Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 1-14
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

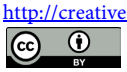
Quasi-Binomial Regression Model for the Analysis of Data with Extra-Binomial Variation

Mohamed M. Shoukri[1], Maha M. Aleid[2]

[1]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London Ontario, Canada
[2]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, KSA
Email: mmshouk@uwo.ca, Shoukri.mohamed@gmail.com, Mahaeid@kfshrc.edu.sa

How to cite this paper: Shoukri, M.M. and Aleid, M.M. (2022) Quasi-Binomial Regression Model for the Analysis of Data with Extra-Binomial Variation. *Open Journal of Statistics*, **12**, 1-14.
https://doi.org/10.4236/ojs.2022.121001

Received: December 17, 2021
Accepted: January 26, 2022
Published: January 29, 2022



Abstract

Objectives: Developing inference procedures on the quasi-binomial distribution and the regression model. **Methods**: Score testing and the method of maximum likelihood for regression parameters estimation. **Data**: Several examples are included, based on published data. **Results**: A quasi-binomial model is used to model binary response data which exhibit extra-binomial variation. A partial score test on the binomial hypothesis versus the quasi-binomial alternative is developed and illustrated on three data sets. The extended logit transformation on the binomial parameter is introduced and the large sample dispersion matrix of the estimated parameters is derived. The Nonlinear Mixed Procedure (NLMIXED) in SAS is shown to be very appropriate for the estimation of nonlinear regression.

Keywords

Quasi-Binomial Distribution, Extra Binomial Variations, Score Test, Quasi-Binomial Regression Model, COVID-19 Case Fatality Data

1. Introduction

In many biological and toxicological experiments, the variable of interest is in the form of counts resulting from binary responses. In such experiments the data may sometimes exhibit greater heterogeneity (variation) than the binomial model. It has long been presumed that an inherent characteristic of data from these types of studies is the tendency for individual experimental units to respond more alike than individuals from other groups, which is commonly known as the "group effect". When the experimental units are animals which are

treated with varying doses of compounds, such group effect is also known as "litter effect". The litters in each group contain varying numbers of live fetuses and some of these have a specific abnormality. To explain the extra-variation caused by the "litter effect", several generalized statistical models have been proposed in the literature. Altham [1] proposed that the analysis of such experiments be based on two-parameter generalizations of the binomial model which allows for the presence of dependent responses within groups and gave two models. Kupper and Haseman [2] suggested a correlated binomial model which is identical to Altham's additive generalization of the binomial model. Williams [3] proposed that the analysis of toxicological studies be based on the beta-binomial model, which is another generalization of the binomial model. However, [1] indicated that the beta-binomial model allows only positive association between the subjects of a group whereas the correlated binomial and the multiplicative generalization of the binomial model allow negative as well as positive associations. A much wider class of family of distributions known as "The generalized Linear Mixed Models" or GLMM [4] is developed and is used extensively in many applications and to deal with overdispersion that exists in count and binary data.

In this paper we show that the quasi-binomial distribution of Consul [5] reviewed by Shenton in [6] can be used as an alternative model for the analysis of overly dispersed dichotomous data. The quasi-binomial (QBD) model has two parameters p and ϕ. The parameter p will be called the binomial parameter and the other parameter ϕ will be called the dispersion parameter. When $\phi = 0$, the quasi-binomial distribution (QBD) reduces to the binomial distribution. Since the binomial distribution hypothesis is the focus of our investigations, it is natural to derive a test statistic for testing the null hypothesis $\phi = 0$.

The paper is structured as follows: in Section 2 we derive the $C(\alpha)$ binomial score test of significance [7] and [8] which is asymptomatically optimal against a QBD alternative and apply the test to some real data in Section 3. In Section 4 we develop a QBD regression model to account for possible extraneous sources of variation. The methods are applied to COVID-19 mortality data.

The flowchart in the Appendix outlines the steps of the model developments and the applications.

2. Quasi-Binomial Distribution and $C(\alpha)$ Binomial Score Test of Significance

A discrete random variable Y is said to have a QBD if and only if its probability function is given from [6] as:

$$P_r(Y=y) = p(y) = \binom{m}{y} p(p+y\phi)^{y-1}(1-p-y\phi)^{m-y}, \quad (1)$$

for $y = 0,1,2,3,\cdots,m$ and zero otherwise and where $0 < p < 1$, $-p/m < \phi < (1-p)/m$. It reduces to the binomial when $\phi = 0$. The r.v. Y

represents the number of successes in m trials such that the probability for the first success is p and that the probability of success in each of the other trials is $p+y\phi$. Thus the probability of success increases or decreases as ϕ is positive or negative and is directly proportional to the number of successes y. All the moments of the QBD are finite and the parameter ϕ has a very substantial effect on the model. The Variance of the QBD is larger or smaller than the variance of the binomial model depending upon $\phi>0$ or $\phi<0$. Consul [9] provided a detailed study of the characteristics of the QBD and gave numerous properties and moment based estimation of the model parameters. The mean μ of the QBD model (1) is given by

$$\mu = mp\left[1+\sum_{j=1}^{m-1}\phi^{j}\left(m-1\right)_{(j)}\right] \tag{2}$$

We shall formulate a $C(\alpha)$ test for testing the binomial model against the QBD alternative. This can be done by testing the null hypothesis $H_0:\phi=0$ against its negation in the presence of the nuisance parameter p. Moran [9] showed that for such problems the $C(\alpha)$ tests, suggested by Neyman [8], are asymptomatically equivalent to tests using the maximum likelihood estimates.

