International Journal of Applied Mathematics & Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 6, Issue 4, Jun - Jul 2017; 19-36
© IASET

GENERALIZED GAMMA REGRESSION MODELS WITH APPLICATION TO CD4 CELL COUNTS DATA OF AIDS PATIENTS

ADEKANMBI D.B

Research Scholar, Department of Statistics, LADOKE Akintola University of Technology,
Ogbomoso, Oyo-State, Nigeria

ABSTRACT

The gamma regression model is a sensible choice of model to analyze responses that are continuous, skewed and take on only positively valued integer outcomes with constant coefficient of variation; of which CD4 cell counts of AIDS patients is a type. CD4 counts may vary by level of formal education, gender, marital status and age of AIDS patients. A detailed theoretical framework of gamma regression was given in this study, and applied to retrospective data set of AIDS patients to determine the relationship between the risk factors and CD4 count of the AIDS patients. Three gamma regression models were considered with three different links for the mean, namely: log, identity and inverse. The choice of link function for the gamma regression is very critical to the accuracy of the model. There appears to be a linear positive effect of sex, level of education, and marital status and a negative effect of age variable on the CD4 counts of the AIDS patients. All the three models showed significant positive impact of sex on the CD4 counts of AIDS patients. The difference between log link and identity link was minimal. The gamma regression model with inverse link function fits poorly, while gamma regression models with an identity link seems to provide a more precise fit to the AIDS data, and was therefore preferred. The result showed that older patients have reduced CD4 cell counts compared to the younger AIDs patients, while males generally have higher CD4 counts than females. However, all the three gamma regression models failed to capture the nature of the observed distribution of the CD4 cell counts. The models were evaluated by the comparison of their deviances and the Akaike Information criterion. Diagnostic evaluation of the models revealed no major problem in the models, except for a few non-influential outlets that were identified. Based on the visual and empirical evidences; the fit of the reciprocal model is therefore preferred for modeling the AIDS data. The results of this study have the potential to be useful for health workers attempting to determine factors associated with improved health of AIDS patients, and for policy makers who are interested in costs and outcomes associated with treatment of AIDS patients.

KEYWORDS: Acquired Immune Deficiency Syndrome, CD4 Cell Count, Gamma Regression, Link Functions, Akaike Information Criterion (AIC)

1.0 INTRODUCTION

AIDS is an acronym for *Acquired Immune Deficiency Syndrome* disease. It is a disease of the immune system that is caused by *Human Immunodeficiency virus*, (HIV). AIDS is the final stage of HIV infection during which time total infectious viruses fight the body system, [6, 12, 29]. HIV could be transmitted from an infected person to an uninfected person by the direct transfer of bodily fluids such as blood products, breast milk, semen and other genital secretions. The identified primary source of transmission is through sexual intercourse, either homosexual or heterosexual, [3, 12, 20, 31]. Presently, there is no vaccine or cure for AIDS. However, an antiretroviral treatment is believed to reduce the risk of infection, but never cures the patient nor alleviates the symptoms, [16, 39]. It was estimated that a total of over 60 million adults were infected in year 2000, with the 63% of them from sub-Sahara Africa, [3, 37, 38]. In Nigeria, the prevalence rate has risen from1. 8% in 1991, 3.8% in 1993, 4.5% in 1995, 5.4% in 1999, 5.8% in 2001, and it has slightly declined to 5% in 2003, and 4.4 % in 2005, [12, 29]. The highest percentage of reported AIDS diagnoses occur in the age group 20-40, which consequently resulted in a reduction in the workforce and loss of productivity, with estimated 4.5 years decrease in life expectancy of this age-group due to the disease, [20, 38]. Women represent the fastest growing group with HIV infection, with the highest rate in sub-Saharan region, [19, 34, 38]. At the end of 2010, it was reported that an estimated 34 million people were living with HIV globally, which includes 3.4 million children less than 15 years, [12].

CD4 cells, also called T-lymphocytes are white blood cells that protect from viral infections by producing antibodies, and are the body's natural defense system against pathogens, infections and illnesses, [5, 13, 17]. Once a person is infected with HIV, the virus attacks and destroys the CD4 cells of the person immune system, which cause the number of cells to decrease over time, [13, 32]. A CD4 cell count is the measurement of the number of blood cells in a cubic millimeter of blood and is the most important laboratory indicator of the health of a person's immune system, [5, 17, 18, 30]. A higher number of CD4 count indicates a stronger immune system to fight HIV and other infections, [42]. It is reasonable to monitor any trends in changes to the CD4 count of an AIDS patient over time. The CD4 count of an uninfected adult who is in good health ranges from 500 cells/mm^3 to 1500 cells/mm^3. People that are HIV positive who have a CD4 cell count over 500 are regarded as being in good health, while those with CD4 count below 200 cells/mm^3 are at significant risk of developing serious illnesses and infections, [5, 17, 18, 29, 30, 40]. Some studies reported sex differential in CD4 cell counts of AIDS patients, indicating that CD4 counts are lower in women compared to men, [13, 16, 19, 21, 32, 35, 36].

