Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 420-436
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Inference Procedures on the Generalized Poisson Distribution from Multiple Samples: Comparisons with Nonparametric Models for Analysis of Covariance (ANCOVA) of Count Data

Maha Al-Eid[1], Mohamed M. Shoukri[2*]

[1]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario Canada
Email: mmshoukr@uwo.ca, *Shoukri.mohamed@gmail.com

How to cite this paper: Al-Eid, M. and Shoukri, M.M. (2021) Inference Procedures on the Generalized Poisson Distribution from Multiple Samples: Comparisons with Nonparametric Models for Analysis of Covariance (ANCOVA) of Count Data. *Open Journal of Statistics*, **11**, 420-436.
https://doi.org/10.4236/ojs.2021.113026

Received: May 18, 2021
Accepted: June 22, 2021
Published: June 25, 2021

Abstract

Count data that exhibit over dispersion (variance of counts is larger than its mean) are commonly analyzed using discrete distributions such as negative binomial, Poisson inverse Gaussian and other models. The Poisson is characterized by the equality of mean and variance whereas the Negative Binomial and the Poisson inverse Gaussian have variance larger than the mean and therefore are more appropriate to model over-dispersed count data. As an alternative to these two models, we shall use the generalized Poisson distribution for group comparisons in the presence of multiple covariates. This problem is known as the ANCOVA and is solved for continuous data. Our objectives were to develop ANCOVA using the generalized Poisson distribution, and compare its goodness of fit to that of the nonparametric Generalized Additive Models. We used real life data to show that the model performs quite satisfactorily when compared to the nonparametric Generalized Additive Models.

Keywords

Count Regression, Over Dispersion, Generalized Linear Models, Analysis of Covariance, Generalized Additive Models

1. Introduction

The Poisson distribution is commonly used to model count data. However, a re-

striction of this distribution is that the response variable must have a mean equal to the variance. This restriction does not often hold true for many biological and epidemiological data. In many applications the variance can be much larger than the mean, a phenomenon known as "over dispersion". This over dispersion may occur due to population heterogeneity, or presence of outliers in the data [1]. An analysis of data with overly dispersed counts can lead to the underestimation of parameter standard error, if overdispersion is ignored. A review of the issue of overdispersion in both binary and count data was reviewed by Hinde and Demetrio [2], and in a more recent review by Hayat and Higgins [3]. Diagnosing and accounting for overdispersion is not a simple issue and should be appropriately dealt with to avoid bias in interpreting the results.

The Negative-Binomial (NB) distribution has been used as an alternative to the Poisson distribution for modeling data that exhibit overdispersion. The NB has two parameters and a variance that is a quadratic function of the mean. NB model has been the model of choice for the analysis of overly dispersed count data. The NB regression was reviewed by Hinde and Demetrio [2]. Joe and Zhu [4] drew a comparison between the NB and a mixture-based generalization of the Poisson distribution.

In this paper we discuss several inferential statistical issues related to a modified form of the Generalized Poisson Distribution (GPD). The GPD distribution was introduced to the statistical literature by Consul and Jain [5] and a detailed account of its properties was given by Consul [6]. The distribution has two parameters, and a variance that is cubic function of the mean. The distribution has been used to analyze data in the fields of genetics [7] as a queuing model [8] [9] [10] and genomics [11]. The modified form of the GPD, which we shall call "Modified Poisson Distribution" (MGPD) was first discussed in [12]. The modification was a double parametric transformation on the original parameters of the GPD. The main purpose of the transformation was to achieve parameters orthogonality [13] and make the MGPD a member of the class of "Generalized Linear Models" [14]. Recently Shoukri and Al-Eid investigated several inference procedures in the two samples situation [15].

This paper has three-fold objectives. In Section 2, we present the model. We then, assume that we have k independent samples and we demonstrate how to construct statistical testing procedures on the dispersion parameters. Specifically, we first validate the hypothesis of homogeneity of dispersion parameters, thereafter we test the significance of the common dispersion parameter. In Section 3 we test the hypothesis of equality of k-means in the presence of overdispersion. When covariates are measured, testing the equality of group means is therefore equivalent to the Analysis of covariance (ANCOVA) in the presence of overdispersion. In Section 4 we use the COVID-19 mortality data to draw a comparison between the MGPD, and the Generalized Additive Models (GAM). We demonstrate the differences between the two analytic strategies and highlight the superiority of the MGPD in the analysis of count data exhibiting over-

dispersion in Section 5. General discussion is presented in Section 6.

2. The Model and Its Parameters Estimation

2.1. Modified Generalized Poisson Distribution

The GPD was introduced by Consul and Jain [5]

$$P(Y=y)=\frac{\lambda_1(\lambda_1+\lambda_2 y)^{y-1}}{y!}\exp\left[-\lambda_1-\lambda_2 y\right]$$
$$\lambda_1>0 \tag{2.1}$$
$$0\preceq\lambda_2<1$$

The GPD whose probability function is given in (2.1) reduces to the well-known Poisson distribution when $\lambda_2=0$. Therefore the parameter λ_2 with the above restriction on its range, is considered the dispersion parameter. Shoukri and Mian [12] employed the parametric transformations:

$$\lambda_1=\mu/(1+\epsilon\mu)$$
$$\lambda_2=\epsilon\lambda_1 \tag{2.2}$$

Using the transformations in (2.2) we therefore have:

$$P(X=x)=\frac{(1+\epsilon x)^{x-1}}{x!}g^x(\mu,\epsilon)\exp\left[-\frac{\mu}{1+\epsilon\mu}\right] \tag{2.3}$$

where $g(\mu,\epsilon)=\frac{\mu}{1+\epsilon\mu}\exp\left[\frac{-\epsilon\mu}{1+\epsilon\mu}\right]$

For fixed ϵ, the function $g(.)$ in (2.3) is the natural parameter transformation which renders the GPD a member of the linear family of exponential class (see; [14]), with a general structure:

$$f(x)=h(x)\exp\left[\phi T(x)-A(\phi)\right] \tag{2.4}$$

We call the transformed GPD, the "Modified Generalized Poisson Distribution" or MGPD.