Let $Y_1, Y_2, \cdots, Y_n$ be n independent random variables where each r. v. Y_i is distributed as a QBD with (m_i, p, ϕ). The likelihood function L is given by (3):

$$L=\prod_{i=1}^{n}\left[\binom{m_i}{y_i}p\left(p+y_i\phi\right)^{y_i-1}\left(1-p-y_i\phi\right)^{m_i-y_i}\right] \tag{3}$$

and, its logarithm (4) equals

$$\ell = \text{constant}+n\ln p+\sum_{i=1}^{n}\left(y_i-1\right)\ln\left(p+y_i\phi\right)+\sum_{i=1}^{n}\left(m_i-y_i\right)\ln\left(1-p-y_i\phi\right) \tag{4}$$

To derive the $C(\alpha)$ test statistic for $H_0:\phi=0$, the first and second partial derivatives of the log-likelihood function ℓ, evaluated at $\phi=0$, are needed.

All summations are from $i=1$ to n in the expressions unless stated otherwise. Differentiating the right hand-side of (4) with respect to the model parameters, and setting $\phi=0$ we get

$$\left.\begin{aligned}\left.\frac{\partial\ell}{\partial\phi}\right|_{\phi=0} &= T_1(p)=p^{-1}\sum y_i\left(y_i-1\right)-q^{-1}\sum y_i\left(m_i-y_i\right)\\ \left.\frac{\partial\ell}{\partial p}\right|_{\phi=0} &= T_2(p)=p^{-1}\sum y_i-q^{-1}\sum y_i\left(m_i-y_i\right)\end{aligned}\right\} \tag{5}$$

where $q=1-p$.

Setting the second equation in (5) to zero and solving for p yields

$$\hat{p}=\sum y_i/\sum m_i \tag{6}$$

as the maximum likelihood estimator of p under $H_0:\phi=0$.

Also, the second partial derivatives are given in (7), (8), (9)

$$\frac{\partial^2\ell}{\partial\phi^2}=-\sum\frac{y_i^2\left(y_i-1\right)}{\left(p+y_i\phi\right)^2}-\sum\frac{y_i^2\left(m_i-1\right)}{\left(1-p-y_i\phi\right)^2}, \tag{7}$$

$$\frac{\partial^2 \ell}{\partial \phi \partial p} = -\sum \frac{y_i (y_i - 1)}{(p + y_i \phi)^2} - \sum \frac{y_i (m_i - 1)}{(1 - p - y_i \phi)^2}, \tag{8}$$

$$\frac{\partial^2 \ell}{\partial p^2} = -np^{-2} - \sum \frac{(y_i - 1)}{(p + y_i \phi)^2} - \sum \frac{(m_i - y_i)}{(1 - p - y_i \phi)^2} \tag{9}$$

Setting $\phi = 0$, the above three equations are obtained in their respective orders as:

$$T_{11}(p) = -p^{-2} \sum y_i^2 (y_i - 1) - q^{-2} \sum y_i^2 (m_i - y_i) \tag{10}$$

$$T_{12}(p) = -p^{-2} \sum y_i (y_i - 1) - q^{-2} \sum y_i (m_i - y_i) \tag{11}$$

and,

$$T_{22}(p) = -np^{-2} - p^{-2} \sum (y_i - 1) - q^{-2} \sum (m_i - y_i) \tag{12}$$

Under the null hypothesis $H_0 : \phi = 0$, the Y_i's are independent binomial variates. Using the expected values of Y_i, Y_i^2 and Y_i^3 for binomial variates one can easily see that $E[T_1(p)] = 0$.

Denoting $-E[T_{11}(p)] = A_{11}(p), -E[T_{12}(p)] = A_{12}(p)$ and $-E[T_{22}(p)] = A_{22}(p)$ we can then show that

$$A_{11}(p) = (2 - 3p) q^{-1} \sum m_i (m_i - 1) + pq^{-1} \sum m_i^2 (m_i - 1), \tag{13}$$

$$A_{12}(p) = q^{-1} \sum m_i (m_i - 1), \tag{14}$$

and

$$A_{22}(p) = (pq)^{-1} \sum m_i. \tag{15}$$

Equations (13), (14), and (15) are in fact the elements of Fisher's information matrix when the null hypothesis $H_0 : \phi = 0$ is true.

