International Journal of Data Science and Analysis
2020; 6(6): 204-212
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200606.14
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Non-linear Approximations of Shape and Location Parameters in the Poisson Inverse Gaussian Model in Analysis of Infectious Count Data

Symon Kamuyu Matonyo, Oscar Ngesa, Anthony Wanjoya

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

Email address:
sykimats@gmail.com (S. K. Matonyo), oscanges@gmail.com (O. Ngesa), awanjoya@gmail.com (A. Wanjoya)

To cite this article:
Symon Kamuyu Matonyo, Oscar Ngesa, Anthony Wanjoya. Non-linear Approximations of Shape and Location Parameters in the Poisson Inverse Gaussian Model in Analysis of Infectious Count Data. *International Journal of Data Science and Analysis*. Vol. 6, No. 6, 2020, pp. 204-212. doi: 10.11648/j.ijdsa.20200606.14

Received: November 13, 2020; **Accepted**: November 21, 2020; **Published**: November 30, 2020

Abstract: Statistical models create a basis for analysis of infectious disesase count. These data sets exhibit unique characteristics such as low counts, delayed reporting, underreporting amoung others. The tendency to model these counts using linear models with their simplicity is common with most research. Further, the assumption of a fixed dispersion in modeling infectious disease counts is quite high. Prediction relating to infectious disease counts have been based on the Poisson model framework. The extension of the Poisson models such NB and PIG distributions have gained popularity over the recent past in modeling count responses showing over dispersion relative to the Poisson distribution. In this study we propose non-linear models for these data sets, modeling the mean and dispersion parameters as additive terms. Negative Binomial (NB) and Poisson Inverse Gaussian (PIG) glm models with a fixed and a varying dispersion parameter and compare them with NB GAM and PIG GAM with both mean and dispersion modeled as additive terms. The model are fitted to over dispersed infectious counts, Salmonella Hadar data set. Residual plots are constructed to explore the quality of fits and analysis goodness of fit is carried out to access the best fitting model. The study results reveal better performance of the PIG models on both the linear and non linear model platforms. Further, modelling both the mean and dispersion proved better as compared to models assuming the dispersion as a constant.

Keywords: Poisson Inverse Gaussian Distribution, General Additive Model, Dispersion, Count Models

1. Introduction

Count data is encountered on daily basis and dealings. The data exhibits unique characteristics such as over-dispersion, under-dispersion, incompleteness, presence of excess zeros among others. More understanding of such data and extraction of important information about it needs some statistical analysis or modelling. Various count data may posses different characteristics and therefore cannot be used with particular count data models. This necessitates the need for a systematic way in choosing of the best model that best describes the data in that, one should test whether the models assumptions are met rather than just picking a model naively. The nature of these data has led to development of various types of statistical models that are of great use in the statistical analysis of this type of data. The models and results vary according to the strength of the distributional assumptions made [4]. Most of these methods have now found their way into major statistical packages, which has greatly encouraged their application in variety of contexts.

Count data are most commonly modeled using regression based on the Poisson regression model [3], this regression has considerable limitation. Poisson distribution has one parameter which limits the variance from varying independently from the mean. Despite the fact that these models have been applied extensively in modelling count events [36], this assumption that the mean and variance of the data are equivalent is rather too restrictive and seldom does occur in observational data. This is a significant drawback since count data are often over- or underdispersed relative to the Poisson variance. More so, when the observed data involves excessive zero counts, over dispersion arises

hence may result in underestimation of the variance of the estimated parameter, resulting in wrong deductions [31].

Over-dispersion in regard to the Poisson model can be modeled by introduction of an additional parameter. For instance, in the negative binomial (NB) distribution and Generalized Poisson (GP) distributions [6], the models enable independent modelling of both, mean and variance by the incorporation of an additional parameter. They enable additional variation within the data to be accounted for by adding a randomly distributed error term, that is based on the Gamma distribution.

Further developments have been in progress with the view of coming up with a model that best analyses count data with their unique characteristics. Extension of the Poisson models such as the Conway-Maxwell-Poisson (COM-Poisson) distribution and its generalized regression model (GLM) have been proved to be exceptionally flexible in modelling count data [15, 14, 32, 13] whereas, the Double-Poisson (DP) and its generalized regression model; DP GLM performed better for count data with high mean scenarios independent of the type of dispersion [41].

The Sichel (SI) distribution is a model suitable for modelling count data that is highly dispersed [40]. This distribution is a compound form of Poisson distribution, that combines the Poisson distribution with generalized inverse Gamma distribution [35]. Recent research on mixed Poisson distributions for analyzing long-tailed count data stated that the SI model provides satisfactory inferences for so many cases compared to other models [16]. However, though the model seems adequate in modelling count data it suffers estimation problem. To overcome this challenge, reparameterization of the model was introduced, [35]. The special form of the SI distribution; the Poisson inverse Gaussian (PIG) in which the shape parameter in the SI model is set to -0.5 provides a better model for analysis of count data with slightly longer tails and excess kurtosis [39]. The PIG model has tractable nature in that its likelihood function is easily obtainable and has a closed form, indicating estimation of parameters as quite simple and less time consuming.

In the recent past, regression models extending modelling to dispersion, other shape parameters and to distributions beyond the exponential family have been so popular [10]. However, though much interest is still focused on modelling the mean, these models allow the flexibility of modelling shape parameter as a function of covariates. The PIG model parametrized in terms of the mean μ and dispersion parameter σ is available as response distribution in already existing regression software [34]. A study conducted on analysis of clinical trials where a secondary outcome was the number of falls that participants experienced while undergoing a drug treatment or usual care (control group: [18]), estimates of the treatment effect on the mean μ were found to be sensitive to the specification of the dispersion model. The sensitivity of the mean model to the dispersion model is of particular concern in the context of clinical trials, since statistical analysis plans do not generally specify modelling the dispersion parameter. A regression model using alternative parametrization of the PIG distribution, where the shape parameter σ is orthogonal to the mean should be considered [21]. But this parametrization of μ and σ leads to μ model estimates that are robust to misspecification of the dispersion model.