Shoukri and Mian [12] showed that a recurrence relation among the r^{th} non-central moments ϑ_r' is such that:

$$\vartheta_{r+1}'=\sigma^2(\mu)\frac{\partial\vartheta_r'}{\partial\mu}+\mu\vartheta_r' \tag{2.5}$$

From (2.5) we can show that:

$$\vartheta_0'\equiv 1,\vartheta_1'\equiv\mu=E(Y),\text{ and }\sigma^2(\mu)=\mu(1+\epsilon\mu)^2\equiv\text{var}(Y) \tag{2.6}$$

That is the variance is a cubic function of the population mean. We shall deal with the situation when $\epsilon>0$.

2.2. Point Estimators

Our approach for parametric estimation in this section will be for a single random sample. If $Y_1,Y_2,\cdots,Y_n$ is a random sample from the GPD (2.3), Consul and Shoukri [16] showed that the unique maximum likelihood estimates of the

parameters exist if and only if the sample variance is larger than the sample mean. Here we shall use the sample moments to obtain estimators for the model parameters (μ, ϵ).

Equating the first two sample moments $(\bar{y}, s^2)$ to their corresponding population moments

$$\bar{y} = \mu$$
$$s^2 = \frac{1}{n}\sum_{i=1}^{i=1}(y_i - \bar{y})^2 = \mu(1+\epsilon\mu)^2$$

and solving for the parameters we get:

$$\tilde{\mu} = \bar{y}$$
$$\tilde{\epsilon} = (s^2)^{1/2}(\bar{y})^{-3/2} - (\bar{y})^{-1} \tag{2.7}$$

The variance of the moment estimators of the mean and the dispersion parameter are respectively given by Shoukri and Al-Eid [15] as:

$$\text{var}(\hat{\mu}) = \mu(1+\epsilon\mu)^2 / n \tag{2.8a}$$

$$v = \text{var}(\hat{\epsilon}) = \frac{(1+\epsilon\mu)^2}{2n\mu^2}\left[1 + 2\epsilon + 3\epsilon^2\mu\right] \tag{2.8b}$$

2.2.1. Homogeneity of Dispersion Parameters

Suppose that we have k independent random samples from (2.3), which we denote $Y_{ij} \sim GPD(\mu_i, \epsilon_i)$ with n_i observations from the i^{th} population $(i = 1, 2, \cdots, k)$.

Wedenote the variance of the estimator of ϵ given in Equation (2.8) by v_i and, let $w_i = 1/v_i$, v_i is given in (2.8b).

Cochran [17] developed a general statistic Q which may be used to test the homogeneity of several population parameters. The Q statistics has asymptotically, chi-square distribution with $k-1$ degrees of freedoms. It is defined as:

$$Q\_esp = \sum_{i=1}^{k} w_i(\hat{\epsilon_i} - \bar{\epsilon})^2 / \sum_{i=1}^{k} w_i \tag{2.9}$$

where

$$\bar{\epsilon} = \sum_{i=1}^{k} w_i\hat{\epsilon_i} / \sum_{i=1}^{k} w_i \tag{2.10}$$

The hypothesis $H_0 : \epsilon_1 = \epsilon_2 = \epsilon_3 = \cdots \epsilon_k = \epsilon$ of homogeneity of dispersion parameters is rejected whenever the statistic Q_esp exceeds $Q_{\alpha,k-1}$, the upper 5% quantile of a chi-square random variable with $k-1$ degrees of freedom.

2.2.2. Testing the Significance of the Common Dispersion Parameter: $H_0 : \epsilon = 0$

Here we develop a test statistic on the null hypothesis of absence of overdispersion. For the case when μ_i's are unknown, a uniformly most powerful test for $H_0 : \epsilon = 0$ (Poisson) versus $H_1 : \epsilon > 0$ (GPD) cannot be obtained, however the locally powerful Neyman's $C(\alpha)$ test can be constructed [18]. The log-likelihood function is given by

$$\ell = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-1\right)\log\left(1+\varepsilon y_{ij}\right)+\sum_{i=1}^{k} n_i \bar{y}_{i.}\left[\log\mu_i - \log\left(1+\varepsilon\mu_i\right)\right] - \sum_{i=1}^{k} n_i\mu_i\left(\frac{1+\varepsilon\bar{y}_{i.}}{1+\mu_i}\right) \tag{2.11}$$

where $\bar{y}_{i.} = \dfrac{y_{i.}}{n_i} = \sum_{j=1}^{n_i} y_{ij}/n_i$.

The locally asymptotically most powerful $C(\alpha)$ test is to reject H_0 for large values of $\left(\partial\ell/\partial\epsilon\right)_{\epsilon=0}$. From (2.11):

$$\left(\partial\ell/\partial\epsilon\right)_{\epsilon=0} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\left(y_{ij}-1\right) - \sum_{i=1}^{k} n_i\mu_i\bar{y}_{i.} - \sum_{i=1}^{k} n_i\mu_i\left(\bar{y}_{i.}-\mu_i\right) \tag{2.12}$$

Therefore, the locally asymptotically most powerful $C(\alpha)$ test is to reject H_0 for large values of T, where

$$T = \left(\partial\ell/\partial\epsilon\right)_{\epsilon=0} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(y_{ij}-\bar{y}_{i.}\right)^2 - y_{ij}\right] \tag{2.13}$$