To test the hypothesis $H_0 : \phi = 0$, one can use the statistic $T_1(p)$ according to Neyman's methodology [7]. Since p is unknown, we can follow Moran's suggestion [8] and use the statistic $T_1(\tilde{p})$, where $\tilde{p}$ is any root-n consistent estimator of p. The maximum likelihood estimator $\hat{p}$, given in (5) is the simplest such estimator. On substituting $\hat{p}$ in (4) and on simplifying, we get

$$T_1(\hat{p}) = (\hat{p}\hat{q})^{-1} \sum (y_i - m_i \hat{p})^2 + \hat{q}^{-1} \sum m_i (y_i - m_i \hat{p}) - \sum m_i \tag{16}$$

It may be noted that when $m_i = m_2 = \cdots = m_n = m$, the expression for $T_1(\hat{p})$ reduces to

$$(\hat{p}\hat{q})^{-1} \sum (y_i - m\hat{p})^2 - mn$$

which is like Fisher's variance test statistic. From Cox and Hinkley [10],

$$Var[T_1(\hat{p})] = A_{11}(p) - A_{12}^2(p) / A_{22}(p) \tag{17}$$

The substitution of $\hat{p}$ for p in (17) gives the functional form of the test statistic, under $H_0 : \phi = 0$, as

$$M^2 = [T_1(\hat{p})]^2 / \widehat{Var}[T_1(\hat{p})]. \tag{18}$$

The statistic M^2 (18) has an asymptotic (for $n \to \infty$) chi-square distribution

with one degree of freedom. Accordingly, the above statistic provides a $C(\alpha)$ a binomial score test which is asymptotically optimal against the quasi-binomial alternative.

3. Examples

We shall now consider two examples. In the first example the data sets are binomially distributed and the test statistic M^2 does not reject the hypothesis of a binomial distribution and in the second example the test statistic M^2 indicates that the data sets are not binomially distributed.

Example 1. Paul [11] discussed a teratological experiment in which pregnant Dutch rabbits were treated with varying doses of a compound. Each litter (group) consisted of a varying number of live fetuses in each rabbit. The number of fetuses in each litter with skeletal of visceral abnormalities were then observed. For illustration, we consider the group, treated with high dose, consisting of $n=17$ litters which gave the following observations:

$$m_i: 9\ 10\ 7\ 5\ 4\ 6\ 3\ 8\ 5\ 4\ 4\ 5\ 3\ 8\ 6\ 8\ 6$$
$$y_i: 1\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ 2\ 0\ 4\ 1\ 1\ 4\ 2\ 3\ 1$$

Since $\sum m_i = 101$ and $\sum y_i = 23$, $\hat{p} = 23/101 = 0.228$.

To test the null hypothesis H_0: The data sets are binomially distributed *i.e.* $\phi \neq 0$ against H_1: The data sets are quasi-binomially distributed *i.e.* $\phi \neq 0$, we compute the following values for (13) to (14) and apply them to (15) and (16).

$$A_{12}(\hat{p}) = \frac{570}{0.772} = 738.342,\ A_{22}(\hat{p}) = \frac{101}{(0.228)(0.772)} = 573.811,$$

$$A_{11}(\hat{p}) = \frac{2-3(0.228)}{0.772}(570) + \frac{0.228}{0.772}(4206) = 2213.84,$$

and

$$\widehat{Var}(T_1(\hat{p})) = 2213.84 - \frac{(738.342)^2}{573.811} = 1263.79.$$

Thus, from (11),

$$M^2 = \frac{(42.781)^2}{(1263.79)} = 1.448$$

Since $P_r(M^2 \geq 1.448) = P_r(X_i^2 \geq 1.448) = 0.22$, the null hypothesis cannot be rejected. Thus, we conclude that the data sets are binomially distributed with $\hat{p} = 0.228$.

4. Quasi-Binomial Regression Model

It is well known that the logistic-linear model is a basis for analyzing regression data or the data from designed experiments when the response variable is measured on the binary scale. The purpose of this section is to modify the QBD so that a finite number of concomitant variables may be included which may account for most of the sources of the extra-binomial variation.

Suppose that the i^{th} response $Y_i\left(1 \le i \le n\right)$ has the QBD given by (1). Also, let $x_{i1}, x_{i2}, \cdots, x_{ik}$ be the values of k explanatory variables associated with the response variable y_i, where the $n \times k$ matrix is of rank k. We now employ the customary logistic transformation on the binomial parameter p as indicated below"

$$p_i = \mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1},$$

where,

$$\theta_i = \ell n\left[p_i\left(1-p_i\right)^{-1}\right] = \sum_{j=1}^{k} x_{ij}\beta_j \tag{19}$$

where $\beta_1, \beta_2, \cdots, \beta_k$ in the right-hand side of (19) are the regression coefficients which are to be estimated along with the parameter ϕ.