The gamma distribution is suitable as a lifetime model, [14, 24]. The gamma distribution can be viewed as a generalization of the exponential distribution with mean $1/\lambda$, which represents the waiting time until the first event occur, where the events are generated by a Poisson process with mean λ. The gamma random variable therefore represents the waiting time until the nth event to occur, [24, 41]. The gamma distribution hangs on the assumption that all waiting times are complete at the end of the study, so that censoring is not allowed. The gamma regression model is applicable if the response has a gamma distribution, [11]. When the distribution is positively skewed and has variance increasing with mean, then gamma distribution is appropriate, [28]. The assumption of constant coefficient of variation (CV) in gamma regression can be verified by grouping the data set into intervals based on the value of estimated mean, and estimate the CV in each interval. Plots of CV against mean should reveal any systematic departure from constancy, [28]. The model has a wide range of application in the medical field, [23, 24].

This study is focused on gamma regression modeling of AIDS cases in Nigeria, to determine factors that are significant in improving the CD4 cell counts of AIDS patients; and to also determine the most suitable link function for the gamma regression for the data set, among the existing link functions for gamma regression models. After the introduction, this study is sectionalized into six sections. In section 2, a detailed description of the data used in this study is presented. In section 3, a full detail of the theory of gamma distribution and gamma regression is presented. A measure of model selection is summarized in section 4. Application of the gamma regression based on the AIDS data is discussed in section 5. Results of application cases base on the three different link functions are also presented and discussed. Finally, in section 6, various issues arising from the study are discussed.

2.0 DATA

Data on AIDS patients used in this study were extracted from the records of the Obafemi Awolowo Teaching Hospitals, Nigeria. In the data set, there were 407 AIDS cases, and the demographic and clinical variables recorded for each patient were gender, level of formal education, marital status, age of patients at death, and the CD4 cell count of each of the HIV patients. The response is therefore the CD4 cell counts of the HIV patients. The response by its nature is always continuous, non-negative positively skewed, and does not exhibit the same variability at all levels.

3.0 METHODOLOGY

Generalised linear models (GLM) are extension of classical linear regression usually formulated with a purpose of predicting the outcome of a response as a function of some linear combination of a set of predictors or explanatory variables, [10, 11, 25, 26, 28]. In order to formulate a GLM, a link function is required, which relates the linear predictors to the predicted mean of response; and also required is a function defining the error probability distribution around the mean. Examples of distributions that belong to the exponential family are Normal, Binomial, Poisson, Gamma, and so on.

Given predictors X_i , the mean of response variable can be expressed in terms of the linear combination of predictors, such that:

$$\eta_i = X_i^T \beta \tag{1}$$

$$\eta_i = g(\mu_i) = x_i^T \beta \tag{2}$$

$$= \beta_0 + \beta_1 X_{i1} + \ldots\ldots + \beta_k X_{ik}$$

Where

η_i : is the linear predictor.

$g(.)$: is the link function.

$\mu_i = E(Y_i | X_i)$

The link function is invertible, so that

$$\begin{aligned} \mu_i &= g^{-1}(\eta_i) = g^{-1}(x_i^T \beta) \\ &= g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots + \beta_k X_{ik}) \end{aligned} \tag{3}$$

The three common link functions for gamma regression are, [10, 11, 28]:

(i) The log link: $g(\mu) = \log(\mu)$

(ii) The identity link: $g(\mu) = \mu$

(iii) The inverse link: $g(\mu) = {}^{1}\!/_{\mu}$

The log link should be used when the effect of the predictors is suspected to be multiples of the mean. So that the effect of the predictors x_j on μ is multiplicative and not additive. The log link is considered because CD4 counts, which is the response variable must be positive and log link yields positive values. The identity link is adequate for modeling variance components which have chi-square distribution. Gamma regression with identity link function is also adequate when the effect of each predictor is considered additive on the original scale. Since $-\infty < \eta < \infty$, the inverse link does not guarantee $\mu > 0$, and this could cause problems which might require restrictions on β, [11]. Essentially, the choice of link function is rather subjective, [11].

3.1 Re-Parameterization of Gamma Distribution

The two-parameter gamma distribution has a density function of the form:

$$f(y;\lambda,\alpha) = \frac{\lambda}{\Gamma(\alpha)}(\lambda y)^{\alpha-1}\exp(-\lambda y) \qquad y > 0, \text{and } \lambda,\ \alpha > 0 \tag{4}$$

Where

α : shape parameter

λ : scale parameter of the distribution.

Then $Y \sim G(\alpha,\lambda)$

The properties of a Gamma distribution are therefore

$$E(Y_i) = \frac{\alpha}{\lambda} = \mu_i \tag{5}$$

$$\mathrm{Var}(Y_i) = \frac{\alpha}{\lambda^2} = \mu^2\left(\frac{1}{\alpha}\right) = \sigma^2(E(Y_i))^2 \tag{6}$$

According to [7, 8], the gamma density function in (4) can be re-parameterized in terms of the mean and shape parameters by setting $\mu = \frac{\alpha}{\lambda}$ then $\lambda = \frac{\alpha}{\mu}$. So that (6) becomes

$$f_Y(y|\mu,\alpha) = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)\left(\frac{\alpha}{\mu}y\right)^{\alpha-1}\exp\left(-{}^{ay}\!\big/_{\mu}\right) \tag{7}$$

$$f_Y(y|\mu,\alpha)dy = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha y}{\mu}\right)^{\alpha}\exp\left(-{}^{ay}\!\big/_{\mu}\right)\frac{1}{y}dy \tag{8}$$

Since $d(\mathrm{In}(y)) = {}^{1}\!\big/_{y}\,dy$

(8) Is referred to as $G(\mu,\alpha)$, which implies that y follows a gamma distribution with mean μ and α as a shape parameter.