The statistic (2.13) is obtained from (2.12) by replacing each μ_i with root n_i consistent estimator, $\hat{\mu}_i$. The simplest $\hat{\mu}_i$ is the maximum likelihood estimator $\hat{\mu}_i = \bar{y}_{i.}$. Moran [18] pointed out that the $C(\alpha)$ test statistic T is asymptotically normal. It can be easily shown that:

$$E(T) = \sum_{i=1}^{k}\left[\left(n_i-1\right)\mu_i\left(1+\varepsilon\mu_i\right)^2 - n_i\mu_i\right] \tag{2.14}$$

and

$$\operatorname{var}(T) = \sum_{i=1}^{k}\left\{2\left(n_i-1\right)\mu_i^2\left(1+\varepsilon\mu_i\right)^4 + \frac{1}{n_i}\left[\mu_i\left(1+\varepsilon\mu_i\right)^4\left(1+3\mu_i+10\varepsilon\mu_i+15\varepsilon^2\mu_i^2\right) - 3\mu_i^2\left(1+\varepsilon\mu_i\right)^4 + \mu_i\left(1+\varepsilon\mu_i\right)^2 - 2\mu_i\left(1+\varepsilon\mu_i\right)^3\right] + \frac{\mu_i}{n_i}\right\} \tag{2.15}$$

Under $H_0:\epsilon=0$, (2.14) and (2.15) reduce respectively to $E^\circ = -\sum_{i=1}^{k}\mu_i$ and

$$v^\circ = \sum_{i=1}^{k}\left(2\left(n_i-1\right)\mu_i^2 + \frac{\mu_i}{n_i}\right) \tag{2.16}$$

The hypothesis $H_0:\epsilon=0$ is rejected whenever:
$Q(\epsilon=0) = \left(T-E^\circ\right)^2/v^\circ$ exceeds $Q_{\alpha,1}$, the upper 5% quantile of a chi-square random variable with one degree of freedom.

3. Testing Equality of Means

Based on the one-way layout data considered in the previous section, we would like to test the null hypothesis $H_0:\mu_1=\cdots=\mu_k=\mu$ against H_a: at least two of the $\mu_i's$ are different, for all $\epsilon>0$. The log likelihood under the hypothesis H_a: is given by (2.11), and will be denoted by ℓ_a, the log likelihood under H_0 will be denoted by ℓ_0 and is obtained by replacing $\mu_i=\mu\left(i=1,2,\cdots,k\right)$ in

(2.11). Under H_a, the maximum likelihood estimator of μ_i is

$$\hat{\mu}_i = \bar{y}_{i.}$$

And the maximum likelihood estimator $\hat{\epsilon}_a$, of ϵ is the non-negative root of

$$\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\frac{(y_{ij}-1)y_{ij}}{1+\hat{\epsilon}_a y_{ij}} - \frac{n_i(\bar{y}_{i.})^2}{(1-\hat{\epsilon}_a\bar{y}_{i.})}\right] = 0 \tag{3.1}$$

Under H_0 the maximum likelihood estimator of the common mean μ is $\hat{\mu} = y../N = \bar{y}$, where, $N = \sum_{i=1}^{k} n_i$.

The maximum likelihood estimator of $\hat{\epsilon}_o$ and ϵ under H_0 is the positive root of

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(y_{ij}-1)y_{ij}}{1+\hat{\epsilon}_o y_{ij}} - \frac{N(\bar{y})^2}{(1+\hat{\epsilon}_o\bar{Y})} = 0 \tag{3.2}$$

Detailed discussion on the necessary and sufficient conditions that (3.1) and (3.2) to have a unique root is given in Consul and Shoukri [16].

Denoting the maximized log likelihood under H_a by L_a, and that under H_0 by L_0, the likelihood ratio test, which has an asymptotic distribution of chi-squared with $(k-1)$ degree of freedom is:

$$\begin{aligned}\lambda &= 2(L_a - L_0)\\ &= 2\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-1)\log\left\{\frac{1+\hat{\varepsilon}_a y_{ij}}{1+\hat{\varepsilon}_0 y_{ij}}\right\} + \sum_{i=1}^{k} n_i\bar{y}_i\log\left\{\frac{\bar{y}_{i.}(1+\hat{\varepsilon}_0\bar{y})}{\bar{y}(1+\hat{\varepsilon}_a\bar{y}_{i.})}\right\}\right]\end{aligned} \tag{3.3}$$

As an alternative to the likelihood ratio test (3.3), we present the Neyman's $C(\alpha)$ statistic which has local optimal properties. Suppose that μ_i can be written as $\mu_i = \mu + \delta_i$ with $\delta_k = 0$. Then testing the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ is equivalent to testing $H_0: \delta_i = o\ (i = 1,2,3,\cdots,k)$, where μ and ϵ are nuisance parameters. We reparametrize (11.2), and denote the resulting function by ℓ^*.

Define $\delta = (\delta_1, \delta_2, \cdots, \delta_{k-1})$, $\tau = (\tau_1, \tau_2)' = (\mu, \epsilon)$.

$$\phi_i(\tau) = \left[\frac{\partial \ell^*}{\partial \delta_i}\right]_{\bar{\delta}=0} \quad i = 1,2,\cdots,k-1$$

$$\Delta_i(\tau) = \left[\frac{\partial \ell^*}{\partial \tau_j}\right]_{\bar{\delta}=0} \quad j = 1,2$$

Let $\hat{\tau}$ be any root-n consistent estimator of τ under the null hypothesis. Moran [18] showed that the $C(\alpha)$ test is based on $F_i(\hat{\tau}) = \phi_i(\hat{\tau}) - \gamma_{i1}\Delta_{i1}(\hat{\tau}) - \gamma_{i2}\Delta_{i2}(\hat{\tau})$, where γ_{i1} and γ_{i2} are the partial regression coefficients of ϕ_i on Δ_1 and Δ_2 respectively. Define the following matrices:

$$P_{ij} = -E\left[\frac{\partial^2 \ell^*}{\partial\delta_i\partial\delta_j}\right]_{\bar{\delta}=0},\ Q_{ij} = -E\left[\frac{\partial^2 \ell^*}{\partial\delta_i\partial\tau_j}\right]_{\bar{\delta}=0},$$

and $R_{ij} = -E\left[\dfrac{\partial^2 \ell^*}{\partial\tau_i \partial\tau_j}\right]_{\bar{\delta}=0}$.

