parameters. Furthermore, the DW regression model can be introduced by relating β to the explanatory variable $f(y_i, q, \beta(x_i)|x_i)$, or by adding a link to other parameters $f(y_i, q(x_i), \beta(x_i)|x_i)$. From (12) q_i can be expressed as;

$$q_i=e^{-e^{x_i\alpha}} \quad (14)$$

From this the conditional probability mass function of the dependent variable Y_i given $\boldsymbol{x_i}$ is given as;

$$f(y_i|x_i)=\left(e^{-e^{x_i'\alpha}}\right)^{y_i^{\beta}}-\left(e^{-e^{x_i'\alpha}}\right)^{(y_i+1)^{\beta}} \quad (15)$$

The likelihood of this equation is maximized numerically using standard non-linear optimization tools.

Interpretation of the regression coefficients

When a regression model has been estimated, we can obtain the mean and quantiles. The mean is obtained from the mean equation by calculating numerically using approximated moments of the DW regression [24].

The quantile equation (8) gives the median. Skewness is common in count data and for this reason, the median is more appropriate than the mean. Conditional median can easily be obtained from the closed form expression of quantiles for the DW regression as;

$$M(x) = \left\lceil \left(-\frac{\log(2)}{\log(q(x))}\right)^{\frac{1}{\beta}} - 1 \right\rceil \quad (16)$$

Taking the log(-log) link function combined with the analytical expression offers a way of interpreting the parameters. Substituting (14) to (16) gives;

$$\log(M(x)+1) = \frac{1}{\beta}\log(\log(2)) - \frac{1}{\beta}x_i'\alpha \quad (17)$$

The regression parameters α are interpreted in relation to the log of the median. This is an analogy with Poisson and NB models for which the parameters are linked to the mean. In particular, $\frac{\log(\log(2)-\alpha_0}{\beta}$ is related to the conditional median when all covariates are set zero, whereas $\frac{-\alpha_p}{\beta}, p = 1, \ldots, p$ can be related to the response corresponding to one unit of X_p, keeping all other covariates constant.

Diagnostic Checking

It is essential to consider a diagnostic analysis to investigate the appropriateness of a model after fitting it. For DW regression, since the response variable is discrete, a residual analysis was performed on the basis of randomized quantile residuals, as developed by [26].

In particular we let;

$$r_i = \phi^{-1}(u_i) \quad (18)$$

where $\phi(.)$ Is the standard normal distribution function and u_i is a uniform random variable on the interval;

$$(a_i, b_i] = \lim_{y \uparrow y_i}[F(y; \hat{q}, \hat{\beta})F(y_i; \hat{q}, \hat{\beta})] \quad (19)$$

$$(a_i, b_i] \approx [F(y_i - 1; \hat{q}_i, \hat{\beta})F(y_i; \hat{q}_i, \hat{\beta})] \quad (20)$$

Apart from the sampling variability in $\hat{q}_i$ *and* $\hat{\beta}$, these residuals follow the standard normal distribution. The validity of the DW model can therefore be assessed using the goodness of fit investigations of normality and residuals. Q-Q plots and normality tests were used.

3.3. Artificial Neural Network Model

Reference [27] defines Artificial Neural Network (ANN) as a parallel connection of a set of nodes referred to as neurons. It represents a function of explanatory variables which is composed of simple building blocks and which may be utilized to provide an approximation of the conditional expectations or, in particular, probabilities in regression. ANN is a non-parametric and data dependent technique. ANN are robust functions and analytical tools for predicting and classification problems that can model very complex non-linear functions to high accuracy levels using a process of learning that is similar to the learning process of the cognitive system of the human brain. [28, 29] are examples of research work where ANN has been applied for count data modelling. Multilayer Perceptron (MLP) is the most used architecture of ANN. MLP adopts back propagation (BP) algorithm as a learning process. This algorithm achieves the learning process by minimizing the sum of squared errors. ANN displays a complex input and output non-linear associations. MLP is made up of one input layer of units, a unit of output layer and a single or more hidden layers. The input units pass their inputs to the units in the first hidden layer or directly to the output units. Each hidden layer adds a constant (bias) to a weighted sum of its inputs and calculates an activation function of the result. This is then passed to hidden units in the next layer or to the output units. ANN adopts a set of input observations x_i and computes outputs y_i using a specified number of layers.