3.2 Gamma Distribution as a Member of the Exponential Family

Gamma density having the density function of the form (7), so that $Y \sim G(\mu,\alpha)$. Rearranging the density according to the exponential form, then

$$f_y(y) = \exp\left\{ \frac{y\left(-\frac{1}{\mu}\right) - \log(\mu)}{\frac{1}{\alpha}} + \alpha\log\alpha + (\alpha-1)\log y - \log\Gamma(\alpha) \right\} \quad \textbf{(9)}$$

So that

$$\theta = -\frac{1}{\mu} \text{ and } a(\varphi) = \varphi,\ \varphi = \frac{1}{\alpha}$$

And

$$b(\theta) = \log(\mu) = \log(-1/\theta) = -\log(-\theta)$$

Then

$$b'(\theta) = -\frac{(-1)}{-\theta} = -1/\theta = \mu$$

$$b''(\theta) = \frac{1}{\theta^2} = \mu^2$$

The canonical parameter is $-\frac{1}{\mu}$, so that the canonical link function is therefore $g(\mu) = -\frac{1}{\mu}$. The dispersion parameter is $\varphi = \frac{1}{\alpha}$. For any gamma distribution, $\mu > 0$, since the distribution is only defined for $y > 0$, [24, 26]. Restriction must therefore be placed on the vector β to ensure the expected value is positive.

3.3 Model Formulation of Gamma Regression

Let $y_i \sim G(\mu_i, \alpha)$ $i = 1,2,.....n$ be independent random variables, then the gamma regression model is given as:

$$\eta_i = g(\mu_i) = x_i^t\beta \quad \textbf{(10)}$$

Where

$x_i = (x_{i1},..........x_{ik})'$: the vector of k covariates. Usually $x_{i1} = 1, \text{for all} i$, so that model has a mean intercept.

$\beta = (\beta_1,..........\beta_k)'$: is the vector of unknown regression parameters, $(k < n)$.

η_i : is a linear predictor.

g: the link function

[7] Proposed that the shape parameter is not constant through the observations and could be modeled following regression structure. So that $y_i \sim G(\mu_i, \alpha)$ $i = 1,2,.....n$ are independent random variables with gamma distribution. The mean and shape parameters follow a regression structure given as:

$$\eta_{1i} = g(\mu_i) = x_i'\beta \tag{11}$$

$$\eta_{2i} = h(\alpha_i) = z_i'\gamma \tag{12}$$

Where

$\beta = (\beta_1,...........\beta_k)'$ and $\gamma = (\gamma_1,........\gamma_j)'$ with k+j <n , are vectors of regression parameters which are related to the mean and dispersion.

G: is the mean link function.

H: is the shape link function.

η_{1i} and η_{2i} : are the linear predictors.

x_i and z_i : are the mean and shape explanatory variables for the i^{th} observation.

Gamma regression has been found adequate in modeling data in which the coefficient of variation is constant, [9, 26, 28].

$$\frac{\sqrt{\operatorname{Var}(y_i)}}{E(y_i)} = \frac{\sqrt{\alpha/\lambda_i^2}}{\alpha/\lambda_i} = \frac{1}{\sqrt{\alpha}} \tag{13}$$

3.4 Parameter Estimation of Gamma Regression Model

Generalised linear models can be fitted to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients, [10, 28]. Given that α is known in $Y \sim G(\mu, \alpha)$, under the re-parameterisation of gamma density function given in (7), the likelihood function of the gamma regression models of (11) and (12) is given by:

$$L = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \left(\frac{\alpha_i}{\mu_i}\right)^{\alpha_i} y_i^{\alpha_i - 1} \exp\left(-\frac{\alpha_i y_i}{\mu_i}\right) \tag{14}$$

The log-likelihood function is therefore,

$$l = \sum_{i=1}^{n} \left\{ -\log[\Gamma(\alpha_i)] + \alpha_i \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - \log(y_i) - \left(\frac{\alpha_i}{\mu_i}\right) y_i \right\} \tag{15}$$

Assuming the systematic components $\mu_i = x_i'\beta$ and $\alpha_i = \exp(z_i'\gamma)$, the components of the score function are therefore: $\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} -\frac{\alpha_i}{\mu_i}\left(1-\frac{y_i}{\mu_i}\right)x_{ij} \qquad j = 1,2,........p$ **(16)**

$$\frac{\partial l}{\partial \gamma_k} = \sum_{i=1}^{n} -\alpha_i\left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right]z_{ik} \qquad k = 1,2,........r \tag{17}$$

These derivatives can then be assembled to give a vector of efficient scores, $u(\beta)$.

$$u(\beta) = \left(\frac{\partial l}{\partial \beta_j}\ \frac{\partial l}{\partial \gamma_k}\right)^T \tag{18}$$

$$u(\beta) = X'W^{(k)}Y$$

Where

X: is the design matrix.

$W^{(k)}$: a diagonal matrix with elements ω_i.