Misspecification of the model may result in realisation of biased estimates, that may in turn lead to erroneous and misleading conclusions. Regression models require that relationship between the variables (response and explanatory) conform to a particular functional form. Omission of important explanatory variables, failure to account for any non-linear components or critical interaction terms, or making measurement errors may lead to misspecification and accordingly bias the parameter estimators of one or more of the predictors in the regression model [5]. The study propose non-linear parameterization of the PIG model in location and dispersion parameters.

2. Literature Review

Counts are non negative integers which represents occurrences of event (s) within a fixed period of time. In numerous scientific and economic contexts the response variable is usually a count which needs to be analyzed in terms of a set of covariates. Common features off these counts is over dispersion or underdipersion with regard to the Poisson assumption. This arises when the variance of the counts data exceeds (overdispersion) or falls short of (underdispersion) the mean.

Regression models for count data

a) The Poisson Regression model

The Poisson model is a benchmark model for count data in much same way as the normal linear model is a bench mark for continuous data [30, 4, 36]. This count model distribution is focused on the number of outcomes of a given event. The model is derived from the Poisson probabilty mass function;

$$f(y_i;\lambda_i)=\frac{e^{-t_i\lambda_i}(t_i\lambda_i)^{y_i}}{y_i!}, y=0,1,\ldots..;\mu>0 \qquad (1)$$

where

y_i=count response

λ_i=rate parameter or predicted count

t_i=time or area in which counts enter the model.

When λ_i is considered as applying to individual counts without considering size or the time, $t_i=1$, where as, when $t_i>1$ it is known as an exposure and usually modeled as an offset.

Estimation of this model is based on log-likelihood parametarization of the Poisson probability distribution, which is focussed at establishing parameter values making the data most likely. The exponential family form is:

$$L(\mu_i;y_i)=\sum_{i=1}^{n}\{y_i\ln(\mu_i)-\mu_i-\ln(y_i!)\}, \qquad (2)$$

where μ_i symbolize the predicted count. The deviance function associated with this equation is applied when the model is estimated as a Generalized Linear Model (GLM). The response variable $y=0,1,2,\ldots$ uses the probability distribution function of the Poisson where:

$$y_i \sim p(\mu_i), y_i = 0,1,\ldots \text{ and } \mathbb{E}(y_i) = \mu_i$$

The mean response is given as $X_i'\beta = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{q-1} X_{iq-1}$ which is a function of linear predictor variables $X_{i1}, \ldots, X_{iq-1}$. The function $\mu(X_i, \beta)$ relates the mean response μ_i to the predictor variables X_i and the regression coefficients β. Hence, if $y_i \sim p(\mu_i)$ then, the mean(response) function is given as $\mu_i = exp(X_i\beta)$. Therefore, the Poisson regression mean response function can be expressed as;

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{q-1} X_{iq-1} \quad (3)$$

and its likelihood function is;

$$L(\beta) = \prod_{i=1}^{n} \frac{\{\mu(X_i,\beta)^{y_i} exp[-\mu(X_i,\beta)]\}}{y_i!} = \frac{\{\prod_{i=1}^{n} [\mu(X_i,\beta)]^{y_i}\} exp[-\sum_{i=1}^{n} \mu(X_i,\beta)]}{\prod_{i=1}^{n} Y_i!}$$

The log likelihood function is given by;

$$\ell = \log L(\beta) = \sum_{i=1}^{n} Y_i \ln[\mu(X_i,\beta)] - \sum_{i=1}^{n} \mu(X_i,\beta) - \sum_{i=1}^{n} \ln(Y_i!) \quad (4)$$

MLEs of $\beta_0, \beta_1, \ldots, \beta_{q-1}$, can be realized through numerical maximization procedures e.g, through the reweighted least square approach. The response function and fitted values $\hat{\mu} = \exp(X_i, b)$ for the regression function $\mu_i = \mu(X_i, \beta) = \exp(X_i', \beta)$ can be obtained from the parameter estimates of β.

The Poisson regression has advantages in that it takes into account the non-negative and discrete data hence, allows for drafting of conclusions on the probability of the occurrence of an event. The model may likewise be used as an option to Cox model in survival analysis, when the risk rates are around steady, amid the perception time frame and given that the danger associated with the occurrence is minimal [1]. However, despite its extensive use the Poisson model seems rather too restrictive due to the assumption that, the mean and variance of the data are equal. This is rarely the case with observation data since most count data in real life are either over- or under-dispersed. This violation of the Poisson necessitates the need for other models like, the Negative Binomial (NB) regression model and the Generalized Poisson (GP) model that are able to capture these characteristics of the data.

b) The Generalized Poisson (GP) Regression Model

The explicit presumptions of the Poisson regression models is that the force of Poisson process is a deterministic capacity of the covariates and the occurrences happen arbitrarily over some time. When handling count data characterised by under- or over-dispersion; where the sample variance is smaller (or larger) compared to the sample mean, the Poisson model leads to biased estimates of the parameters [7]. For over-dispersed data sets, the GP regression models provide a better alternative for the Poisson regression model [25, 28, 23, 24, 26]. The GP and NB regression models have been applied in order to handle over-dispersed count data in place of the Poisson model [25, 12]. For instance, where the type of dispersion exhibited by the data set is already known to be overdispersed then, one can either use the GP or NB regression models to model such data, otherwise, for an unknown type of dispersion the decision ought to be the GP model as it is more adaptable. The General Poisson regression can be expressed as;

$$g_i(z_i : \mu_i, x) = \left[\frac{\mu_i}{1+\alpha\mu_i}\right]^{z_i} \left[\frac{1+\alpha\mu_i}{z_i!}\right]^{z_i - 1} \exp\left(\frac{-\mu_i(1+\alpha z_i)}{1+\alpha\mu_i}\right) \quad (5)$$

with mean and variance;

$$\mathbb{E}(z_i|x_i) = \mu_i \; and Var(z_i|x_i) = \mu_i(1+\alpha\mu_i)^2$$

where,

$$z_i = 0,1,2,3,\ldots$$

$$\mu_i = \mu_i(x_i) = \exp(x_i, \beta)$$

$x_i = (p-1)$ dimensional vector of predictors and

$\beta = p$ dimension vector of parameters.