Here, we replaceτ by its estimator $\hat{\tau}$ in *F*, *P*, *Q*, and *R*, the $C(\alpha)$ test statistic is given by

$$F'\left(p - QR^{-1}Q'\right)^{-1} F \tag{3.5}$$

The asymptotic distribution of the test statistic given in (3.5) will be that of a chi-square with $k-1$ degrees of freedom.

Now, there are two possible root-n consistent estimators of τ, under H_0:

The first is the maximum likelihood estimator $\hat{\tau} = \left(\bar{y}, \hat{\epsilon_0}\right)'$, which on substitution we get $\Delta_j\left(\hat{\tau}\right) = 0$ $\left(j = 1,2\right)$, and hence $F_j\left(\hat{\tau}\right) = \phi_i\left(\hat{\tau}\right)$. Accordingly, (3.5) reduces to

$$C^2 = \sum_{i=1}^{k} \frac{n_i\left(\bar{y}_{i.} - \bar{y}\right)^2}{\bar{y}\left(1 + \hat{\varepsilon}_0 \bar{y}\right)^2} \tag{3.6}$$

The hypothesis of equality of population means is thus rejected whenever C^2 exceeds $Q_{\alpha,k-1}$, the upper 5% quantile of a chi-square random variable with $k-1$ degrees of freedom. For more details we refer the reader to [12].

4. ANCOVA: The Generalized Poisson Regression

It is well-known that ANOVA and regression are related techniques that are concerned with testing the differences in group means after adjusting for the confounding effects of potential risk factors and covariates. Since the MGPD is a member of the linear exponential family (for fixed ϵ) Shoukri and Mian [12] expressed the expectation μ_i of y_i as:

$$\eta\left(\mu_i\right) = X_i^T \beta \tag{4.1}$$

In Equation (4.1), X_i is a set of measured $\left(P+1\right)$ covariates, and a subset of theses covariates defines a set of indicators (dummy) variables to identify categorical effects. The transformation $\eta(\cdot)$ is a monotone, differentiable function named "the link function". To estimate $\beta_0, \beta_1, \cdots, \beta_p$, and ϵ we construct the log-link so that:

$$\mu_i\left(x\right) = \exp\left[X_i^T \beta\right]$$

The logarithm of the likelihood function will be proportional to

$$\begin{aligned} \ell = \ell\left(\beta, \varepsilon\right) &= \sum_{i=1}^{k}\left(y_i - 1\right)\ln\left(1 + \varepsilon y_i\right) + \sum_{i=1}^{k} y_i \ln \mu_i\left(x\right) \\ &- \sum_{i=1}^{k} y_i \ln\left(1 + \varepsilon\mu_i\left(x\right)\right) - \sum_{i=1}^{k} \frac{\mu_i\left(x\right)\left(1 + \varepsilon y_i\right)}{1 + \varepsilon\mu_i\left(x\right)} \end{aligned} \tag{4.2}$$

The first and second partial derivatives are given by:

$$\frac{\partial \ell}{\partial \varepsilon} = \dot{\ell}_{\varepsilon} = \sum_{i=1}^{k} \frac{y_i\left(y_i - 1\right)}{1 + \varepsilon y_i} - \sum_{i=1}^{k} \frac{y_i \mu_i\left(x\right)}{1 + \varepsilon\mu_i\left(x\right)} - \sum_{i=1}^{k} \frac{\mu_i\left(x\right)\left(y_i - \mu_i\left(x\right)\right)}{\left(1 + \varepsilon\mu_i\left(x\right)\right)^2} \tag{4.3}$$

$$\frac{\partial \ell}{\partial \beta_r} = \dot{\ell}_r = \sum_{i=1}^{k} \frac{\left(y_i - \mu_i(x)\right)}{\left(1+\varepsilon\mu_i(x)\right)^2} x_{ir} \tag{4.4}$$

$$\frac{\partial^2 \ell}{\partial \varepsilon \partial \beta_r} = \ddot{\ell}_{\varepsilon r} = -2\sum_{i=1}^{k} \frac{\mu_i(x)\left(y_i - \mu_i(x)\right)}{\left(1+\varepsilon\mu_i(x)\right)^3} x_{ir} \tag{4.5}$$

$$\frac{\partial^2 \ell}{\partial \beta r \partial \beta s} = \ddot{\ell}_{rs} = -\sum_{i=1}^{k} \left[\frac{\mu_i(x)}{\left(1+\varepsilon\mu_i(x)\right)^2} x_{ir}x_{is} + 2\varepsilon \frac{\left(y_i - \mu_i(x)\right)\mu_i(x)}{\left(1+\varepsilon\mu_i(x)\right)^3} x_{ir}x_{is} \right] \tag{4.6}$$

$$\frac{\partial^2 \ell}{\partial \varepsilon^2} = \ddot{\ell}_{\varepsilon r} - \sum_{i=1}^{k} \frac{y_i^2\left(y_i - 1\right)}{\left(1+\varepsilon y_i\right)^2} + \sum_{i=1}^{k} \frac{y_1\mu_i^2(x)}{\left(1+\varepsilon\mu_i\right)^2} + 2\sum_{i=1}^{k} \frac{\mu_i^2(x)\left(y_i - \mu_i(x)\right)}{\left(1+\varepsilon\mu_i(x)\right)^3} \tag{4.7}$$

Taking the expected value of the negative of the second partial derivatives we get the Fishers' information matrix I, whose elements are:

$$-E\left[\ddot{\ell}_{rs}\right] = I_{rs} = \sum_{i=1}^{k} \frac{\mu_i(x)}{\left(1+\varepsilon\mu_i(x)\right)^2} x_{ir}x_{is},\ r,s,=1,2,\cdots,p \tag{4.8}$$

$$-E\left[\ddot{\ell}_{\varepsilon s}\right] = I_{\varepsilon r} = 0 \tag{4.9}$$

From Consul and Shoukri [16] we get:

$$-E\left[\ddot{\ell}_{\varepsilon\varepsilon}\right] = i_{\varepsilon\varepsilon} = 2\left(1+2\varepsilon\right)^{-1} \sum_{i=1}^{k} \frac{\mu_i^2(x)}{\left(1+\varepsilon\mu_i(x)\right)^2} \tag{4.10}$$

The asymptotic distributions of the regression estimators can be established using the results in [12].