The likelihood function will be given by

$$L = \prod_{i=1}^{n}\left[\binom{m_i}{y_i}\left(\frac{\mathrm{e}^{\theta_i}}{1+\mathrm{e}^{\theta_i}}\right)\left(\frac{\mathrm{e}^{\theta_i}}{1+\mathrm{e}^{\theta_i}}+y_i\phi\right)^{y_1-1}\left(\frac{1}{1+\mathrm{e}^{\theta_i}}-y_i\phi\right)^{m_i-y_i}\right] \tag{20}$$

Taking the log of the likelihood function (20) we get the log-likelihood function in (21)

$$\begin{aligned}\ell = &\sum \ell n\left[\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1}\right] + \sum\left(y_i-1\right)\ell n\left[\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1}+y_i\phi\right] \\ &+\sum\left(m_i-y_i\right)\ell n\left[\left(1+\mathrm{e}^{\theta_i}\right)^{-1}-y_i\phi\right]+\text{constant},\end{aligned} \tag{21}$$

where the summations are for $i=1$ to n and θ_i is defined in (19).

Differentiating ℓ, given in (21) partially with respect to $\beta_r, r=1,2,\cdots,k$, and ϕ, we have the following system of $(k+1)ML$ equations:

$$\begin{aligned}\dot{\ell}_r = \frac{\partial \ell}{\partial \beta_r} = &\sum x_{ir} - \sum \mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1} x_{ir} \\ &+\sum\left(y_i-1\right)\frac{\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-2}}{\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1}+y_i\phi}x_{ir} \\ &-\sum\left(m_i-y_i\right)\frac{\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-2}}{\left(1+\mathrm{e}^{\theta_i}\right)^{-1}-y_i\phi}x_{ir}=0,\ r=1,2,\cdots,k\end{aligned} \tag{22}$$

and

$$\dot{\ell}_\phi = \frac{\partial \ell}{\partial \phi} = \sum \frac{y_i\left(y_i-1\right)}{\mathrm{e}^{\theta_i}\left(1+\mathrm{e}^{\theta_i}\right)^{-1}+y_i\phi} - \sum \frac{y_i\left(m_i-y_i\right)}{\left(1+\mathrm{e}^{\theta_i}\right)^{-1}-y_i\phi} = 0. \tag{23}$$

The second partial derivatives are given by (where $q_i = 1-p_i$)

$$\dot{\ell}_{\phi\phi} = \frac{\partial^2 \ell}{\partial \phi^2} = -\sum \frac{y_i^2\left(y_i-1\right)}{\left(p_i+y_i\phi\right)^2} - \sum \frac{y_i^2\left(m_i-m_i\right)}{\left(1-p_i-y_i\phi\right)^2}$$

$$\dot{\ell}_{\phi r} = \frac{\partial^2 \ell}{\partial \phi \partial \beta_r} = -\sum \frac{y_i\left(y_i-1\right)p_iq_i}{\left(p_i+y_i\phi\right)^2}x_{ir} - \sum \frac{y_i\left(m_i-y_i\right)p_iq_i}{\left(1-p_i-y_i\phi\right)^2}$$

and

$$\dot{\ell}_{rs} = \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = -\sum p_i q_i x_{ir} x_{is} + \sum \frac{y_i - 1}{p_i + y_i \phi} p_i q_i (1 - 2p_i) x_{ir} x_{is} - \sum \frac{y_i - 1}{(p_i - y_i \phi)^2} p_i^2 q_i^2 x_{ir} x_{is} - \sum \frac{(m_i - y_i) p_i^2 q_i^2}{(1 - p_i - y_i \phi)^2} x_{ir} x_{is}$$

$\mathcal{L}$ for $r, s = 1, 2, \cdots, k$.

The expectations of the negatives of the above second partial derivatives would give the elements of the Fisher's information matrix. For these we use some results from [9] on inverse moments of the QBD. Thus

$$\begin{aligned} I_{\phi\phi} &= \sum E\left[\frac{Y_i^2 (Y_i - 1)}{(p_i - Y_i \phi)^2}\right] + \sum E\left[\frac{Y_i^2 (m_i - Y_i)}{(1 - p_i - Y_i \phi)^2}\right] \\ &= \sum_{i=1}^{n} \frac{m_i (m_i - 1) p_i \left[2q_i + (m_i - 1) p_i\right]}{\left[q_i - (m_i - 1)\phi\right](p_i + 2\phi)} \end{aligned} \tag{24}$$

$$\begin{aligned} I_{\phi r} &= \sum E\left[\frac{Y_i (Y_i - 1)}{(p_i - Y_i \phi)^2}\right] p_i q_i x_{ir} + \sum E\left[\frac{Y_i (m_i - Y_i)}{(1 - p_i - Y_i \phi)^2}\right] p_i q_i x_{ir} \\ &= \sum_{i=1}^{n} \frac{m_i (m_i - 1)(1 - (m_i - 3)\phi) p_i^2 q_i x_{ir}}{(p_i + 2\phi)(1 - p_i - m_i \phi) + \phi}, \quad r = 1, 2, \cdots, k \end{aligned} \tag{25}$$

$$I_{rs} = \sum_{i=1}^{n} \left[1 - \frac{(m_i - 1) p_i}{p_i - 2\phi} + \frac{(1 + \phi - m_i \phi) p_i}{q_i - m_i \phi + \phi}\right] m_i p_i q_i^2 x_{ir} x_{is}, \tag{26}$$

where $r, s = 1, 2, 3, \cdots, k$.