For this study involving data characterized by skewness, ANN with a median neuron model in the hidden layer is considered. The architecture of this model is shown in Figure 1. This figure represents a median neural model structure with k^{th} order and *m*-input. The weights between the inputs and the hidden layers is represented by W and is a matrix with *m* by *k* dimension. The hidden layer has *k* neurons which represents the order of the network. The incoming signals in the hidden layer constitutes the output of the neuron.

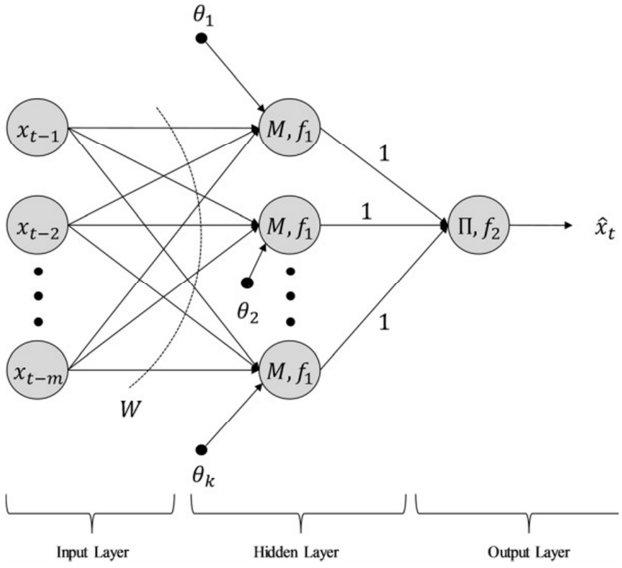

Figure 1. *Architecture of the median neuron model.*

Let $s_1, s_2, \cdots, s_m$ be incoming signals and y_i (i=1, 2, $\cdots$, k) be the output of the i^{th} neuron, this can be represented as;

$$y_i = Median(s_1, s_2, \ldots, s_m) \quad (21)$$

$$f_i(y_i) = y_i \tag{22}$$

The activation function in the hidden layer neurons are linear and $\theta_1, \theta_2, \dots, \theta_k$ are the bias terms. The output layer has a single neuron and multiplicative neuron model was used in this layer. The weights of the hidden and output layers are taken as 1 and the bias term is taken as 0. The activation function is a sigmoid activation function in the output layer. The output is calculated as;

$$f_2(y) = \frac{1}{1+e^{-y}} \tag{23}$$

The algorithm of how the output for a learning sample is computed is given below with $x_1, x_2, \cdots, x_m$ representing the input values of the learning sample.

The hidden layer neurons outputs are computed using the incoming signals given as;

$$h_j = f_1\left(Median\left\{w_{1j}x_1 + \theta_j, w_{2j}x_2 + \theta_j, \dots, w_{mj}x_m + \theta_j\right\}\right) j = 1, 2, \dots, k \tag{24}$$

The network output is computed using the outputs of the hidden layer and a sigmoid activation function.

$$o = f_2\left(\prod_{j=1}^{k} h_j\right) = \frac{1}{1+exp\left(\prod_{j=1}^{k} h_j\right)} \tag{25}$$

Model Development

The development of ANN with a three-layer network structure of a back propagation (BP) learning algorithm involves various steps which include, scaling and normalizing of raw data to an appropriate format, data division, determining the number of input nodes, hidden layers, hidden nodes, output nodes and also determining an activation function. There is also training by applying BP algorithm and finally evaluating the model.

Data normalization is done to smooth data so as to give better data generalization and improve performance. Normalization function is based on maximum and minimum values as suggested by [28]. The normalization formula is given as;

$$X_{new} = \frac{X_t - X_{min}}{X_{max} - X_{min}}(D_{max} - D_{min}) + D_{min}$$

Where;

X_t is the value that was normalized

X_{min} is the minimum value of the statistic variable

X_{max} is the minimum value of the statistic variable

D_{min} and D_{max} are minimum and maximum values needed for normalization.