Y: a vector with elements y_i^*

The components of the Hessian matrix are:

$$\frac{\partial^2 l}{\partial \beta_k \beta_j} = \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2}\left(1-\frac{2y_i}{\mu_i}\right)x_{ij}x_{ik} \qquad j,k = 1,2,........p \tag{19}$$

$$\frac{\partial^2 l}{\partial \gamma_k \beta_j} = \sum_{i=1}^{n} -\alpha_i\left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right]z_{ik} \qquad k = 1,2,........r \tag{20}$$

$$\frac{\partial^2 l}{\partial \gamma_k \gamma_j} = \sum_{i=1}^{n} -\alpha_i\left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right]z_{ik} \qquad k = 1,2,........r \tag{21}$$

The expectations on both sides of the equation (19), (20) and (21) yield

$$-E\left(\frac{\partial^2 l}{\partial \beta_k \beta_j}\right) = \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2} x_{ji}x_{ik} \qquad j,k = 1,2,....p \tag{22}$$

$$-E\left(\frac{\partial^2 l}{\partial \gamma_k \beta_j}\right) = 0 \quad k = 1,2,....r \text{ and } j = 1,2......p \tag{23}$$

$$-E\left(\frac{\partial^2 l}{\partial\beta_k\beta_j}\right)=\sum_{i=1}^{n}\alpha_i^2\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i)-\frac{1}{\alpha_i}\right]z_{ij}z_{ik} \qquad j,k=1,2,...r \tag{24}$$

The Fisher information matrix is a block diagonal matrix, with one of the blocks corresponding to the mean regression parameters, while the other corresponds to the shape regression parameters. It follows that the Fisher information is

$$I(\beta)=\begin{bmatrix}-E\left(\frac{\partial^2 l}{\partial\beta_k\beta_j}\right) & -E\left(\frac{\partial^2 l}{\partial\gamma_k\beta_k}\right)\\ -E\left(\frac{\partial^2 l}{\partial\gamma_k\beta_k}\right) & -E\left(\frac{\partial^2 l}{\partial\beta_k\beta_j}\right)\end{bmatrix} \tag{25}$$

$$I(\beta)=\begin{bmatrix}\sum_{i=1}^{n}\frac{\alpha_i}{\mu_i^2}x_{ji}x_{ki} & 0\\ 0 & \sum_{i=1}^{n}\alpha_i^2\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i)-\frac{1}{\alpha_i}\right]z_{ij}z_{ki}\end{bmatrix} \tag{26}$$

In matrix form, this is

$$I(\beta)=X'W^{(k)}X \tag{27}$$

An iterative algorithm to obtain the maximum likelihood estimates of the gamma regression parameters has been proposed, [8, 28]. Given the parameters values $\left(\beta^{(k)},\gamma^{(k)}\right)'$, the mean vectsincef the regression parameter is updated from:

$$\beta^{(k+1)}=\beta^{(k)}+I^{-1}\left(\beta^{(k)}\right)u\left(\beta^{(k)}\right) \tag{28}$$

$$\beta^{(k+1)}=\beta^{(k)}+\left(X'W_1^{(k)}X\right)^{-1}X'W_1^{(k)}Y \tag{29}$$

Where

$W_1^{(k)}$: is a matrix with diagonal elements of $w_i^{(k)}=\left(\mu_i^2\right)^{(k)}\big/\alpha_i^{(k)}$

Given initial values $\left(\beta^{(k+1)},\gamma^{(k)}\right)'$ the shape parameters $\gamma^{(k+1)}$ could be updated from

$$\gamma^{(k+1)}=\left(Z'W_2^{(k)}Z\right)^{-1}X'W_2^{(k)}Y \tag{30}$$

Where

$W_2^{(k)}$: is a diagonal matrix with elements $w_i^{(k)}=1\big/d_i^{(k)}$

$$d_i = \alpha_i^{-2}\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i) - \frac{1}{\alpha_i}\right]^{-1}$$

$$\dot{y}_i = \eta_{2i} - \frac{1}{\alpha_i}\left[\frac{\partial^2}{\partial\alpha^2}\log\Gamma(\alpha_i) - \frac{1}{\alpha_i}\right]^{-1}\left[\frac{\partial^2}{\partial\alpha^2}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right]$$

Given the initial values of β then $\beta^{(k+1)}$ could be obtained from (29); and given the initial values of β and γ then $\gamma^{(k+1)}$ could be obtained from (30). The process should be repeated until convergence is reached.

3.5 Gamma Regression Residual Analysis

The main purpose of residual analysis in a generalized linear model is to identify model mis-specification or outliers. Since a well fitting model is a prerequisite for reliable inferences, it is therefore necessary to inspect the quality of fit provided by the gamma regression model, after fitting the model to a set of data.

3.5.1 Residual Deviance

The lack of fit in gamma regression is measured by deviance. It provides a measure of the discrepancy between the model and the data. A large value of deviance indicates a poor fit, while a small value of deviance indicates a good fit, [25]. Given that

$$Y \sim G(\mu_i, \alpha) \text{ Independent } \mu_i = \exp(x_i^t\beta) \quad \textbf{(31)}$$

$$l(\mu, \alpha, y) = \sum_{i=1}^{n}\left\{\alpha\left[\frac{-y_i}{\mu_i} - \log(\mu_i)\right] - \log(\Gamma(\alpha)) + \alpha\log(\alpha y_i) - \log(y_i)\right\} \quad \textbf{(32)}$$

$$l(y, \alpha, y) = \sum_{i=1}^{n}\{\alpha[-1 - \log(y_i)] - \log(\Gamma(\alpha)) + \alpha\log(\alpha y_i) - \log(y_i)\} \quad \textbf{(33)}$$

$$-2(l(\hat{\mu}, \alpha, y) - l(y, \alpha, y)) = 2\left[\sum_{i=1}^{n}\alpha\left(-1 - \log(y_i) + \frac{y_i}{\mu_i} + \log(\hat{\mu}_i)\right)\right] \quad \textbf{(34)}$$

$$= -2\sum_{i=1}^{n}\alpha\left[\log\left(\frac{y_i}{\hat{\mu}_i}\right)\right] - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \quad \textbf{(35)}$$

Then the deviance residual is therefore

$$D(y, \hat{\mu}) = -2\sum_{i=1}^{n}\left[\log\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right] \quad \textbf{(36)}$$

$$D(y, \hat{\mu}) \sim \varphi\chi^2_{n-p}$$

Where $\varphi = 1/\alpha$ and $\hat{\mu}_i = g^{-1}(x_i'\beta)$

Reject model (10) at 0.05 significance level if $\frac{D(y,\hat{\mu})}{\hat{\varphi}} > \chi^2_{n-p,1-0.05}$

So that a calculated deviance that exceeds the upper $100(1-\alpha)$ percent point of the χ^2_{n-p} distribution indicates a poor fit to the data at the 100α percent significant level.