For $\alpha = 0$ the above model reduces to a Poisson model implying that the GP model is general form of the standard Poisson model. A comparison of the two models in terms of their capability to model over-dispersed count data showed that the GP regression out do the Poisson regression comparing their log-likelihood values [11]. Despite the fact that these models are able to capture overdispersion in count data, the models may not be adequate in handling data characterised by too many zeros.

Other regression models for handling count data

c) The Double Poisson (DP) Model and its extensions.

The Double Poisson regression model is a model within the frameworks of the double exponential family proposed by [9]. The distribution is derived from a mixture of two Poisson distributions, $P(\mu)$ and $P(z)$, i.e,

$$f(z, \mu, \theta) = K(\mu, \theta)[P(\mu)]^{\theta}[P(z)]^{1-\theta} \quad (6)$$

where,

θ =dispersion parameter

$K(\mu, \theta)$ =normalizing constant. The exact values of this constant are dependent on the values of μ and θ. The density function of the model is expressed as,

$$P(Z = z) = f_{\mu,\theta}(z) = K(\mu, \theta) f_{\mu,\theta}(z) \quad (7)$$

and the probability mass function is given by;

$$P(Z = z) = f_{\mu,\theta}(z) = \theta^{1/2} \exp(-\theta\mu) \left(\frac{e^{-z} z^z}{z!}\right) \left(\frac{e\mu}{z}\right)^{\theta z}, z = 0,1,2,\ldots \quad (8)$$

with mean and standard deviation (the exact density $f_{\mu,\theta}(z)$) given as;

$$\mathbb{E}(Z) \approx \mu, \quad (9)$$

$$SD(Z) \approx \left(\frac{\mu}{\theta}\right)^{1/2} \quad (10)$$

The normalizing constant $K(\mu,\theta)$ is approximately equal to 1 and can be obtained as;

$$\frac{1}{k(\mu,\theta)} = \sum_{z=0}^{\infty} f_{\mu,\theta}(z) \approx 1 + \frac{1-\theta}{12\mu\theta}\left(1 + \frac{1}{\mu\theta}\right) \quad (11)$$

This constant ensures that the density adds up to unity. This double poiusson model takes into account overdispersion ($for\ \theta > 0$) and underdispersion ($when\ \theta < 0$) and reduces to a Poisson distribution whenever $\theta = 1$. The estimates of the parameters μ and θ can be gotten through maximum likelihood approach. Due to the fact that the model has capability of handling over- and underdispersed count data, it has been applied in various research and several extensions developed.

Though this models seems attractable in its ability to handle over- and underdispersed counts, it is still limited in that the normalizing constant has no closed form hence leading to non exact results [36, 22]. The incorporation of this constant increases the non-linearity which results in difficulties in achieving the MLEs.

d) Conway-Maxwell (COM) Poisson models

The COM-Poisson is a generalization of the Poisson model, enabling for over- and underdispersed count data to be modelled and having a probabilty mass function given by;

$$P\{Y = y|\lambda, r\} = \frac{\lambda^y}{(y!)^r} \cdot \frac{1}{Z(\lambda,r)}, y = 0,1,2,\ldots \quad (12)$$

$$Z(\lambda, r) = \sum_{n=0}^{\infty} \frac{\lambda^n}{(n!)^r}, \lambda > 0\ and\ r \geq 0 \quad (13)$$

where,

y is a random count of a discrete nature

$\lambda =$ centering parameter (approximately equal to the expected value in many cases)

$r =$shape parameter for the model

Setting $r = 0, \lambda < 1$ results in the geometric distribution; while when $r \to \infty$ the model converges to a Bernoulli Distribution otherwise, when $r = 1$ results in Poisson model. The adaptability of the COM-Poisson model greatly extends its utilization for count occurences. The mean and the variance of the distribution are given as;

$$\mathbb{E}(Y) = \frac{\partial \log Z}{\partial \log \lambda} \quad (14)$$

$$Var(Y) = \frac{\partial^2 \log Z}{\partial \log^2 \lambda} \quad (15)$$

Despite having some attractable characteristics and being more robust it has limitations in its applicability as a basis for GLMs since the parameters λ and r lack a clear centering parameter [15]. Actually, while λ almost equals the mean for r close to 1, this differs at a greater extent from the expected value for small values of r. When the data is over-scattered, r would be required to be small and along these lines a COM-Poisson dependent on the original formulation would be extremely hard to decipher and use for the over-scattered count data. To overcome this challenge introduced a reparametrization of μ to $\mu = \lambda^1/r$ was introduced in the original distribution to provide a clear centering parameter [15].