In the data division step, data is divided into two parts, training set and testing set. The training set is used for model formulation and testing set is used for prediction. Data can be divided by a ratio such as 70%: 30%, 80%: 20% and 90%: 10%. A ratio of 70%: 30% is appropriate model and avoid over-fitting which was used in this study. The number of hidden nodes is done via a trial and error method. The hidden neuron has the ability to influence the error on the nodes to which their output is connected. Error is used to estimate the stability of a neural network. Better stability is indicated by minimal error. Excessive hidden neurons cause over fitting, that is, the neural network overestimates the complexity of the problem. The purpose of the hidden layers to detect the features to capture data pattern and to perform the complicated non-linear mapping between the input and output variables.

The learning rate and momentum is also considered where the value is a range from 0 to 1. A choice of the learning rate and momentum is very sensitive and a simple way to choose this is by trial and error method. There is a pre-defined stopping criterion; which acts as the core part of ANN. The criterion can either be that the number of iterations has been reached or the total sum of square errors is lower than a pre-determined value.

3.4. Performance Measures

The objective of each of the used models is to fit an accurate model for over-dispersed count data. The adequacy of the Artificial neural network and Discrete Weibull regression model is assessed using mean squared error (MSE) and root mean squared error (RMSE). An MSE value that is closer to zero, indicates a more useful model fit. The MSE is calculated as;

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2 \tag{26}$$

The RMSE is given as;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2} \tag{27}$$

The best model is the one with the least root mean squared error.

4. Data Simulation

This study used simulated data to fit the DW regression model and the ANN model. In the simulation process, different sample sizes 100, 250, 500 and 1000 were considered. For this case, over-dispersion has been considered as it is the most common form of dispersion for count data. Two covariates were simulated. The first one following a normal distribution N (0, 1) and the other one uniformly distributed with parameters (0, 1.5). The true regression parameters are assumed to be as follows;

$$\alpha = (\alpha_0, \alpha_1, \alpha_2) = (2, 0.5, 0.4) \tag{28}$$

Selection of the shape parameter β of the DW regression model is made in a way that over- dispersion of the data is achieved. In this study β was taken to be 0.9. The parameter q is calculated from each X using (14). The simulation of this data is iterated 1000 times.

To ensure that the generated data is dispersed in relation to the Poisson Model, the mean and variance of the dependent variable were obtained ad used to calculate the dispersion

coefficient. These summaries are given in Table 1. It can be seen that the *Dip* is greater than one for all the simulated sample sizes. This is an indication that the data is over-dispersed relative to the Poisson model.

Table 1. *Mean, variance and dispersion coefficient for various sample sizes.*

N	Mean	Variance	Dip
100	18.710	799.360	42.724
250	16.648	350.301	21.042
500	17.044	682.359	40.035
1000	16.969	560.669	33.041

To graphically illustrate the distribution of the simulated data, histogram for the different 1000 sample size is given in Figure 2.