3.5.2 Standardised Residual

The standardized residual identify any observations that give a disproportionately large contribution to the deviance. For gamma regression, the standardized residual is defined as follows:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{Va\hat{r}(y_i)}} \tag{37}$$

Where

$$Var(y_i) = \frac{\hat{\mu}^2}{\hat{\alpha}_i} \tag{38}$$

4.0 MODEL SELECTION: AKAIKE INFORMATION CRITERION (AIC)

For competing generalized linear regression models, the best model can be determined by taking into account the number of parameters, using the Akaike Information Criterion (AIC). AIC measure describes the tradeoff between bias and variance in model construction, or between accuracy and complexity of the model, [11, 41]. It is a measure of fit that penalizes for the number of parameters.

$$\begin{aligned} AIC &= D + 2p \\ &= -2l_{mod} + 2p \end{aligned} \tag{39}$$

Where

D: deviance statistic

p: number of parameters in the linear predictor of the model under consideration.

l_{mod}: Log-likelihood of the fitted model.

When models differ in terms of their link functions or predictors, comparing AIC statistic is straightforward. However the same data should be fitted by models that are being compared. Smaller values of AIC indicate better fit, and thus the AIC can be used to compare models, whether nested or not, [25].

5.0 RESULTS OF GAMMA REGRESSION ANALYSIS OF AIDS DATA

Figure 1 is the scatter plot matrix of the CD4 cell counts against the predictors which shows clear evidence of a continuous positively skewed curve; and the variance increases rapidly with the mean, especially with *age*. This therefore suggests that a gamma regression might be an appropriate model for the data. In fact, since the response variable, CD4 cell counts are positively skewed continuous variable, suggesting that gamma regression should be an appropriate model structure for the data.

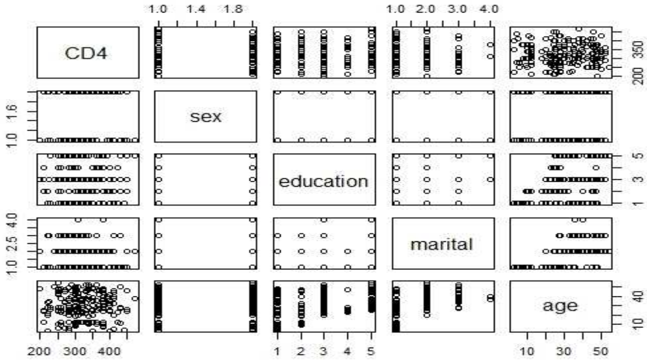

Figure 1: Scatter Plot Matrix of the AIDS Variables

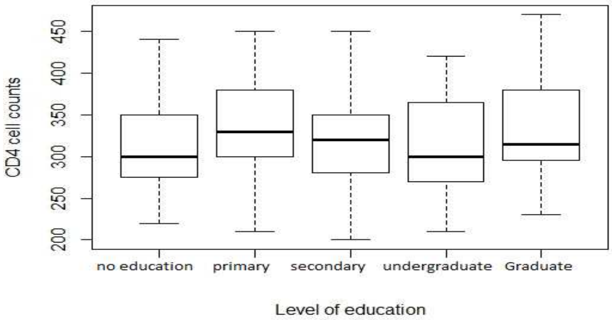

Figure 2: Box Plot for CD4 Counts Data, by Level of Education

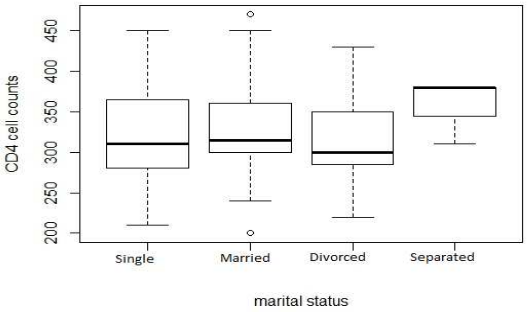

Figure 3: Box Plot of CD4 Counts Data by Marital Status

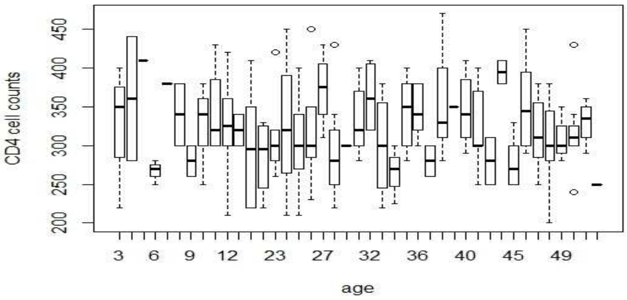

Figure 4: Box Plot of CD4 Counts Data by Age of Patients

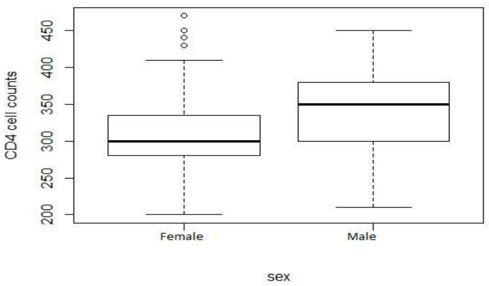

Figure 5: Box Plot of CD4 Counts Data by Sex

Table 1: MLE of the Parameter, Standard Errors, and P-Values for Gamma Regression Models Fitted to the AIDS Data