e) The Sichel (SI) Distribution

The Sichel (SI) distribution, which is a compound Poisson model was introduced by [33]. The distribution combines the Poisson distribution with the generalized inverse Gaussian distribution. This resultant combination is in particular very useful in modelling count data that show overdispersion in respect to the Poisson model and has produced satisfactory results in many instances where other count models were inadequate. The distribution contains three parameters $0 < \theta < 1, \sigma > 0$ and $-\infty < \gamma < \infty$ and can be expressed as;

$$p(y|\mu,\sigma,\gamma) = \frac{(\mu/d)^y K_{y+\gamma}(\alpha)}{K_\gamma(1/\sigma) y!(\alpha\sigma)^{y+\gamma}} \quad (16)$$

where,

$y =$dependent variable

$\mu =$expected value of the observations

$\sigma =$scale parameter

$\gamma =$shape parameter

$\alpha^2 = \sigma^{-2} + 2\mu(d\sigma)^{-1}$ and

$$d = \frac{K_{\gamma+1}(1/\sigma)}{K_\gamma(1/\sigma)}$$

The fucntion $K_\gamma(t)$ is referred to as the modified Bessel function expressed as;

$$K_\gamma(t) = \frac{1}{2}\int_0^\infty X^{\gamma-1} \exp\left(-\frac{1}{2}t(x + x^{-1})\right) dx \quad (17)$$

The central moments of the SI distribution are given as;

$$\mathbb{E}(Y) = \mu, Var(Y) = \mu^2[2\sigma(\gamma+1)/d + 1/d^2 - 1]$$

As $\sigma \to \infty$ and $\gamma > 0$, the distribution converges to a NB distribution. For, $\gamma = -0.5$, the distribution reduces to a Poisson-inverse Gaussian with an expected value μ and variance $\mu + \mu^2\sigma$ [39]. The likelihood function of the PIG model is effectively realistic and has a closed form, showing estimation of parameters very straightforward and nearly takes no time [27]. Various research in the fields of medicine and transport have demonstrated that PIG provides a better fits in modeling count data with longer tails and excess Kurtosis [39]. Recently, an orthogonal parametrization of this dispersion model was proposed and its performance tested on a clinical trial data [20].

3. Methodology

a) The Poisson Inverse Gaussian (PIG) Regression Model

The PIG regression model is a special form of the Sichel distribution proposed by [8]. In this model, the shape parameter in the Sichel distribution is set to be a constant, that is, $\gamma = -\frac{1}{2}$. Therefore, the model is characterized by two parameters in contrast with the SI model. This model is more tractable and very useful when handling data exhibiting longer tails compared to those of a NB model [8, 35]. A variety of Poisson mixture distributions that can be used to handle count data have been proposed, [29]. The distributions are characterized by the two parameters μ and σ where the expected value and the variance are given as $\mathbb{E}(Y) = \mu$ and $Var(Y) = \mu(1 + \sigma\mu)$. Under these parametrization the probabilty denstity function function is expressed as;

$$f_\lambda(y|\mu,\sigma)=\sqrt{\frac{2}{\pi\sigma}}(1+2\mu\sigma)^{\frac{1}{4}}e^{\frac{1}{\sigma}}\frac{\left(\mu/\sqrt{(1+2\mu\sigma)}\right)^y}{y!}K_{y-0.5}\left(\sqrt{(1+2\mu\sigma)/\sigma}\right),y=0,1,\ldots \quad (18)$$

with the inverse Gaussian distribution $\lambda\sim IG(\mu,\sigma)$ expressed as;

$$f_\lambda(\lambda|\mu,\sigma)=(2\pi\sigma^2\lambda^3)^{-\frac{1}{2}}\exp\left(-\frac{(\lambda-\mu)^2}{2\mu^2\sigma^2\lambda}\right),\lambda>0$$

and $K_\gamma(t)$ is a modified Bessel function of a third kind. This parametrization was first used in literature with $\tau=\sigma$, [8]. As $\sigma\to 0$ the distribution converges to a Poisson distribution. Under this parametrization the distribution can be communicated as a multiplicative arbitrary effect model. For instance, given;

$$Y\sim Poisson(t\mu\varepsilon)\ and\ \varepsilon\sim IG(1,\sigma) \quad (19)$$

where t represents the moddel offset then we can have the model as;

$$Y\sim PIG(t\mu,\sigma)$$

The PIG distribution has been presented in various literature texts in different parametrization. It was first introduced with parameters α and λ with $\alpha>0$ and $0<\lambda<1$, [33]. Under this parametrization the distribution's expected value is expressed as $\mathbb{E}(Y)=\frac{\alpha\lambda(1-\lambda)^{\frac{1}{2}}}{2}$ with a variance $Var(Y)=\alpha\lambda(2-\lambda)/(4\sqrt{(1-\lambda)^3})$. However, the asymptotic correlation of the Likelihood estimators $\hat{\alpha}$ and $\hat{\lambda}$ are emphatically negative under this representation. To solve this limitation an orthogonal representation (μ,α) was developed, [35]. This results in a pdf of the following format;

$$f(y|\mu,\alpha)=\left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}}\exp\{\sqrt{(\mu^2+\alpha^2)}-\mu\}\frac{\left(\mu(\sqrt{\mu^2+\alpha^2}-\mu)\right)^y}{\alpha^y y!}K_{y-\frac{1}{2}(\alpha)},\mu>0,\alpha>0 \quad (20)$$

with a mean and variance expressed as;

$$\mathbb{E}(Y)=\mu\ and\ Var(Y)=\mu(1+\mu)/\left(\sqrt{\mu^2+\alpha^2}-\mu\right)$$

In this representation σ is inversely related to α , i.e, $\alpha=\frac{\sqrt{1+2\mu\sigma}}{\sigma}$ and as the value of $\alpha\to\infty$ the distribution converges to Poisson. It is not possible to express this model as a multiplicative arbitrary effect model as in the previous parametrization (19).