Our approach to the data analysis when the main interest is comparing group means in the presence of potential risk factors and confounders is summarized in three steps. In the first step we use the MLE to estimate the regression parameters using Equation (4.2), without including the groups as independent variable. In the second step, we extract the residuals (E) of the generalized Poisson regression model, defines as:

E = Observed dependent variable – predicted value of the dependent variable

In the final step we test the normality and variance homogeneity of E. Thereafter, we use nonparametric ANOVA with the residuals being the dependent variable, and the groups being the independent variables to complete the ANCOVA testing.

5. Data Analyses

Al-Gahtani *et al.* [19] analyzed COVID-19 case fatality data collected retrospectively from the start of the of the epidemic to December 2-2020, the day the Pfizer vaccine was approved by the American Center for Disease Control (CDC). The data were collected from 120 countries grouped into 15 regions [19] as shown in **Table 1**. We will reanalyze the data such that:

The response variable is the aggregate number of COVID-19 deaths which we

denote by "y". We shall use different set of covariates, and these are:

- Region: The factor variable which is the main effect.
- The other covariates are:

1) X_1 =log (percentage of obese personsin a country reported in 2018) [21] [22];

2) X_2 = log (population density) [23];

3) X_3 = log (number of people with colorectal cancer in a country reported in 2017) [24];

4) X_4 = log (Chronic Kidney Disease—case fatality in a country as reported in 2017) [20].

In **Figure 1** we show the histogram of (y), the aggregate of COVID-19 deaths during the study period. The distribution is positively skewed with variance much larger than the mean.

Direct calculations from the summary statistics given in **Table 2** give:

$Q\_esp = 0.00004$, and the corresponding p-value = 0.999. Therefore, the hypothesis of homogeneity of dispersion parameters is supported by the data. Moreover, $Q\ (\epsilon = 0)$ is quite large and the corresponding p-value = 0.00001.

Table 1. Adaptedtable: Countries and the corresponding Regional classification as given in https://doi.org/10.1016/s0140-6736(20)30045-3 [20]. In the first column we have the countries, in the second column we have Region or group name followed by the number of countries within the group. In the last column we have Region code.

Countries	Region name	Region Code
Peru, Ecuador, Bolivia	Andean. Latin (3)	10
Kazakhstan, Georgia, Armenia, Azerbaijan, Kyrgyzstan, Uzbekistan, Tajikistan	Central Asia (7)	2
Czechia, Romania, Hungary, Serbia, Bulgaria, Croatia, Slovakia, Bosnia, Slovenia, North-Macedonia, Albania, Montenegro	Central Europe (12)	5
Brazil, Columbia, Mexico, Panama, Costa Rica, Guatemala, Honduras, Venezuela Paraguay, El-Salvador	Central Latin America (10)	11
Dominican Republic, Puerto Rico, Jamaica	Caribbean (3)	9
Ethiopia, Kenya, Uganda, Zambia, Madagascar. Mozambique, Angola, French Guinea	CESSA (8)	13
Indonesia, Philippine, China, Myanmar, Malaysia, Sri Lanka, French Polynesia, Maldives	East Asia (8)	1
Russia, Poland, Ukraine Belarus, Lithuania, Latvia, Estonia	East Europe (7)	6
Japan, Singapore, Republic Korea, Australia	HIAP (4)	4
Iran, Iraq, Turkey, Morocco, Saudi Arabia, Israel, Jordan, United Arab, Kuwait, Qatar, Lebanon, Oman, Egypt, Occupied Palestine, Tunisia, Bahrain, Algeria, Libya, Afghanistan, Sudan	MENA (20)	12
India, Bangladesh, Pakistan, Nepal	South Asia (4)	3
Chile, Argentina, Uruguay	South Latin America (3)	8
South Africa, Namibia, Zimbabwe	SSSA (3)	15
USA, France, Spain, UK, Italy, Germany, Belgium, Netherland Canada, Switzerland, Portugal, Austria, Sweden, Greece, Denmark, Ireland, Norway, Luxemburg, Finland, Cyprus	West Europe and North America (20)	7
Mali, Nigeria, Ghana, Cameroon, Ivory Coast Senegal, Guinea, Cape Verde	WSSA (8)	14

CSSA = Central Sub-Saharan-Africa; MENA = Middle East and North Africa; HIAP = High Income Asian Pacific; WSSA = Western Sub-Saharan-Africa; SSSA = Southern Sub-Saharan Africa.

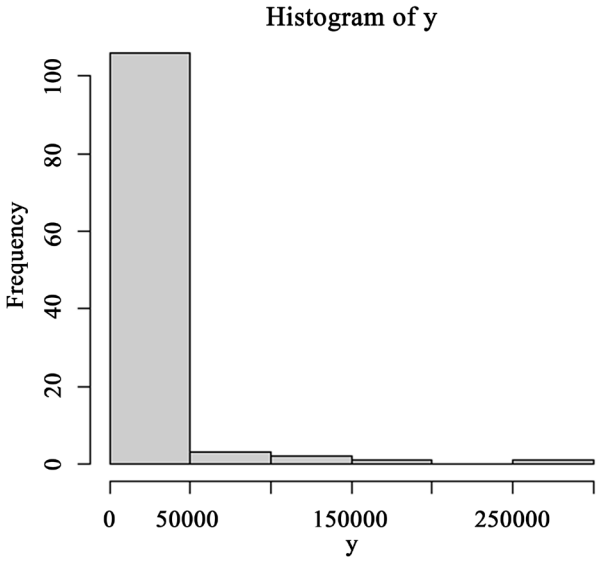

Figure 1. Histogram COVID-19 deaths (y). It is skewed to the right showing clear over dispersion in the data.