Equations (24), (25), (26) are the elements of Fisher's information matrix. From [12], and based on the large sample theory of the likelihood estimation, we can establish the asymptotic normality of $\hat{\Lambda} = (\hat{\beta}, \hat{\phi})$; that is

$$\mathcal{L}\left[\sqrt{n}(\hat{\Lambda} - \Lambda) \to N_{k+1}(0, \Sigma)\right]$$

in law. The large sample variance covariance matrix is given by

$$\Sigma = n \begin{bmatrix} I_{rs} & I_{r\phi} \\ I_{\phi r} & I_{\phi\phi} \end{bmatrix}^{-1}.$$

In testing hypothesis about parameters in a logit model, one generally uses large sample tests. The choice is between the likelihood ratio test and other consistent tests which are asymptotically equivalent to the likelihood ratio test under the null hypothesis [8], in contrast to the likelihood-ratio test which requires fitting the model under both the null and alternative hypotheses). Now, to test the null hypothesis $H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$, the Wald statistic given in (27) is

$$W = \frac{(\hat{\phi})^2}{AV_0(\hat{\phi})}, \tag{27}$$

In (27) $AV_0(\hat{\phi})$ is the asymptotic variance of $\hat{\phi}$, evaluated under the null hypothesis H_0. Under H_0, the statistic W has the same asymptotic (for large

samples) X_i^2 distribution as the likelihood ratio statistic. Equivalently, $H_0 : \phi = 0$ is rejected whenever the value of

$$\hat{\phi} \Big/ \sqrt{A\hat{V}_0\left(\hat{\phi}\right)} > Z_{1-\alpha},$$

where $Z_{1-\alpha}$ is the standard normal deviate for α-level of significance, and $A\hat{V}_0\left(\hat{\phi}\right)$ denotes the large sample variance of $\hat{\phi}$, under H_0, and after all other parameters are replaced by their maximum likelihood estimates.

5. Applications of the QBD Regression

1) Clinical trial results

One group of 16 pregnant female rats was fed a control diet during pregnancy and lactation and a second group of 16 pregnant female rats was given a diet treated with a chemical. Weil [13] published clinical trial data on the number m of pups alive at 4 days and the number y of pups that died at the end of 21 days lactation period for each litter. The fractions y_i/m_i for the two groups are given below:

Control: 0/13, 0/12, 0/9, 0/9, 0/8, 0/8, 1/13, 1/12,
1/10, 1/10, 1/9, 2/13, 1/5, 2/7, 3/10, 3/10.

Treated: 0/12, 0/11, 0/10, 0/9, 1/11, 1/10, 1/10, 1/9,
1/9, 1/5, 2/9, 3/7, 5/10, 3/6, 7/10, 7/7.

We apply the quasi-binomial regression model to the above data with 16 replications in each group and take

$$\theta_i = \sum_{j=1}^{2} x_{ij}\beta_j, \ i = 1, 2, \cdots, 32$$

where $x_{i1} = 1$ and $x_{i2} = 0$ when the subject is in the control group and $x_{i2} = 1$ when it is in the treatment group.

The maximum likelihood estimates of $\left(\beta_1, \beta_2, \phi\right)$ were obtained by simultaneously solving the system of equations.

$\dot{\ell}_r = 0$ and $\dot{\ell}_\phi = 0$, given in (14) and (15), with the help of NLMIX procedure in SAS (version 9.4). ML estimates are

$$\hat{\beta}_1 = -2.5135\left(0.3435\right), \hat{\beta}_2 = 0.6595\left(0.4307\right)$$

and

$$\hat{\phi} = 0.0517\left(0.0114\right)$$

The numbers in the brackets are the large sample standard deviations. Both $\hat{\beta}_1$ and $\hat{\phi}$ are highly significant (p-value < 0.001).

2) Example 2: Multiple regression (risk factors associated with COVID 19 case fatality)

The novel coronavirus disease (COVID-19) pandemic affected every country in our world and imposed tremendous strains on the world economies and the health care systems.

During the 2901-2020 year over 5000 research papers have been published

and the fundamental aim has been to understand the mechanism of spread of the virus and the main risk factors leading to associated mortality. Many of these reports on the COVID-19 pandemic suggested that the coronavirus was associated with more serious chronic diseases and mortality regardless of country and age. Other reports suggested that those with underlying comorbidities, including obesity, type 2 diabetes, heart, and kidney diseases are at high risk of infection and death. Therefore, there is a need to understand how common comorbidities and other factors are associated with the risk of death due to COVID-19 infection. Our investigation aims at exploring this relationship. Specifically, our fundamental aim is to explore the relationship between the aggregate numbers of deaths among total number of reported COVID-19 cases.