The aspect of the PIG model based on orthogonal parametrization of the mean and shape parameter was investigated and the model appllied on data from clinical trial for the number of falls that patients experienced while undergoing a drug treatments [20]. The PIG distribution is represented as a respeonse distribution with;

$$g(\mu)=x'\beta;\ h(\sigma)=\omega'\delta$$

The study assumes an orthogonal parametrization proposed and introduce general additive models for both the mean and the dispersion parameters in the PIG model assuming;

$$g(\mu)=x'\beta;\ h(\alpha)=\omega'\gamma \quad (21)$$

b) Model Formulation

Let Y be a response variable from the exponential family distributions and assuming Y is parameterized by θ then the General additive model (GAM) for the distribution parameters can be expressed as;

$$g(\theta)=\alpha+\sum_{j=1}^{m}f_j(x_j)+\varepsilon \quad (22)$$

where α is a constant term, $f_j's$ are unspecified smooth functions of the covariates $x_j,j=1,2,\ldots,m$, $g(\cdot)$ is a smooth monotonic link function which is known and $\varepsilon\sim N(0,\sigma^2)$ is error term [17, 37]. The model above can be represented otherwise using basis expansions for each smoother with its corresponding penalty and estimation carried out by penalized regression approaches with appropriate degree of smoothness for the $f_j's$ estimated from the data via marginal likelihood maximization or cross validation approaches [37]. Under this approach a smooth function f_j can be represented by a set of q spline basis functions $b_{ji}(x)$ hence;

$$f_j=\sum_{i}^{q}b_{ji}(x)\cdot\beta_{ji}$$

where β_{ji} is the smoothing coefficient associated with the j^{th} function. $\hat{\beta}$ is estimated by maximizing the penalized log-likelihood;

$$\ell_p(\beta)=\ell(\beta)-\frac{1}{2}\sum_{j=1}^{m}\lambda_j\beta'D_j\beta \quad (23)$$

where $\ell.$ is the log-likelihood function,λ_j is the smoothing parameter for the jth function f_j and D_j is known matrix of coefficients [38].

In this study we seek to develop GAMs, for both the mean and dispersion parameters under the orthogonal parametrization, that is, $Y\sim PIG(\mu,\alpha)$. We assume a mean and dispersion model of the form;

$$g(\mu_i)=\beta_0+\sum_{j=1}^{m}f_j(x_{ij})+\varepsilon \quad (24)$$

$$h(\sigma_i)=\gamma_0+\sum_{j=1}^{m}s_j(x_{ij}) \quad (25)$$

where,

$g(\mu_i)=\log(\mu_i)$ and $h(\sigma_i)=\log(\sigma_i)$ are the monotonic link functions

$x_i's$ are observed data variables

$\varepsilon = \log t_i$ is the offset term for an observation time t_i

c) Model specification

A parameter-observational model for infectious diseases takes the following functional form;

$$\mu_t = exp(\eta_t + \alpha y_{t-1}) \quad (26)$$

where the model parameters μ_t is the mean infections per week and η_t and αy_{t-1} are the endemic and is epidemic components respectively [19]. The endemic component η_t is considered to be parameter driven. Due to the fact that most infectious disease data exhibit seasonality, $log(\eta_t)$ is modeled as the sum of S harmonic waves having different frequencies and an intercept term as;

$$\eta_t = \nu + \beta t + \sum_{s=1}^{s} (\gamma_s sin(\omega_s t) + \delta_s cos(\omega_s t)) \quad (27)$$

where,

s=number of harmonics

$\omega_s = \frac{2s\pi}{q}$ are fourier frequencies with q as the base frequency. The epidemic component αy_{t-1} is an observational driven process through α. For values of α between 0 and 1, the model depicts occasional epidemic outbreaks which a branching process with immigration. The process is ergodic in situations where $\alpha < 1$ and has an exponential increase for values of $\alpha > 1$. However, in cases where $\alpha = 0$, the model reduces to a parameter driven formulation without an epidemic incidences.

The model serves a link for infectious disease counts mean and the explanatory variables. The dispersion parameter is modeled as a function of the explanatory variables to explain extra variation in the data. The study modelled mean infection model as a function of additive terms:

$$log(\mu_t) = f(\eta_t) + f(y_{t-1}) \quad (28)$$

where,$f(\eta_t)$ is the endemic additive term given as:

$$f(\eta_t) = 1 + f(t) + f(sin(\omega_s t) + f(cos(\omega_s t)) \quad (29)$$

and $f(y_{t-1})$ is the epidemic component considered to be observational driven. Further the dispersion in the data was modelled as additive terms of the both the endemic and epidemic components under the Poisson inverse model as:

$$log(\sigma_t) = f(\eta_t) + f(y_{t-1}) \quad (30)$$

where parameters are as earlier defined.

d) Description of the Data (Salmonella Hadar Data)

The Salmonella Hadar data set contained 295 observations of the disease recorded over a period of six years (2001-2006) in Germany. The distribution of the responses, in figure 1, depicts a highly peaked data with a long tail and skewed to the right with a kurtosis and skewness coefficients of 7.3 and 1.86 respectively. The ratio of variance to the mean was 3.45 an indication of presence of over-dispersion in the data [2]. A Poisson glm model fiited on the data recorded a dispersion of 1.96. This showed that the data was over-dispersed relative to the Poisson distribution.