Table 2. Summary measures of COVID-19 deaths: group sizes (n), means (m), standard deviation (s) and the estimates of the dispersion parameter (eps) per group.

Region	n	m	s	eps
1 An. Latin	3	19,496	14,448	0.005
2 C. Asia	7	1350	836	0.016
3 C. EUROP	11	3515	3577	0.018
4 C. Latin	10	33,158	59,252	0.01
5 Caribbean	3	976	1177	0.037
6 CESSA	7	640	658	0.039
7 E. ASIA	6	5452	6495	0.016
8 E. Europe	7	10,488	15,213	0.014
9 HIAP	4	920	934	0.032
10 MENA	19	5986	11,027	0.024
11 S. Asia	4	38,606	66,403	0.009
12 S. Latin	2	27,080	16,476	0.004
13 SSSA	3	7358	12,372	0.019
14 W. Eur	20	28,061	59,609	0.013
15 WSSA	7	692	806	0.043

Therefore, the hypotheses that the common dispersion is not significantly different from zero is not supported by the data. The C^2-statistic is quite large as well, and the corresponding p-value is near zero, therefore the hypothesis of equality of mean counts in all regions (aggregate COVID-19 deaths) is also not supported by the data.

We shall write a function using the R-program for the estimation of the regression parameters. The iteration process requires staring points. We obtain the staring points by first fitting the classical Poisson regression, which is done using

the following code:

$$\text{out1} = \text{GLM}\ (y \sim x_1 + x_2 + x_3 + x_4, \text{data} = \text{data2}, \text{family} = \text{Poisson}).$$

Having obtained the parameter estimates from the Poisson regression, we use them to start the iteration process and obtain final estimates as shown in the Appendix.

The MGPD regression results are summarized in Table 3.

The correlation between the observed and predicted COVID-19 death counts is (0.758).

Figure 2 gives the Q-Q plot of the quantiles of model residuals exhibiting close agreement among with the quantiles of the standard normal distribution.

To complete the ANCOVA testing we use theKruskal-Wallis test whereby residuals of the MGPD regression model are used as dependent variables and the "Regions", or groups as independent variables. The results are summarized as follows:

Kruskal-Wallis chi-squared = 14.936, p-value = 0.1344.

Therefore, after adjusting for the covariates, there is not sufficient evidence to

Table 3. Maximum likelihood estimation of the MGPD regression using R.

Parameter	Estimate	SE	Z	p value
1-Intercept	−0.299	1.138	−0.262	0.7900
2-X_1	0.659	0.186	3.545	0.0004
3-X_2	0.489	0.131	3.729	0.0002
4-X_3	0.425	0.107	3.400	0.0002
5-X_4	−0.385	0.090	−4.280	0.00002
6-ϵ	0.028	0.002	13.537	0.000001

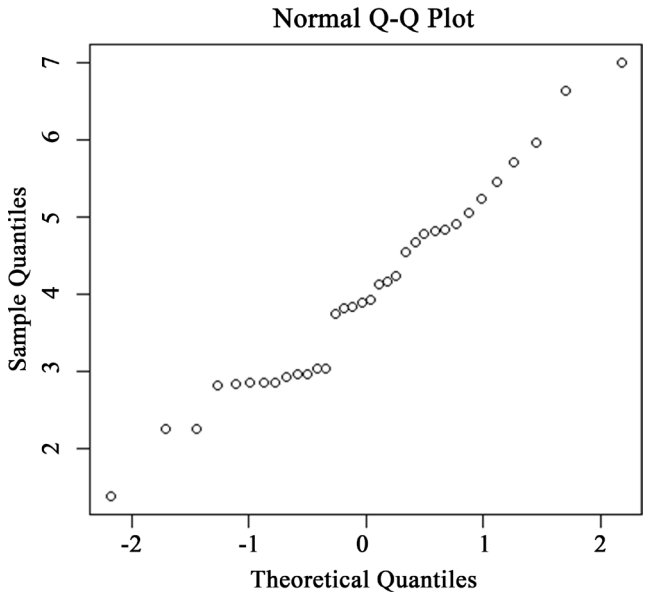

Figure 2. Plot of the quantiles of the residuals of the MGPD regression model against the quantiles of the standard normal distribution.

reject the hypothesis of equality of mean counts in COVID-19 deaths among the "Regions".

6. Nonparametric Regression Modeling: Generalized Additive Models

The Generalized Additive Models (GAM) are recent developments that are becoming popular as modeling techniques. It is nonparametric in nature and, even though less powerful, it is quite robust against departure from the assumptions required by classical GLM regression models. The GAM allow us to include non-linear smoothers into the modeling strategy. In mathematical terms GAM solve the following equation:

$$g(\mu_i) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5) \tag{6.1}$$

The $f_j(x_j)$ are smooth functions to be estimated. Equation (6.1) seems complex, but it is very simple to understand. The first thing to notice is that with GAM we are not necessarily estimating the response directly, *i.e.* we are not modelling y. In fact, as with GLM we have the possibility to use link functions to model non-normal response variables (and thus perform Poisson or logistic regression) [14]. Therefore, the term $g(\mu)$ is simply the transformation of y needed to "linearize" the model. When we are dealing with a normally distributed response this term is simply replaced by y. The second part of the equation, where we have two terms: the parametric and the non-parametric part. In GAM we can include all the parametric terms we can include in Linear Model or GLM, for example linear or polynomial terms. The second part is the non-parametric smoother that will be automatically fitted, and it is the key point of GAMs. A complete and lucid account of the GAM theory can be found in [25] [26] [27].

We fitted the GAM to the data using the R-package "GAM", and the next two lines are the needed code:

```
library(gam);
agam=gam(y~Region+x1+x2+x3+x4,data=data2).
```

Call: gam(formula = y ~ code + x1 + x2 + x3 + x4, data = data2).

The following results are obtained from the GAM fitting to the data:

1) Null Deviance: 135512150629 on 112 degrees of freedom;

2) Residual Deviance: 74451674616 on 96 degrees of freedom.