The WHO website [14] provided detailed account of the number of COVID-19 cases by country, which we accessed on December 2-2020. We included in the study the cumulative number of COVID-19 cases and the associated death counts by country as of December 2-2020. We excluded countries that had cumulative counts less than 10,000 cases. We denote the number of cases pe-country by m, and the corresponding deaths denoted by y. The data base has 112 countries, we divided them into regions according to the classification given in data source number [15]. The most referenced risk factors are:

1) X_1 = log (percentage of obese persons in a country reported in the year (2018) [17].

2) X_2 = log (population density) [18] [19] [20].

3) X_3 = log (number of people with colorectal cancer in a country reported in the year (2017) [21].

4) X_4 = log (Chronic Kidney Disease-case fatality in a country as reported in (2017) [15] [16].

Note that we used the log (factor) to stabilize the variance. The data are summarized in **Table 1**.

The histogram of y is given in **Figure 1**, showing the severe skewness in the distribution.

Figures 2-5 are the box plots of the risk factors. The plot shows that the distributions are evenly distributed among regions, except for X_3.

Table 1. Summary statistics of the COVID-19 cases (m), deaths among cases, and the four covariates.

	N	Minimum	Maximum	Mean	Std. Deviation
m	113	10,129	13,385,755	555,864.71	1,657,855.674
y	113	29	266,043	12,972.13	34,784.047
LOG_CKD_CASE_FATALITY	113	3.44	6.50	5.0616	0.51962
LOG_COLOREC_CANCER	113	3.69	12.98	8.0393	1.69123
LOG_OBESITY	113	1.28	3.63	2.8430	0.61066
LOG_POPDENSITY	113	6.57	16.65	12.3378	1.91657

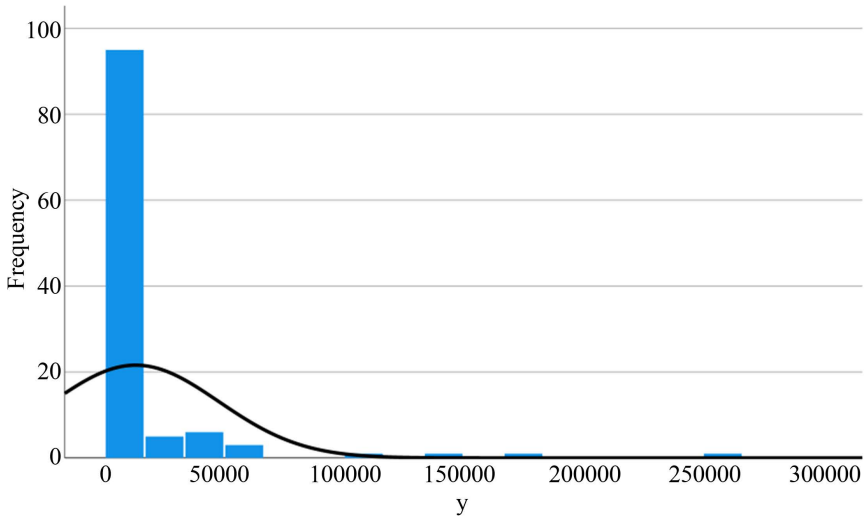


Figure 1. Histogram of the number of deaths.

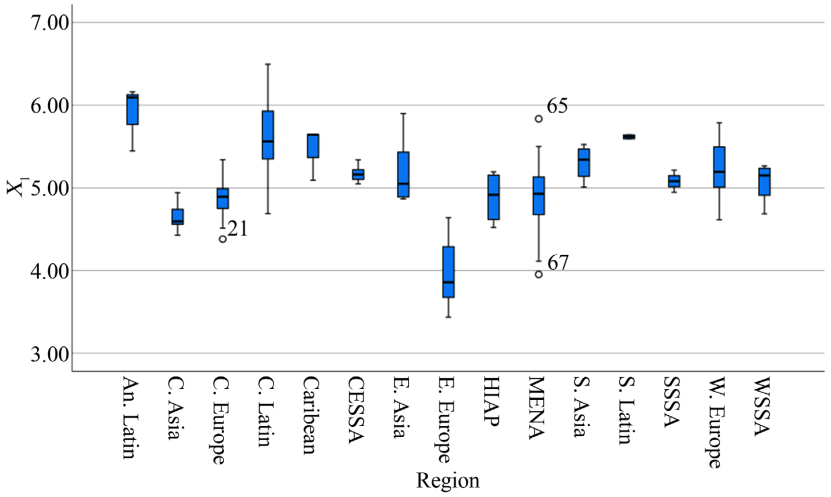


Figure 2. Box plot of X_1 by region.