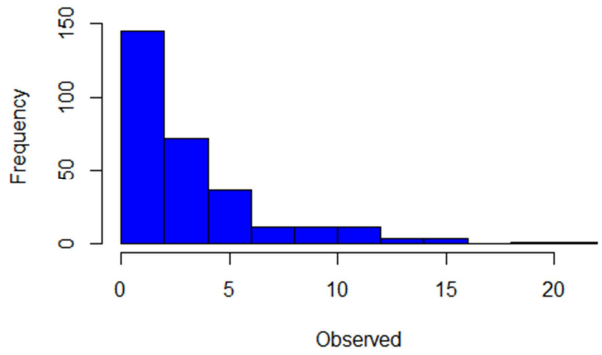

Figure 1. *Salmonella Hadar data distribution.*

4. Results and Discussions

a) Linear models

Generalized linear models (GLM) were fitted for both data sets under the Poisson, Negative Binomial and PIG distributions and model performance examined. Goodness of fit of the models was examined based on values of Akaike Information Criterion (AIC), global deviance and Bayesian Information Criterion (BIC). Model (s) considered to have best fit for the data were those that recorded the smallest values of these measures. Table 1 shows the summary results for the Poisson, Negative Binomial (NBI) and Poisson Inverse Gaussian (PIG) linear model fits with both fixed and varying dispersion parameters for the study data.

Table 1. *Modelling results for Poisson, NB and PIG glm models.*

Model	**PO**	**NB**		**PIG**	
Parameters	**Fixed**	**Fixed**	**Varying**	**Fixed**	**Varying**
ν	0.9861 (0.0670)	1.0138 (0.0950)	0.9804 (0.1001)	1.0276 (0.0976)	0.9986 (0.1010)
β	-0.0011 (0.00034)	-0.0016 (0.0005)	-0.0014 (0.0005)	-0.0016 (0.0005)	-0.0014 (0.0005)
γ	-0.1497 (0.0478)	-0.1426 (0.0643)	-0.1409 (0.0639)	-0.1481 (0.0654)	-0.1449 (0.0649)
δ	-0.4848 (0.0523)	-0.4807 (0.0702)	-0.4974 (0.0724)	-0.4888 (0.0709)	-0.5038 (0.0726)
α	0.0764 (0.0074)	0.0862 (0.0126)	0.0875 (0.0125)	0.0838 (0.0124)	0.0853 (0.0122)
$\tau: \lambda$	-	-1.500 (0.194)	-	-1.4009 (0.2092)	
γ_0	-	-	-0.7193 (0.7636)	-	-0.6916 (0.7211)
γ_1	-	-	-0.1782 (0.1729)	-	-0.1647 (0.1633)
	-	-	-	-	-
$GlobalDev.$	1313.972	1240.889	1239.903	1237.772	1236.804
AIC:	1323.972	1253.903	1252.889	1250.804	1249.772
SBC:	1342.39	1279.688	1274.991	1276.589	1271.874

*Numbers in brackets are the standard errors for the estimates.

From these summary results it is evident that the NB and PIG models had similar estimates. For instance, the past observations of the disease had a positive relationship with the current infection frequency. Further, the PIG model had better fits for both the fixed and varying dispersion parameters. The models with a varying dispersion parameter showed better fits for the data compared to those with a fixed dispersion. This is evident from the small values of AIC, SBC and Global deviance values. This is an indication that the Inverse Gaussian part in the PIG model has a better flexibility to handle over dispersed data common in infectious disease compared to the Gamma part present in NB models. Efficiency of the PIG model is facilitated by the varying dispersion parameters in the model that help to capture variability within the data.

b) Non-linear model fits

General additive models were fitted for Poisson, NB and PIG models. General additive term models for the mean assuming the dispersion was fixed were fitted to the data under the models and results compared. Further both the mean and dispersion were modeled as smooth additive terms and model performance investigated. Table 2, shows the summary results for the AIC, SBC and global deviance for the fitted models.

Table 2. *GAM models AIC, SBC and Global deviance comparisons.*

model	parameter modelled	AIC	SBC	Global deviance
Poisson	Log (μ_i)	1238.596	1274.97	1218.846
NB (m1)	Log (μ_i)	1212.634	1252.028	1189.702
NB (m2)	Log (μ_i); log (σ_i)	1211.622	1251.414	1188.845
PIG (m1)	Log (μ_i)	1209.564	1250.46	1187.36
PIG (m2)	Log (μ_i); log (σ_i)	1199.047	1249.795	1164.435

Evidently, from the observed summary results the PIG GAM model produced a better fitted model for the study data. This is indicated by the small values of AIC (1199.047, 1209.564) recorded for the model compared to the NB model AICs (1212.634, 1211.622). Modeling the mean and the dispersion parameters as additive terms increased the model performance further by far. This is clear indication that modeling the mean parameter alone may result in model misspecification, [20]. A comparison of the AIC of the additive models and their linear counterparts showed that the additive models had better fits compared to the glm models previously fitted. This implies that, non-linear additive models are best suited to fit overdispersed count data as they produced better fits compared to the linear models.

c) Residual Analysis

For any fitted model, true residuals r_i have standard normal distribution irrespective of the model distribution. The randomized quantile residual summary for the additive models are shown in Table 3.

Table 3. *Randomized normal quantile residuals for the additive models.*

GAM Model	Mean	Variance	Skewness	Kurtosis
NB (m1)	0.011	1.002	0.220	3.42
NB (m2)	-0.0009	0.951	0.258	3.48
PIG (m1)	-0.0073	0.981	0.133	3.28
PIG (m2)	0.002	0.942	0.044	2.874

The normalized quantile residuals for the fitted PIG models behave well. Their means are nearly zero, variance nearly one, kurtosis is nearly 3 and the values fall inside the acceptance region, except for model skewness. The model residuals shows some right skewness. The residual distributions for NB the models suggests a leptokurtic behavior which is slightly skewed to the right. The residuals from the PIG model with additive terms for the mean parameter are leptokurtic and skewed to the right. The figures 2 and 3 show the residual worm plots for the additive models with fixed dispersion and an additive dispersion model respectively.