From which: The correlation between observed counts and predicted counts is: $(1-74451674616/135512150629)^{1/2} = 0.671$.

The GAM results are shown in **Table 4**, and the Q-Q plot of the model residuals is given in **Figure 3**, showing that the model residuals are not as close to normality as the residuals of the MGPD regression model.

7. Discussion

In this paper we demonstrated the use of the MGPD as a model for the ANCOVA. We used a two-steps approach. In the first step we used the regression models to

Table 4. The results of fitting the GAM: ANOVA for Parametric Effects.

Source	DF	Sum Sq	Mean Sq	F value	Pr. (>F.)
Region	12	1.5856e+10	1.3213e+09	1.7038	0.078
x_1	1	1.9335e+09	1.9335e+09	2.4932	0.118
x_2	1	2.3000e+10	2.3000e+10	29.6566	3.963e−07***
x_3	1	1.1518e+10	1.1518e+10	14.8522	0.0002
x_4	1	8.7527e+09	8.7527e+09	11.2859	0.001
Residuals	96	7.4452e+10	7.7554e+08		

***significant at level of significance less than 0.00001.

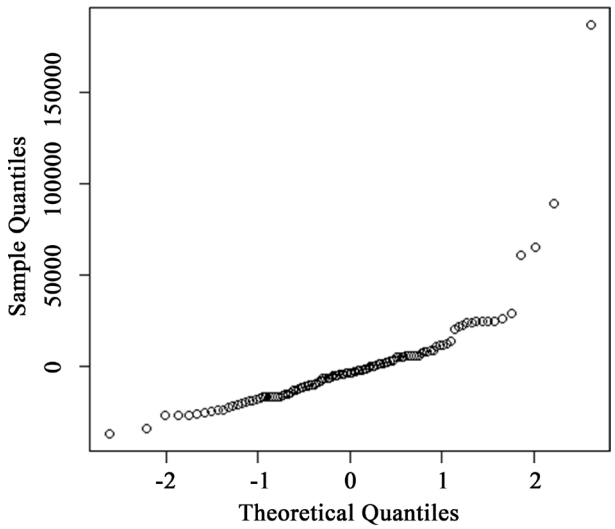

Figure 3. Plot of the quantiles of the residuals of the GAM nonparametric regression model against the quantiles of the standard normal distribution.

assess the influence of possible confounders and covariates on the outcome of interest. Thereafter we extracted the model residuals and used these residuals as a dependent variable of a nonparametric ANOVA, with the groups being the independent predictors. We note that while there was a significant difference among the group means in the univariate analysis, such difference was not significant in the second step of the ANCOVA. We note that the MGPD regression model showed high correlation (0.758) between the observed counts and the model based predicted counts, indicative of a good fit by the model to the given data. On using the Q-Q plot, model residuals are shown to have close agreement to the empirical quantiles of the standard normal distribution. This shows that the model is quite reliable as a predictive tool, and that the distribution of the estimated regression parameters is that of a multivariate normal.

For the sake of comparison, we fitted the data using the GAM, a nonparametric regression approach. This approach deals with the covariates as factors. The GAM model showed that after adjusting for the covariates within the same model, there are no significant differences among regions. These findings are in agreement with those based on the MGPD regression. The GAM did not pro-

duce estimate for the dispersion parameter ϵ. The measure of goodness of fit of the GAM was (0.671), which is much lower than that of the MGPD. The MGPD model has several advantages when compared to the GAM. First, The GAM cannot be used as a predictive tool, while the MGPD model can be used to predict the mean of the response variable. Second, the residuals of the MGPD regression model have a distribution that is almost normal. This emphasizes the reliability of the likelihood based statistical estimation of the model parameters. Finally, our two-steps approach to data fitting makes helps avoiding both overfitting and possible multicollinearity.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cox, D.R. (1983) Some Remarks on Overdispersion. *Biometrika*, **70**, 269-274. https://doi.org/10.1093/biomet/70.1.269

[2] Hinde, J. and Demetrio, C.G.B. (1998) Overdispersion: Models and Estimation. *Computational statistics and Data Analysis*, **27**, 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3

[3] Hayat, M.J. and Higgins, M. (2014) Understanding Poisson Regression. *Journal of Nursing Education*, **53**, 207-215. https://doi.org/10.3928/01484834-20140325-04

[4] Joe, H. and Zhu, R. (2005) Generalized Poisson Distribution: The Property of Mixture of Poisson and Comparison with the Negative Binomial Distribution. *Biometrical Journal*, **47**, 219-229. https://doi.org/10.1002/bimj.200410102

[5] Consul, P.C. and Jain, G.C. (1970) On the Generalization of Poisson Distribution. *Annals of Mathematical Statistics*, **41**, 1387.

[6] Consul, P.C. (1989) Generalized Poisson Distribution. Marcel Dekker Inc., New York.

[7] Janardan, K.G. and Schaeffer, D.J. (1977) Models for the Analysis of Chromosomal Aberrations in Human Leukocytes. *Biometrical Journal*, **19**, 599-612. https://doi.org/10.1002/bimj.4710190804

[8] Tanner, J.C. (1961) A Derivation of Borel Distribution. *Biometrika*, **40**, 222-224. https://doi.org/10.1093/biomet/48.1-2.222

[9] Consul, P.C. and Shoukri, M.M. (1988) Some Chance Mechanisms Related to a Generalized Poisson Probability Model. *American Journal of Mathematical and Management Sciences*, **8**, 181-202. https://doi.org/10.1080/01966324.1988.10737237

[10] Jiang, H. and Wong, W.H. (2009) Statistical Inferences for Isoform Expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032. https://doi.org/10.1093/bioinformatics/btp113

[11] Srivastava, S. and Chen, L. (2010) A Two-Parameter Generalized Poisson Model to Improve the Analysis of RNA-seq Data. *Nucleic Acids Research*, **38**, e170. https://doi.org/10.1093/nar/gkq670