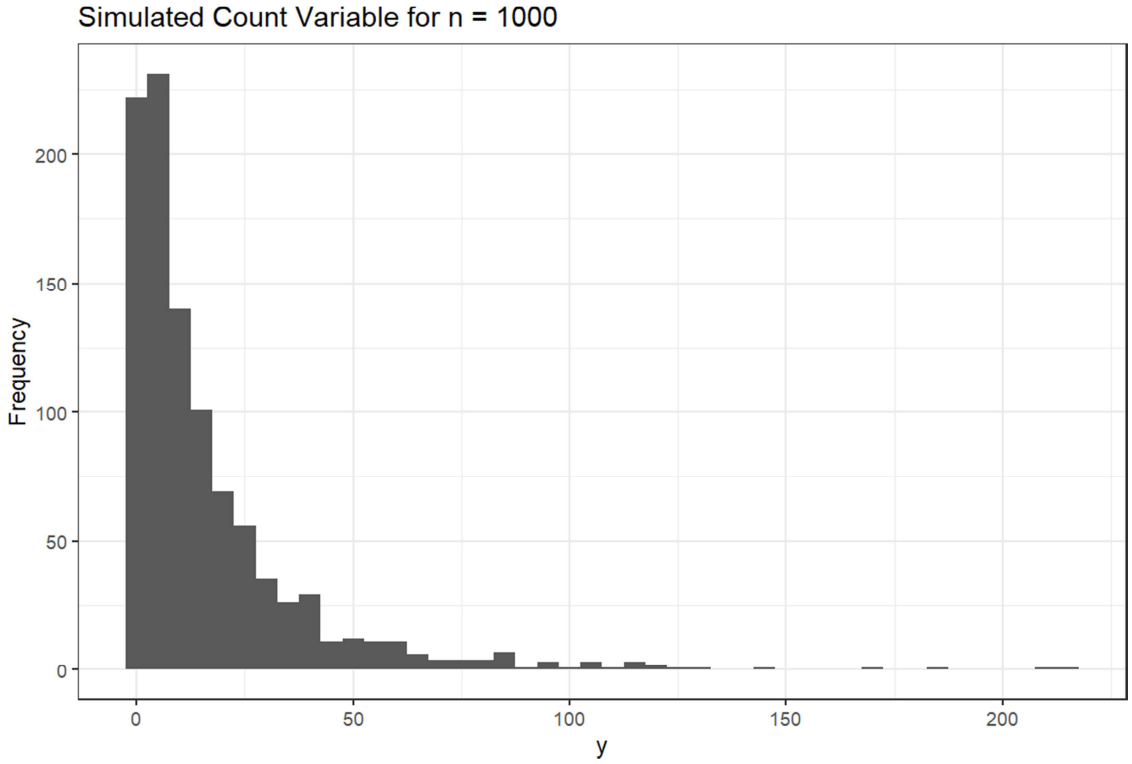

Figure 2. *Histogram showing the distribution of simulated count variable for 100 sample size.*

5. Results and Discussions

5.1. Fitting DW Regression Model

DW regression model is fitted on the simulated data with the varying sample sizes. The estimates of the parameter values are obtained and summarized in Table 2. From this table it can be seen that there are minimal variations from the true values of the parameters. It is also important to note that as the sample size increases, the variations from the true values decreases. The p-values of the parameters are less than 0.05 indicating that all the parameter estimates are significant at 0.05 level of significance.

Diagnostic analysis was performed to investigate the appropriateness of the fitted model. The normality of the residuals was tested using the Kolmogorov-Smirnov test. This test returned a p-values given in Table 3. All the P-values are greater than 0.05 indicating that the residuals follow a normal distribution. Furthermore, a simulated 95% envelope added to Q-Q plots were plotted as shown in Figure 3. From the plots, majority of the points lie within the envelope's bounds hence a good model fit.

Table 2. *Estimates from the fitted DW regression model.*

N	Parameter	Estimate	Std. error	P-value
100	α_0	2.1326	0.3056	<0.001
	α_1	0.5113	0.2406	0.0336
	α_2	0.4574	0.1067	<0.001
	β	0.9483	0.0749	<0.001
250	α_0	2.3374	0.2033	<0.001
	α_1	0.4478	0.1579	0.0046
	α_2	0.3700	0.0664	<0.001
	β	0.9923	0.0527	<0.001
500	α_0	1.8712	0.1334	<0.001
	α_1	0.5350	0.1083	<0.001
	α_2	0.4096	0.0487	<0.001
	β	0.8767	0.8767	<0.001
1000	α_0	2.0223	2.0223	<0.001
	α_1	0.4813	0.4813	<0.001
	α_2	0.3960	0.3960	<0.001
	β	0.8980	0.8980	<0.001