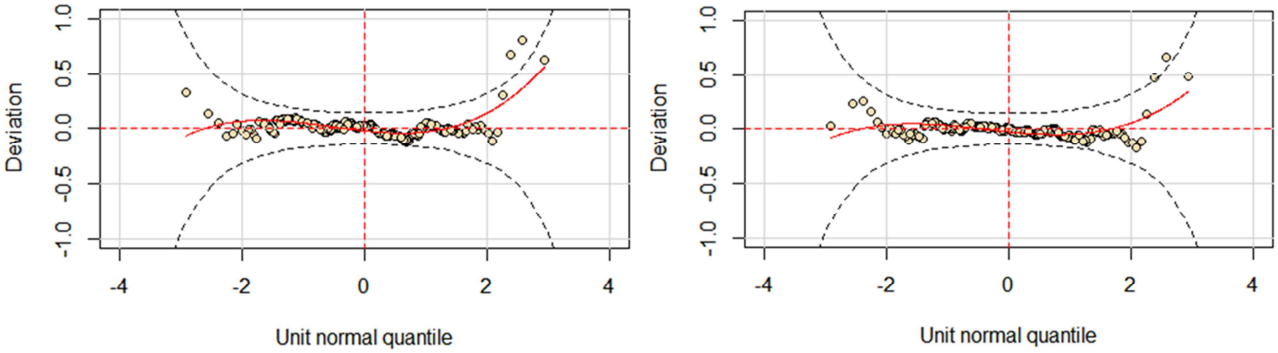

Figure 2. *Residual worm plots for NB and PIG GAM models with fixed dispersion.*

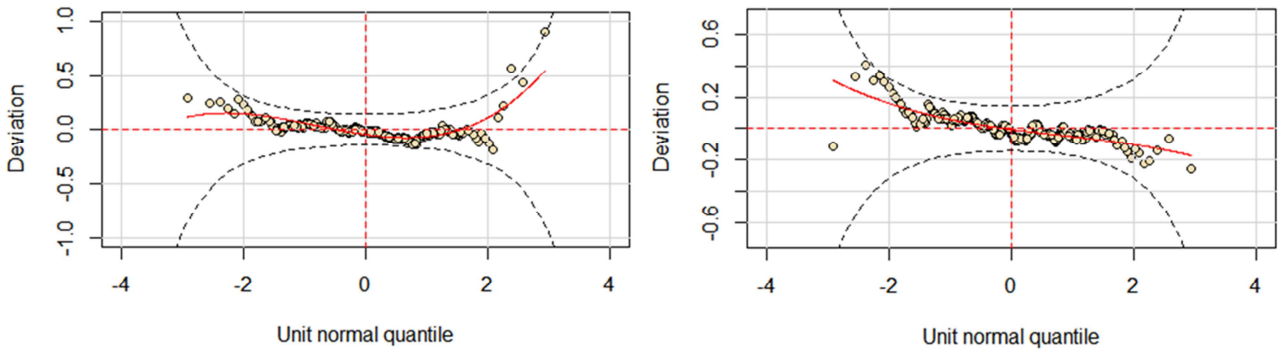

Figure 3. *Residual worm plots for NB and PIG GAM models with additive dispersion.*

The residual worm plots show that the model fits the data fairly well. However, the plots for the additive models where the dispersion parameter was assumed to be fixed indicate the model did not fit the data very well as shown by some points lying outside the confidence bounds. An in depth insight to the residual worm plots for both the fixed dispersion models and additive dispersion reveal some misfits of the penalized curves fitted to the residual points for given data. This serves as an explanation to the disparities seen from the summary of randomized quantile residuals for the models. The misfits observed for these models were as follows: (0.5, 23.5) and (74.5, 98.5) for the skewness and kurtosis of the residuals, (98.5, 122.5) for kurtosis and (245.5, 270.5) for the variance of the residuals in the eleventh interval for the NB model with a fixed dispersion parameter. The PIG model with fixed dispersion had misfits in the intervals; (0.5, 23.5), (68.5, 90.5) and (226.5, 249.5) for the skewness coefficient and (0.5, 23.5) for the kurtosis. Though the additive dispersion models showed improved model fits of the residual worm plots, an in depth analysis of these plots had evidence of misfits in some intervals. The NB additive model had misfits in: (74.5, 98.5) for skewness coefficient, (0.5, 24.5), (74.5, 98.5) and (172.5, 196.5) for peak coefficients. The PIG additive model recorded misfits at the intervals, (0.5, 24.5) and (74.5, 98.5) for the skewness and kurtosis of the model residuals.

5. Conclusions

The study applied non linear models for both the mean and dispersion parameters of the PIG distribution to overdispersed infectious disease count data. General additive models of the PIG and NB distribution were fitted to infectious disease counts. The linear model fits for counts indicated that PIG glm models with varying dispersion parameter had better performance in fitting the data as opposed to the other fits. More so, PIG GAM overall provided better model results compared to NB GAM models and the linear forms of the two models. This could be attributed to the flexibility of the Inverse Gaussian part of present in this model. Further, the findings showed that models having both the mean and dispersion parameters as varying terms had better performance as compared to models with a fixed dispersion paramater. This is an indication that modeling the mean parameter in a distribution assuming the dispersion may have resulted in the poor performance of these models. From the study results the PIG and NB models had almost similar estimates for the smooth terms and linear coefficients implying that these models can be applied alternatively depending on the data structures. This study proposes the PIG GAM models as a better distribution for modelling overdispersed infectious disease count.

References

[1] Norman E. Breslow and N. E. Day. Statistical methods in cancer research. vol. 2. the design and analysis of cohort studies. *Lyon, France: International Agency for Research on Cancer* 1987.

[2] M. G. Bulmer. Principles of statistics dover publications. *New York*, 1979.

[3] A. Colin Cameron and Pravin K. Trivedi. Essentials of count data regression. *A companion to theoretical econometrics*, 331, 2001.