Effect	Method	Estimate	SE	P-Value
Intercept	Log	5.7207	0.0398	2e-16***
	identity	304.7857	12.5586	2e-16***
	inverse	0.0033	0.00013	2e-16***
Sex	Log	0.0741	0.0250	0.0034**
	identity	24.0057	8.0886	0.0034**
	inverse	-0.0002	0.0007	0.0036**
Age	Log	-0.0011	0.0015	0.4962
	identity	-0.3512	0.4932	0.4773
	inverse	0.000003	0.000005	0.5150
Level of Education				
Primary Education	Log	0.0605	0.0374	0.1072
	identity	20.0782	12.0473	0.0972
	inverse	0.00018	0.00012	0.1187
Secondary Education	Log	0.0225	0.0365	0.5371
	identity	7.9535	11.5383	0.4914
	inverse	0.000006	0.00011	0.5827
Tertiary	Log	0.0145	0.0522	0.7811
	identity	5.4073	16.3549	0.7413
	inverse	0.00004	0.00017	0.8187
Graduate	Log	0.0408	0.0398	0.3071
	identity	13.3783	12.7832	0.2966
	inverse	0.00012	0.00011	0.3204
Marital Status				
Married	Log	0.0522	0.0409	0.2028
	identity	16.9143	13.0257	0.1956
	inverse	0.00016	0.00013	0.2097
Divorced	Log	0.0317	0.0497	0.5248
	identity	10.4324	15.7862	0.5095
	inverse	0.00009	0.00016	0.5381
Separated	Log	0.1160	0.1057	0.2741
	identity	38.2207	37.1342	0.3046
	inverse	0.00035	0.00030	0.2430

Significant Codes: $0^{(***)}$ $0.001^{(**)}$

Table 2: Residual Deviance and AIC for the Gamma Models

	Residual Deviance	Residual D.F	AIC
Model 1	6.1044	197	2265.4
Model 2	6.1014	197	2265.3
Model 3	6.1075	197	2265.5

The degree of symmetry of the predictors can be judged from the compound box plots in Figures 1. Figures 2 shows the box plots for the CD4 counts by level of education of patients, and are all positively skewed with no outliers, except for the secondary group which is negatively skewed. Figure 3 is the box plot of CD4 counts by marital status of AIDS patients. Noticeable from the box plot is that the CD4 counts tend to be a little higher for single and married compared with the divorced marital status group. There are obvious outliers in the CD4 counts for married, and each marital status, distribution shows evidence of right skewed, except the separated status that is so skewed that the minimum, median and third quartile are all equal to this marital group. Figure 4 shows the compound box plot for CD4 counts by age of AIDS patients. For all the ages, the distribution is positively skewed, except for ages 26, 32, 36, 38 and 49, and outliers can be seen at ages 23, 26, 28 and 50. Figure 5 which is the box plot for CD4 counts by sex shows that the distribution of CD4 counts for female is positively skewed, with a few outliers while that of males is negatively skewed. Males appear to have higher CD4 counts than females across the time period considered.

In the analysis, the response variable is the CD4 counts of AIDS patients, while level of formal education of patients, marital status, and age sex is the predictors. Table 1 provides the result of fitting the three gamma regression models to the AIDS data, giving the estimates of the regression parameters, their standard errors, and their corresponding p-values for the three gamma regression models. Model1 contain information for the log link fit, model2 contain information for the identity link, while model3 contain information for the inverse link fit. However, all the three gamma regression models failed to capture the nature of the observed distribution of the CD4 counts. The estimated coefficients of the three models are noticeably different from each other. The three models can be evaluated based on their residual deviance and AIC-based model selection. In fact the three models yield similar results with little difference in their residual deviance and their AIC values. As shown in Table 2, the residual deviance of model1 is 6.1044 on 197 d.f and AIC of 2265.4, model2 yields residual deviance of 6.1014 on 197 d.f with AIC of 2265.3, while the residual deviance of model3 is 6.1075 with AIC of 2265.5. The model with smaller AIC is preferred, so that model2 with identity link fit is preferred above the inverse and the log link fit. Among all the predictors considered, the three gamma models consistently indicate that only *sex* has significant impact on the CD4 counts of the AIDS patients. For model1, there are exp (0.0741) = 1.0769 times as much CD4 for males relative to females with other variables held fixed. The comparable figure for *sex* in the identity link gamma model is 24.01. Age, level of education of patients and their marital status are not significant in predicting the CD4 counts of AIDS patients.

The usual diagnostics were performed for the three models, as shown in Figures 6 respectively. The plots of the jackknife deviance residuals against the fitted values shown on the top left panel in the Figures 7, revealed no appearance of systematic trend. The plot on the top right of Figures 8 is the normal QQ plot of the standardized deviance residuals, and shows that the standardized residuals are normally distributed. On the left side of the bottom panel is the plot of the Cook statistics against the standardized leverages, which identifies three observations to the right of the vertical line as observations with likely high leverage compared to the variance of the raw residual at that point. The last plot on the right side of the bottom panel shows the cook statistics plotted against case number, which also revealed that some observations are influential on the models.