Table 3. *Kolmogorov-Smirnov test for normality P-values.*

N	KS. Test (P-value)
100	0.9830
250	0.9355
500	0.7039
1000	0.1687

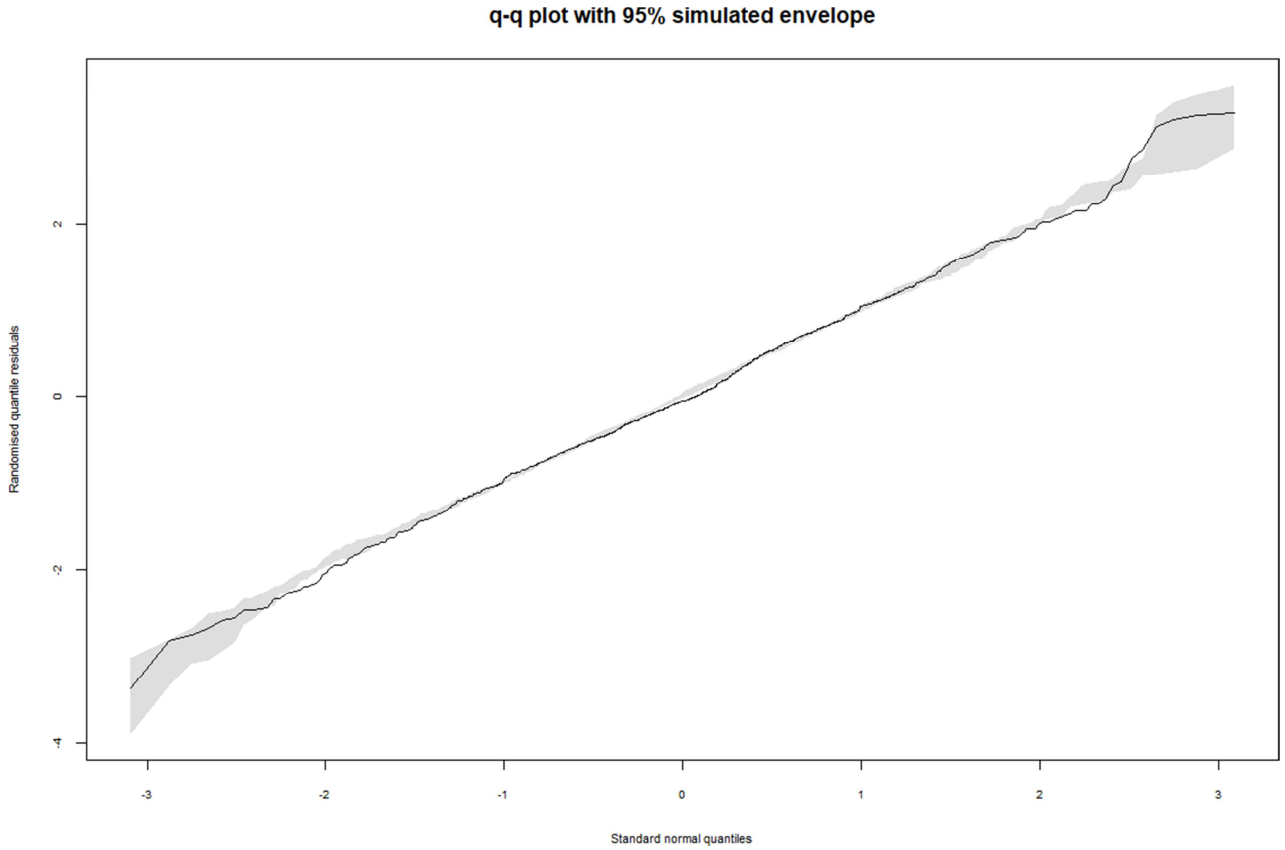

Figure 3. *Q-Q plot of randomized quantile residuals of the discrete Weibull regression model for 1000 sample size.*

5.2. Fitting ANN Model

ANN model is fitted on the data using median neuron model. A model with one hidden layer was fitted on the data. The selection of the number of nodes in the hidden layer was done through a trial and error method. Since the data had two input nodes, models with 1, 2 and 3 hidden nodes were tried and the one that produced minimal error was selected since it was considered to be stable. Models with one hidden node were selected for a sample size of 100 and 250. For the 500 and 1000 sample sizes, models with 3 and 2 hidden nodes were chosen respectively.

5.3. Performance Measures Comparison for Simulated Data

Table 4. *Summary of performance measures of the two models with varying sample sizes.*

N	Statistic	DW	ANN
100	MSE	2917.591	1026.521
	RMSE	54.0147	32.039
250	MSE	1157.197	1161.286
	RMSE	34.018	34.078
500	MSE	2203.939	1058.831
	RMSE	46.946	32.539
1000	MSE	1789.242	1350.225
	RMSE	42.299	36.745

The main aim of this study was to assess the performance of the two fitted models. Mean squared errors and residual mean squared errors were used to asses these models. The summary of this statistics is given in Table 4. From the results it can be seen that the MSE and RMSE is least for the ANN model in all the different sample sizes except for when n=250, these values are a higher for the DW regression model.

5.4. Application on Real Data

The ability of DW regression model and the ANN model to handle over-dispersion automatically was tested by applying this model on an over-dispersed data set. The data set used in this study is from German Health Survey. The data is available in the COUNT R package [31]. The data is comprised of three variables saved as a data frame. The data is saved as badhealth. The variables of the data set include;

1. numvisit - the number of visits made to the doctor and ranges from 0 to 40
2. badh - an indicator variable where 1 represents a patient claiming to be in bad health and 0 not.
3. age - the age of patient and ranges from 20 to 60 years.