[12] Shoukri, M.M. and Mian, I.U.H. (1991) Some Aspects of Statistical Inference on the Lagrange (Generalized) Poisson Distribution. *Communication in Statistics: Computations and Simulations*, **20**, 1115-1137.

https://doi.org/10.1080/03610919108812999

[13] Cox, D.R. and Reid, N. (1987) Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **49**, 1-39. https://doi.org/10.1111/j.2517-6161.1987.tb01422.x

[14] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. Chapman and Hall, London. https://doi.org/10.1007/978-1-4899-3242-6

[15] Shoukri, M.M. and Al-Eid, M. (2020) Inference Procedures on the Ratio of Modified Generalized Poisson Distribution Means: Applications to RNA_SEQ Data. *International Journal of Statistics in Medical Research*, **9**, 41-49. https://doi.org/10.6000/1929-6029.2020.09.05

[16] Consul, P.C. and Shoukri, M.M. (1984) Maximum Likelihood Estimation of the Generalized Poisson Distribution. *Communications in Statistics*, *Theory and Methods*, **13**, 1533-1547. https://doi.org/10.1080/03610928408828776

[17] Cochran, W.G. (1954) Some Methods for Strengthening the Common X2 Tests. *Biometrics*, **10**, 417-451. https://doi.org/10.2307/3001616

[18] Moran, P.A.P. (1970) On Asymptotically Optimal Test of Composite Hypotheses. *Biometrika*, **57**, 47-55. https://doi.org/10.1093/biomet/57.1.47

[19] Al-Gahtani, S., Shoukri, M. and Al-Eid, M. (2021) Predictors of the Aggregate of COVID-19 Cases and Its Case-Fatality: A Global Investigation Involving 120 Countries. *Open Journal of Statistics*, **11**, 259-277. https://www.scirp.org/journal/ojs https://doi.org/10.4236/ojs.2021.112014

[20] Cockwell, P. and Fisher, L.-A. (2017) Global, Regional, and National Burden of Chronic Kidney Disease, 1990-2017: A Systematic Analysis for the Global Burden of Disease Study. *The Lancet*, **395**, 709-733.

[21] Rottoli, M., Bernante, P., Garelli, S. and Gianella, M. (2020) How Important Is Obesity as a Risk Factor for Respiratory Failure, Intensive Care Admission and Death in Hospitalized COVID-19 Patients? Results from a Single Italian Center. *European Journal of Endocrinology*, **183**, 389-397. https://doi.org/10.1530/EJE-20-0541

[22] https://worldpopulationreview.com/en/country-rankings/obesity-rates-by-country

[23] Rashed, E.A., Kodera, S., Gomez-Tames, J. and Hirata, A. (2020) Correlation between COVID-19 Morbidity and Mortality Rates in Japan and Local Population Density, Temperature, and Absolute Humidity. *International Journal of Environmental Research and Public Health*, **17**, 5447. https://doi.org/10.3390/ijerph17155477

[24] GBD Colorectal Cancer Collaborators (2019) The Global, Regional, and National Burden of Colorectal Cancer and Its Attributable Risk Factors in 195 Countries and Territories, 1990-2017: A Systematic Review for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology and Hepatology*, **4**, 913-933.

[25] Hastie, T.J. and Tibshirani, R.J. (1990) Generalized Additive Models. Chapman & Hall/CRC, London.

[26] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511755453

[27] Wood, S.N. (2017) Generalized Additive Models: An Introduction with R. Second Edition, CRC Press, Boca Raton. https://doi.org/10.1201/9781315370279

Appendix A

R-Code fitting the Generalized Poisson using the method of maximum likelihood.

###Notations: b_i are the regression parameters, mu is the mean function, "n" is the number of observations "ll" denotes the ###log-likelihood function, eta is the linear predictor ###

```
llik= function(y,par){
b0=par[1]
b1=par[2]
b2=par[3]
b3=par[4]
b4=par[5]
k=par[6]
n=length(y)
eta=b0+b1*x1+b2*x2+b3*x3+b4*x4
mu=exp(eta)
ll= sum(y*log(mu/(1+(k*mu))))+sum((y-1)*log(1+(k*y))
    +((-mu*(1+(k*y)))/ (1+(k*mu)))-lgamma(y+1))

return(-ll)
}
res=optim(par=c(1.4,.84,.07,.95,-.37,.1),llik,y=y,method="BFGS",hessian=T)
theta=res$par
theta
#CALCULATING THE STANDARD ERRORS OF MLE
out2=nlm(llik,theta,y=y,hessian=TRUE)
summary(out2)
plot(data2$y,resid(out2))
data_new=data.rame(data2$y,resid(out2))
fish=out2$hessian
solve(fish)
element=diag((solve(fish)))
se=sqrt(element)
se
z=theta/se
out.GMPD=data.frame(theta,se,z)
out.GMPD
#### FINAL ESTIMATES -0.30007804 0.65884589 0.48926214 0.42536091
-0.38469265
data2$y_hat=exp(-.3+.66*data2$x1+.5*data2$x2+.425*data2$x3-.384*data2$x4)
data2$y
data2_error=data.frame(data2$y,data2$y_hat)
cor(data2$y,data2$y_hat) #####0.76
```

```
data2_error=data2$y-data2$y_hat
data2$response=sqrt((data2_error)^(1/3))
qqnorm(data2$response)
### SHAPITO WILK TEST OF NORMALITY####
shapiro.test(data2$response)
leveneTest(data2$response~data2$Region)
aov_result=aov(data2$response~data2$Region)
####ANOVA ON THE RESIDUALS WITH REGION BEING
THE INDEPENDENT VARIABLEUSING KRUSKAL_WALLIS####
levels(data2$Region)
aov_result=aov(data2$response~data2$Region)
summary(aov_result)
boxplot(data2_error~data2$Region,xlab="Region",ylab="GPD
Residuals",main="CODID-19 Deaths")
kruskal_result=kruskal.test(data2$response~data2$Region)

###END OF CODE####.
```