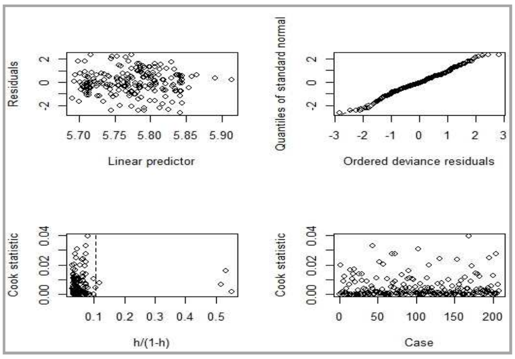

Figure 6: Residual Plot for Model 1

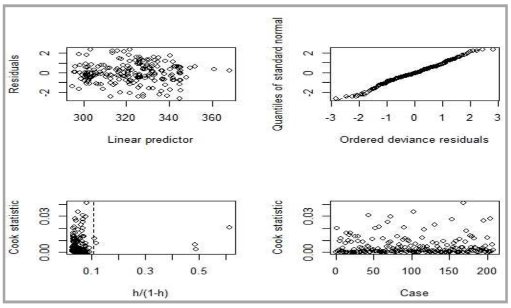

Figure 7: Residual Plot for Model 2

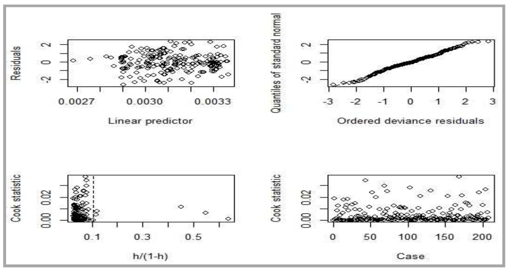

Figure 8: Residual Plot for Model 3

6.0 DISCUSSIONS

Generally, GLMs are the regression modeling of non-normal data. Gamma regression is a reasonable model for the AIDS data since the response variable, CD4 cell counts is a continuous, positive outcome with constant coefficient of variation. Three gamma models were considered for the AIDS data. Before fitting the gamma models to the AIDS data, outliers were detected. Regardless of the distribution of CD4 counts, for the inverse link, the estimates of the parameters of the model were very poor. This link was deemed a poor choice for the data, but the link could adjust in a better way for different data set. Comparison of the fit indices of the three models revealed that the gamma regression model with an identity link provided a better fit compared to the gamma regression model with log-link, and with inverse link.

The gamma regression model is not without some limitations. Generally, GLMs tend to be inefficient in the presence of heavier tails. Gamma model is not consistent when the data are heteroscedastic in nature. Specification of an adequate gamma regression model for a data set requires a careful examination of the characteristics of the data. Further studies of AIDS patients with large cases and most predictive variables are therefore recommended to confirm the results. Despite the limitations, the results are valuable in understanding the role of CD4 cell counts in response to the risk factors.

CONCLUSIONS

The results of the gamma regression analysis of AIDS patients have the potential to be useful to medical doctors and health workers, attempting to determine factors that could help in improving the CD4 counts of AIDS patients, and thereby achieve improved health of the patients. The result of this study should be valuable to policy makers in evaluating the cost effectiveness of factors that could impact positively on the health of AIDS patients.

REFERENCES

1. Agresti, A. An introduction to categorical data analysis. John Wiley and Sons, Inc., 1996
2. Alioum A, Leroy V, Commenges D, Dabis F, Salamon R. The effect of gender, age, transmission category and antiretroviral therapy on the progression of human immunodeficiency virus infection using multistate Markov models. Epidemiology 1998; 9:605-612
3. Bendavid E, Bharttacharya J. The president's emergency plan for AIDS relief in Africa, an evaluation of outcomes. Annals of International Medicine 2009; 150:688-695.
4. Cakmakyapan, S., and Goktas, A. A comparison of binary and probit models with a simulation study. Journal of Social and Economic Statistics, 2013 ;2 (1): 1-17.
5. CD4 count. Available at https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/. Accessed April 5, 2016.
6. Centres for Control and Prevention disease HIV/AIDS-United States, 1981-2000. MMWR Morb Mortal Wkly Rep 2001; 50: 430-434.
7. Cepeda, E and Gammerman, D. Bayesian methodology for modeling parameters in the two parameter exponential family. ESTADISTICA, 2005; 57 (168): 93-105.
8. Cepeda-Cuervo, E. Modelagem de variabilidade em modelos lineeares generalizados. Unpublished Ph.D thesis, Mathematics Institute, Universidade Federal Rio de Janeiro, 2001.
9. Christensen, R. Log-linear models. New-York: Sringer-Verlag, 1990.
10. Dobson, A.J. An introduction to generalized linear models. 2010. CRC Press.
11. Faraway, J.J. Extending the linear model with R, generalized linear, mixed effects and non-parametric regression models. New-York: Chapman and Hall/CRC, 2006.
12. Global HIV/AIDS Response: Epidemic update and health sector progress towards universal access. Progress report 2011 by World Health Organization (WHO), Joint United Nations program on HIV/AIDS (UNAIDS), and UNICEF.
13. Gupta SB, Gilbert RL, Brady AR, Livingstone SJ, Evans BG. CD4 cell counts in adults with newly diagnosed HIV infection: results of surveillance in England and Wales, 1990-1998. CD4 Surveillance Scheme Advisory Group, AIDS 2000; 14:853-861.
14. Gupta, S.S and Groll, P.A. Gamma distribution in acceptance sampling based on life tests. Journal of America Statistical Association 1961, 56: 942-970.