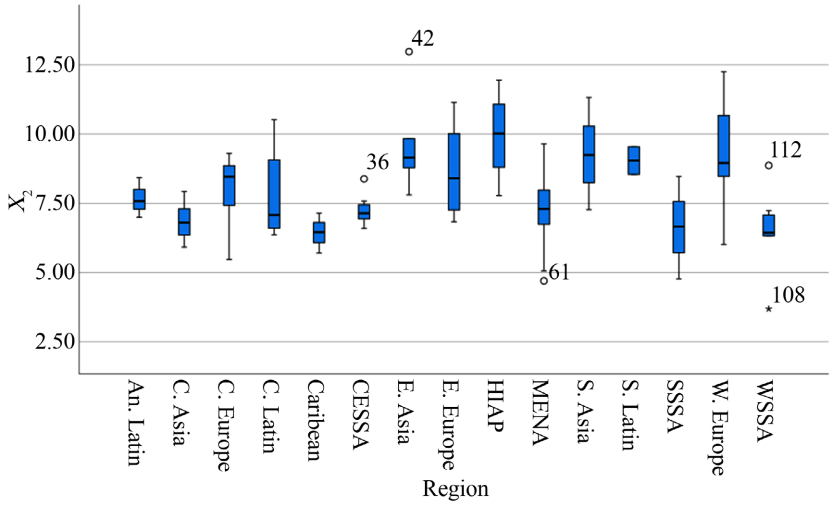


Figure 3. Box plot of X_2 by region.

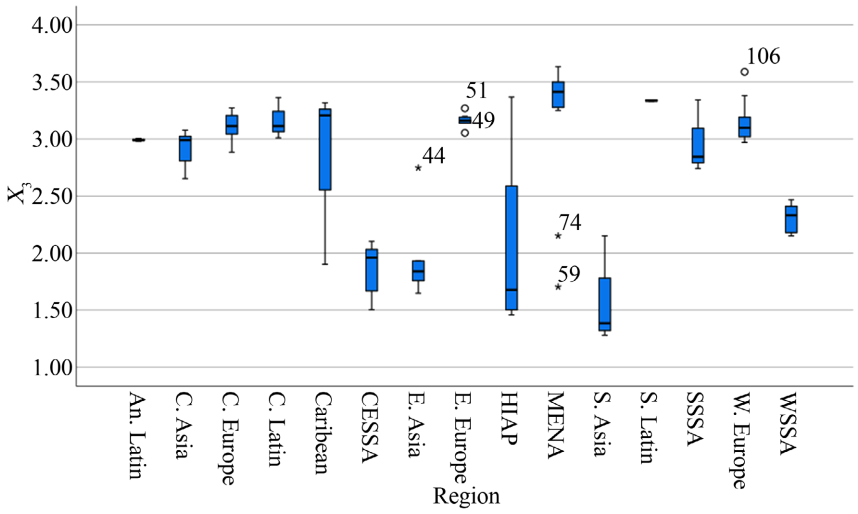


Figure 4. Box plot of X_3 by region.

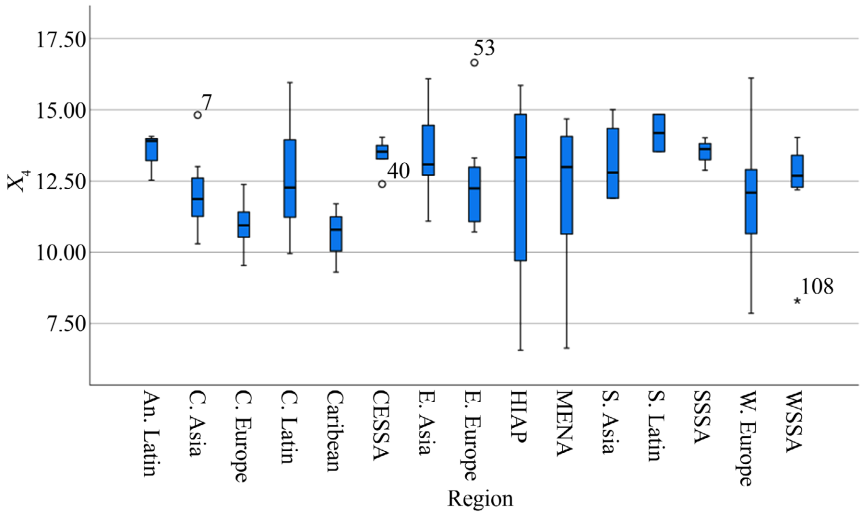


Figure 5. Box plot of X_4 by region.

We estimated the average case-fatality rate as:

$$\hat{p} = \sum y_i / \sum m_i = 0.023.$$

Moreover, the other quantities are given as:

$$A_{11}(p) = 8.49 \times 10^{19},\ A_{12}(p) = 3.51 \times 10^{14},\ A_{22}(p) = 2772 \times 10^{6},$$

$$T_1(\hat{p}) = -1.1 \times 10^{12},\ Var\left[T_1(\hat{p})\right] = 4.05 \times 10^{19}.$$

Hence $M^2 = \left[T_1(\hat{p})\right]^2 \Big/ \widehat{Var}\left[T_1(\hat{p})\right] = 29932.22$, and we therefore reject the binomial hypothesis. We used the SAS NLMIXED procedure to fit the QB regression model. The results are shown in **Table 2**.