[4] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[5] Wansu Chen, Lei Qian, Jiaxiao Shi, and Meredith Franklin. Comparing performance between log-binomial and robust poisson regression models for estimating risk ratios under model misspecification. *BMC medical research methodology*, 18 (1): 63, 2018.

[6] Prem C Consul and Gaurav C Jain. A generalization of the poisson distribution. *Technometrics*, 15 (4): 791–799, 1973.

[7] David R. Cox. Some remarks on overdispersion. *Biometrik*, 70 (1): 269–274, 1983.

[8] C Dean, JF Lawless, and GE Willmot. A mixed poisson–inverse-gaussian regression model. *Canadian Journal of Statistics*, 17 (2): 171–181, 1989.

[9] Bradley Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81 (395): 709–721, 1986.

[10] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. Regression models. In *Regression*, pages 21–72. Springer, 2013.

[11] Felix Famoye, John T. Wulu, and Karan P. Singh. On the generalized poisson regression model with an application to accident data. *Journal of Data Science*, 2 (2004): 287–295, 2004.

[12] Felix Famoye. Restricted generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 22 (5): 1335–1354, 1993.

[13] Royce A. Francis, Srinivas Reddy Geedipally, Seth D. Guikema, Soma Sekhar Dhavala, Dominique Lord, and Sarah LaRocca. Characterizing the performance of the conway-maxwell poisson generalized linear model. *Risk Analysis: An International Journal*, 32 (1): 167–183, 2012.

[14] Srinivas Geedipally and Dominique Lord. Effects of varying dispersion parameter of poisson-gamma models on estimation of confidence intervals of crash prediction models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061): 46–54, 2008.

[15] Seth D. Guikema and Jeremy P. Goffelt. A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal*, 28 (1): 213–223, 2008.

[16] Pushpa Lata Gupta, Ramesh C. Gupta, and Ram C. Tripathi. Score test for zero inflated generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 33 (1): 47–64, 2005.

[17] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1 (3): 297–318, 1986.

[18] Robert A Hauser, Stephane Heritier, Gerald J Rowse, L Arthur Hewitt, and Stuart H Isaacson. Droxidopa and reduced falls in a trial of parkinson disease patients with neurogenic orthostatic hypotension. *Clinical neuropharmacology*, 39 (5): 220, 2016.

[19] Leonhard Held, Michael Höhle, and Mathias Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical modelling*, 5 (3): 187–199, 2005.

[20] Gillian Z. Heller, Dominique-Laurent Couturier, and Stephane R. Heritier. Beyond mean modelling: Bias due to misspecification of dispersion in poisson-inverse gaussian regression. *Biometrical Journal*, 2018.

[21] Gillian Z. Heller, Dominique-Laurent Couturier, and Stephane R. Heritier. Beyond mean modelling: Bias due to misspecification of dispersion in poisson-inverse gaussian regression. *Biometrical Journal*, 61 (2): 333–342, 2019.

[22] Joseph M. Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[23] Naratip Jansakul. Fitting a zero-inflated negative binomial model via r. In *Proceedings 20th International Workshop on Statistical Modelling. Sidney, Australia*, pages 277–284, 2005.

[24] B. M. Golam Kibria. Applications of some discrete regression models for count data. *Pakistan Journal of Statistics and Operation Research*, 2 (1): 1–16, 2006.

[25] Jerald F Lawless. Negative binomial and mixed poisson regression. *Canadian Journal Of Statistics*, 15 (3): 209–225, 1987.

[26] Scott J. Long, J. Scott Long, and Jeremy Freese. *Regression models for categorical dependent variables using stata*. Stata press, 2006.

[27] Julie M. Rickard. Factors influencing long-distance rail passenger trip rates in great britain. *Journal of Transport Economics and policy*, pages 209–233, 1988.

[28] Martin Ridout, John Hinde, and Clarice G. B. DeméAtrio. A score test for testing a zero- inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics.*, 57 (1): 219–223, 2001.

[29] R. A. Rigby, D. M. Stasinopoulos, and C. Akantziliotou. A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution. *Computational Statistics & Data Analysis*, 53 (2): 381–393, 2008.

[30] J Scott Long. Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, 7, 1997.

[31] George A. F. Seber and Alan J. Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[32] Kimberly F. Sellers, Sharad Borle, and Galit Shmueli. The com-poisson model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28 (2): 104–116, 2012.

[33] H. S. Sichel. On a family of discrete distributions particularly suited to represent long- tailed frequency data. In *Proceedings of the Third Symposium on Mathematical Statistics*. SACSIR, 1971.

[34] Mikis D. Stasinopoulos, Robert A. Rigby, Gillian Z. Heller, Vlasios Voudouris, and Fernanda De Bastiani. *Flexible regression and smoothing: using GAMLSS in R*. Chapman and Hall/CRC, 2017.

[35] Gillian Z. Stein, Walter Zucchini, and June M. Juritz. Parameter estimation for the sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, 82 (399): 938–944, 1987.

[36] Rainer Winkelmann. *Econometric analysis of count data*. Springer Science & Business Media, 2008.

[37] Simon N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2006.

[38] Simon N. Wood. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (3): 495–518, 2008.

[39] Liteng Zha and Yajie Zou. The poisson inverse gaussian (PIG) generalized linear Regression model for analyzing motor vehicle crash data. 2014.

[40] Yajie Zou, Dominique Lord, Yunlong Zhang, and Yichuan Peng. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. *Transportation research record*, 2392 (1): 11–21, 2013.

[41] Yaotian Zou, Dominique Lord, and Srinivas Reddy Geedipally. Over-and under- dispersed count data: Comparing conway-maxwell-poisson and double-poisson distributions. In *Transportation Research Board 91st Annual Meeting, Washington, DC, USA*. Citeseer, 2012.