15. Herna'n MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000; 11: 561-570.

16. Hisada M, Stuver SO, Okayama A, Mueller NE. Gender difference in skin reactivity to purified protein derivative among carriers of HTLV-1 in Japan. J Acquir Immue Defic Syndr 1999; 22:302-307

17. HIV and AIDS Information: CD4 cell counts. Available at http://www.aidsmap.com/CD4-cell-counts/page/1044596/. Accessed April 5, 2016.

18. HIV treatment: Understanding your lab work (Blood Tests) -POZ. Available at https://www.poz.com/basics/hiv-basics/understanding-lab-work. Accessed April 5, 2016.

19. Institute of Medicine. Exploring the biological contributions to human health: does sex matter? 2001: http://www.nap.edu/catalog/10028.html.

20. Ivers L, Kendrick D, Doucette K. Efficacy of antiretroviral therapy programs in resource-poor settings, a meta-analysis of the published literature. Clin Infect Dis 2005;41:217-224.

21. Junghans C, Ledergerber B, Chan P, Weber R, Egger M. Sex differences in HIV-1 viral load and progression to AIDS. Lancet 1999; 353:589.

22. Justice AC, Feinstein AR, Wells CK. A new prognostic staging system for the acquired immunodeficiency syndrome. N Engl J Med 1989; 320:1388-1393.

23. Kalbfleisch, J.D. The statistical analysis of failure time data. 2^{ed}. New-York: Wiley-Interscience, 2002.

24. Krishnamoorthy, K. Handbook of statistical distributions with applications. Florida: Chapman and Hall/CRC, 2006.

25. Krzanowski, W.J. An introduction to statistical modeling. New-York: Oxford University Press, 1998.

26. Lawal, B. Categorical data analysis with SAS and SPSS Applications. New-Jersey: Lawrence Erlbaum Associates, 2003.

27. Level and pattern of HIV-1-RNA viral load over age: differences between girls and boys? AIDS 2002; 16:97-104.

28. McCullagh, P. and Nelder, J. Generalized linear models (2^{ed}). London: Chapman and Hall, 1989.

29. National Department of Health. National antiretroviral treatment guidelines, 1st ed. 2004. Available at www.doh.gov.za/docs/factsheets/guidelines/artguide04-f.html. Accessed, January 19, 2016.

30. O'Brien WA, Hartigan PM, Daar ES, Simberkoff MS, Hamilton JD. Changes in plasma HIV RNA levels and CD4^{+} lymphocyte counts predict both response to antiretroviral therapy and therapeutic failure. VA Cooperative Study group on AIDS. Ann Intern Med 1997; 126: 939-945.

31. Peterson P, Sharp B, Gekker G, Portoghese P, Sannerud K, Balfour HJ. Morphine promotes the growth of HIV-1 in human peripheral blood mononuclear cell cocultures. AIDS 1990; 4: 869-873.

32. Portales P, Clot J, Corbeau P. Sex differences in HIV-1 viral load due to sex difference in CCR5 expression. Ann Intern Med 2001; 134:81-82.

33. Prabhala RH, Wira CR. Influence of estrous cycle and estradiol on mitogenic responses of splenic T-andB-lymphocytes. Adv Exp Med Biol 1995; 371A:379-381.

34. Rosenberg PS, Biggar RJ. Trends in HIV incidence among young adults in the United States. JAMA 1998; 279:1894-1899.

35. Sayles JN, Wong MD, Cunningham WE. The inability to take medications openly at home: Does it help explain gender disparities in HAART use? J Women's Health 2006; 15:173-181.

36. Sterling T, Lyles C, Vlahov D, Astemborski J, Margolick J, Quinn T. Sex differences in longitudinal human immunodeficiency virus type 1 RNA levels among seroconverters. J Infect Dis 1999; 180:666-672.

37. The ART-LINC collaboration and ART_CC groups. Mortality of HIV-1-infected patients in the first year of antiretroviral therapy, comparison between low-income and high income countries. Lancet 2006; 367:817-824.

38. UNAIDS. Global report. UNAIDS Report on the global AIDS epidemic. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS (UNAIDS) 2010.

39. Vassiliadou N, Tucker L, Anderson D. Progesterone-induced inhibition of chemokine receptor expression on peripheral blood mononuclear cells correlates with reduced HIV-1 infect ability in viro. J Immunol 1999; 162:7510-7518.

40. Verhofstede C, Reniers S, Van Wanzeele F, Plum J. Evaluation of proviral copy number and plasma RNA level as early indicators of progression of HIV-1 infection: correlation with virological and immunological markers of disease. AIDS 1994; 8:1421-1427.

41. Wackerly, D.D., Mendenhall III, W., and Scheaffer, R.L. Mathematical statistics with applications. Pacific Grove: Wad worth Group, Duxbury and Brooks/Cole, 2002.

42. Weinstock H, Sweeney S, Satten GA, Gwinn M. HIV seroincidence and risk factors among patients repeatedly tested for HIV attending sexually transmitted disease clinics in the United States, 1991 to 1996. STD Clinic HIV Seroincidence Study Group. J Acquir Immune Defic Syndr Hum Retrovirol 1998; 19:506-512.