We note that the fitting algorithm produces variance covariance matrix of the estimated regression parameters (not shown here).

The Nonlinear Mixed Model procedure (NLMIXED) is an iterative algorithm and its convergence, which can be slow, depends heavily on the starting.

Table 2. Results of the quasi-binomial regression for the COVID-19 case fatality data.

Parameter	Estimate	Standard Error	95% confidence	limits
b_0	−6.3423	0.0122	−6.3662	−6.3183
b_1	0.4180	0.0020	0.4145	0.4222
b_2	−0.0960	0.0007	−0.0974	−0.0946
b_3	0.2063	0.0010	0.2039	0.2087
b_4	0.0560	0.0007	0.0547	0.0574
phi	0.0050	0.0020	0.001	0.0090

6. Discussion

For observed data sets which exhibit variation greater than what is expected under the hypothesized model, the researchers often try to determine the sources of this phenomenon which is known as over-dispersion. There are three broad categories of such sources of over dispersion: 1) genuine or significant over-dispersion or under-dispersion which may be accounted for by generalizations of the known distribution, 2) the apparent over-dispersion is due to some outliers, which may be detected by residual analysis by some other diagnostic method, 3) poor choice of some of the explanatory variables. Therefore, it seems appropriate that one should apply a model which includes a dispersion parameter as well as a reasonable number of carefully chosen covariates and variates. The fitting of the QBD regression model can be tricky, and one may adopt one of the algorithms described in [22] and [23].

Acknowledgements

The authors thank anonymous reviewers for their constructive comments.

Conflicts of Interest

None declared by both authors.

References

[1] Altham, P.M.E. (1978) Two Generalizations of the Binomial Distribution. *Applied Statistics*, **27**, 162-167.

[2] Kupper, L.L. and Haseman, J.K. (1978) The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics*, **34**, 67-76.

[3] Williams, D.A. (1975) The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics*, **31**, 949-952.

[4] Breslow, N.E. and Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, 9-25. https://doi.org/10.1080/01621459.1993.10594284

[5] Consul, P.C. (1974) A Simple Urn Model Dependent upon Predetermined Strategy. *Sankhyā*: *The Indian Journal of Statistics*, *Series B*, **36**, 391-399.

[6] Shenton, L.R. (2006) Quasi Binomial Distribution. Wiley StatsRef: Statistics Refer-

ence Online.

[7] Neyman, J. (1959) Optimal Asymptotic Tests of Composite Statistical Hypotheses. In: Grenander, V., Ed., *Probability and Statistics*, John Wiley & Sons, New York, 13-34.

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

[9] Consul, P.C. (1990) On Some Properties and Applications of Quasi-Binomial Distribution. *Communications in Statistics—Theory and Methods*, **19**, 477-504. https://doi.org/10.1080/03610929008830214

[10] Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. Chapman and Hall, London.

[11] Paul, S.R. (1982) Analysis of Proportions of Affected Fetuses in Teratological Experiments. *Biometrics*, **38**, 361-370.

[12] Rao, C.R. (1973) Linear Statistical Inference and Applications. John Wiley & Sons Inc., New York.

[13] Weil, C.S. (1970) Selection of the Valid Number of Sampling Units and a Consideration of Their Combination in Toxicological Studies Involving Reproduction, Teratogenesis or Carcinogenesis. *Food and Cosmetic Toxicology*, **8**, 177-182. https://doi.org/10.1016/S0015-6264(70)80337-6

[14] WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/table

[15] Chueh, T.-I., Zheng, C.-M., Hou, Y.-C. and Lu, K.-C. (2020) Novel Evidence of Acute Kidney Injury in COVID-19. *Journal of Clinical medicine*, **9**, 3547. https://doi.org/10.3390/jcm9113547

[16] GBD Chronic Kidney Disease Collaboration (2020) Global, Regional, and National Burden of Chronic Kidney Disease, 1990-2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet*, **395**, 709-733. https://doi.org/10.1016/s0140-6736(20)30045-3

[17] Diabetes Prevalence (% of Population Ages 20 to 79)—Country Ranking. https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings

[18] https://openknowledge.worldbank.org/bitstream/handle/10986/32383/9781464814914.pdf

[19] 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**, Article No. 5447. https://doi.org/10.3390/ijerph17155477

[20] Population Density and Population Counts. https://data.worldbank.org/

[21] GBD 2017 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. https://doi.org/10.1016/S2468-1253(19)30345-0

[22] Boateng, E. and Abaye, D. (2019) A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, **7**, 190-207. https://doi.org/10.4236/jdaip.2019.74012

[23] Deng, J. and Lu, Q.J. (2018) Fuzzy Regression Model Based on Fuzzy Distance Measure. *Journal of Data Analysis and Information Processing*, **6**, 126-140. https://doi.org/10.4236/jdaip.2018.63008

Appendix: Flow Chart for the Manuscripts