The response variable for this study is taken to be numvisit which is a count variable with a sample mean of 2.353 and a sample variance of 11.98. The variance is larger than the mean indicating an over-dispersion relative to the Poisson model. The histogram showing the distribution of the data is given in Figure 4.

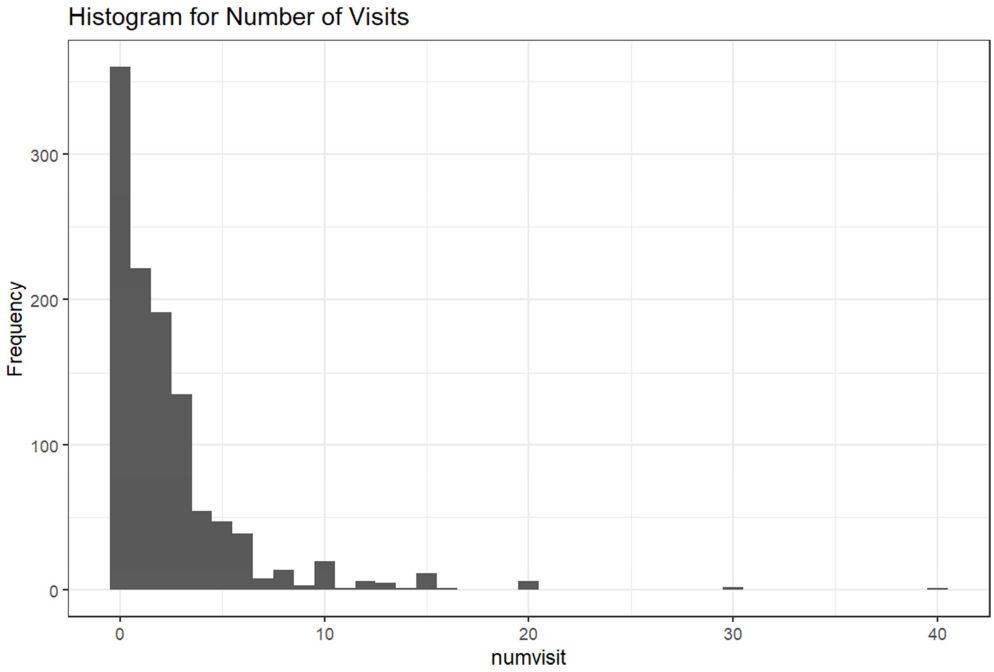

Figure 4. *Histogram of the badhealth data.*

The two models are fitted on the data starting with DW regression model whose model summaries are given in Table 5. The DW regression model fitted on the data resulted in P-values of estimates less than 0.05 indicating that the explanatory variables are significant at 5% level of significance. The model residuals are examined to ascertain the suitability of the model. Figure 5 of randomized quantile residuals shows that the residuals followed a normal distribution with many points falling within the simulated 95% envelope. Furthermore, the Kolmogorov-Smirnov test p-value is 0.06626.

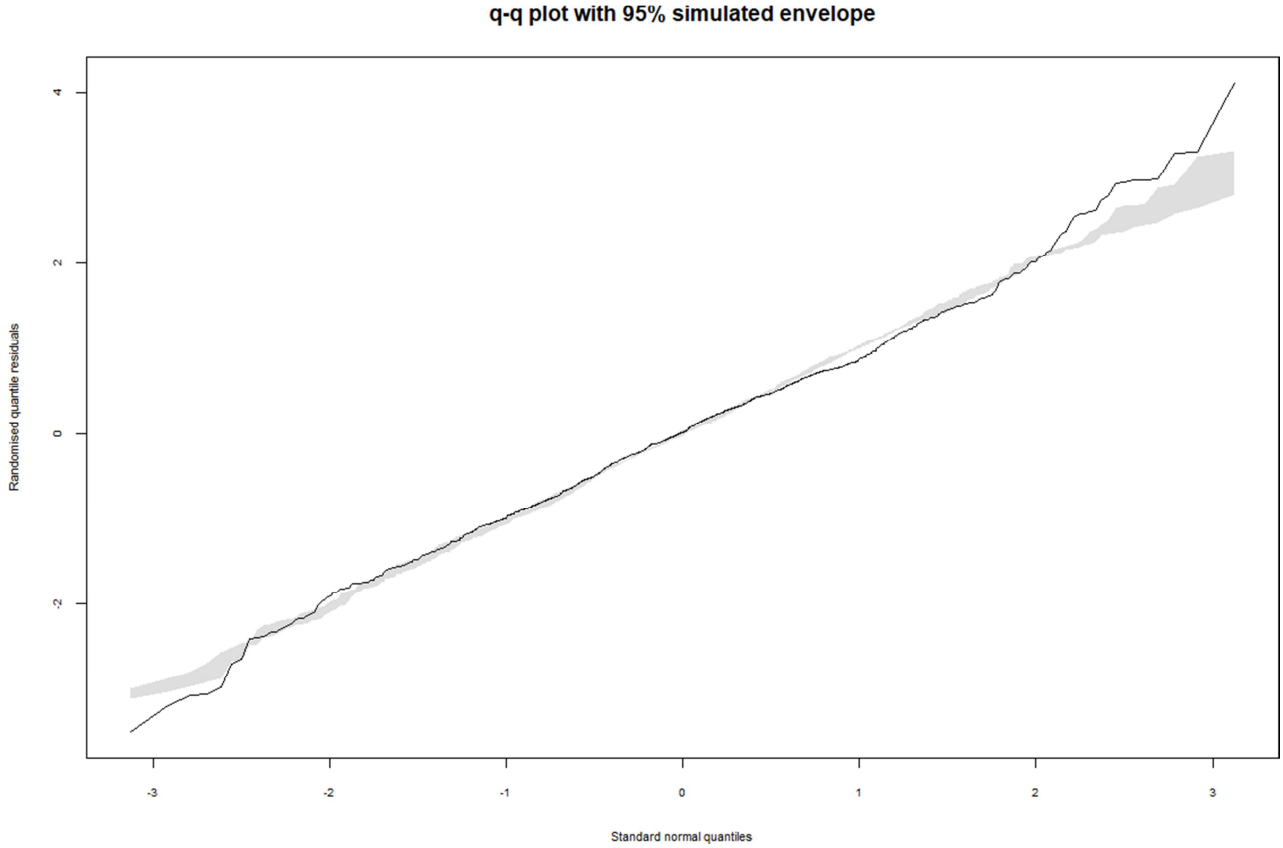

Figure 5. *Q-Q plot of randomized quantile residuals of the DW regression model fitted on badhealth data.*

Table 5. *Summary from DW regression model.*

	Estimate	Std. error	P-value
Intercept	0.390940	0.133024	0.00329
Badh	1.099217	0.112536	<0.001
Age	0.006794	0.003365	0.04347
Beta	0.988614	0.026497	<0.001

The ANN model was fitted to the data with two input nodes, one hidden layer with 3 hidden nodes and one output node. Three hidden nodes were selected after a trial and error on four different hidden nodes.

5.5. Performance Measures comparison for Real Data

The objective of each of these two methods was to fit an accurate model on the badhealth data and make a prediction using the fitted model. The adequacy of the DW regression model and the Artificial Neural Network model is assessed on the basis of MSE and RMSE. The results are summarized in Table 6. The Artificial Neural Network model performs better than the DW regression model since it had minimal values of MSE and RMSE. The RMSE for DW regression model is 69.0668 and that of ANN is 35.5652.

Table 6. *Summary of performance measures of the two models on badhealth data.*

Statistic	DW	ANN
MSE	4770.227	1264.884
RMSE	69.0668	35.56520

6. Conclusions and Recommendations

This study aimed at comparing the performance of DW regression model and ANN models. The two models were applied on a simulated data set and a real data set from German Health survey. The over-dispersion relative to the Poisson regression was considered in categorizing data as over-dispersed. The DW regression model has an attractive feature that is similar to the flexibility of the continuous Weibull distribution. DW regression model uses conditional quantiles unlike generalized linear models which uses the conditional mean as a central to interpretation. This property is useful for count data since they have a highly skewed distribution. The ANN model considered in this study is a high order neural network with a robust architecture. It considered a median neuron network so that it could adequately handle the skewed nature of over-dispersed count data.

The performance of the two models were done based on MSE and RMSE. From the results, ANN model with the median neuron generally outperformed the DW regression model both on the simulated data and the application on badhealth data set. Future research should try different robust statistics of ANN such as trimmed mean and also it can be useful to consider both parameters q and β of the DW regression as functions of the covariates.

References

[1] Karlaftis, M. G. and Tarko, A. P. (1998). Heterogeneity considerations in accident modeling. Accident Analysis & Prevention, 30 (4): 425–433.

[2] Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

[3] Chin, H. C. and Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, 35 (2): 253–259.

[4] Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation research part A: policy and practice*, 44 (5): 291–305.

[5] Hauer, E. (1997). Observational before/after studies in road safety. Estimating the effect of highway and traffic engineering measures on road safety.

[6] Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S., Boatwright, P., et al. (2006). Conjugate analysis of the conway-maxwell-poisson distribution. *Bayesian analysis*, 1 (2): 363–374.

[7] Consul, P. and Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 21 (1): 89–109.

[8] Sellers, K. F., Shmueli, G., et al. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, 4 (2): 943–961.

[9] Smith, D. and Faddy, M. (2016). Mean and variance modeling of under-and overdispersed count data. *Journal of Statistical Software*, 69 (6): 1–23.

[10] Sáez-Castillo, A. and Conde-Sánchez, A. (2013). A hyper-poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*, 61: 148–157.

[11] Chanialidis, C., Evers, L., Neocleous, T., and Nobile, A. (2018). Efficient bayesian inference for com-poisson regression models. *Statistics and Computing*, 28 (3): 595–608.

[12] Klakattawi, H., Vinciotti, V., and Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy*, 20 (2): 142.

[13] Lee, A. H., Stevenson, M. R., Wang, K., and Yau, K. K. (2002). Modeling young driver motor vehicle crashes: data with extra zeros. *Accident Analysis & Prevention*, 34 (4): 515–521.

[14] Berhanu, G. (2004). Models relating traffic safety with road environment and traffic flows on arterial roads in addis ababa. *Accident Analysis & Prevention*, 36 (5): 697–704.

[15] Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37 (1): 35–46.

[16] Lord, D. (2006). Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, 38 (4): 751–766.

[17] Lord, D., Geedipally, S. R., and Guikema, S. D. (2010). Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis: An International Journal*, 30 (8): 1268–1276.

[18] Winkelmann, R. and Zimmermann, K. F. (1995). Recent developments in count data modelling: theory and application. *Journal of economic surveys*, 9 (1): 1–24.

[19] Oh, J., Washington, S. P., and Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, 38 (2): 346–356.

[20] Hilbe, J. M. (2011). *Modeling count data*. Springer.

[21] Nakagawa, T. and Osaki, S. (1975). The discrete weibull distribution. *IEEE Transactions on Reliability*, 24 (5): 300–301.

[22] Roy, D. (2004). Discrete rayleigh distribution. *IEEE Transactions on Reliability*, 53 (2): 255–260.

[23] Sato, H., Ikota, M., Sugimoto, A., and Masuda, H. (1999). A new defect distribution metrology with a consistent discrete exponential formula and its applications. *IEEE Transactions on Semiconductor Manufacturing*, 12 (4): 409–418.

[24] Barbiero, A. (2015). Discreteweibull: Discrete weibull distributions (type 1 and 3), r package version 1.1.

[25] Da Silva, M. F., Ferrari, S. L. P., and Cribari-Neto, F. (2008). Improved likelihood inference for the shape parameter in weibull regression. *Journal of Statistical Computation and Simulation*, 78 (9): 789–811.

[26] Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5 (3): 236–244.

[27] Gichuhi, A. W. (2008). Nonparametric changepoint analysis for bernoulli random variables based on neural networks.

[28] Yunos, Z. M., Ali, A., Shamsyuddin, S. M., Ismail, N., et al. (2016a). Predictive modelling for motor insurance claims using artificial neural networks. *Int. J. Advance Soft Compu. Appl*, 8 (3).

[29] Haghani, S., Sedehi, M., and Kheiri, S. (2017). Artificial neural network to modeling zero- inflated count data: Application to predicting number of return to blood donation. *Journal of research in health sciences*, 17 (3): E1–4.

[30] Ke, J. and Liu, X. (2008). Empirical analysis of optimal hidden neurons in neural network modeling for stock prediction. In *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, volume 2, pages 828–832. IEEE.

[31] Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.