Jurnal Matematika UNAND
Vol. **13** No. **4** Hal. 270 – 286
Edisi Oktober 2024
ISSN : 2303–291X
e-ISSN : 2721–9410
©Departemen Matematika dan Sains Data FMIPA UNAND
https://doi.org/10.25077/jmua.13.4.270-286.2024

ANALYSIS FACTORS AFFECTING COVID-19 MORTALITY USING COUNT REGRESSION

NISWATUL QONA'AH[a,b]*; T. MARTIN WALUKUSA[c]

[a] *Department of Statistics, National Cheng-Kung University, Taiwan,*
[b] *Statistics Study Program, Sebelas Maret University, Indonesia,*
[c] *Department of Statistics, Feng Chia University, Taiwan,*
email : niswatulqonaah@staff.uns.ac.id, mtlu@o365.fcu.edu.tw

Accepted May 29, 2024 Revised July 26, 2024 Published October 31, 2024

Abstract. *The 2019 novel coronavirus known as the 2019-nCoV or simply COVID-19 has been declared by the World Health organization (WHO), in first quarter of 2020, as a world pandemic and a public health emergency of international concern. Alas, many details related to the COVID-19 have remained unsolved completely. The success of government strategies in fighting the COVID-19 relays mainly on the results from epidemiological or statistical studies. Statistical models play a major role in providing reliable results based on appropriate analyses. Traditional (one-part) models, mixture models and mixed-effects models for counts are used to investigate effects of the WHO-regions and Cumulated COVID-19 cases on the outcome variable COVID-19 new deaths tolls. Overall result reveals there is a strong association between number of new deaths COVID-19 with predictors including the WHO regions and cumulated cases. Besides, models that account for the over-dispersion feature have smallest AICs and have reasonable regression model fits.*

Keywords: COVID-19, WHO, over-dispersion, Zero-inflation, Mixed-effects, Death counts

1. Introduction

The new coronavirus (2019-nCoV) is a viral pneumonia disease that is a serious illness threat to human life. The 2019-nCoV pandemic is primarily known as COVID-19 [1]. The dramatic spike in new COVID-19 numbers of cases and deaths compelled the WHO to declare the outbreak a public health emergency with international attention in January 2020 and then as a pandemic in March 2020. Since then, international organizations including the UN, WHO, etc., the private sector, the government, and the general public are all engaged in the fight against the COVID-19 pandemic at all costs. The increase in the number of deaths around the world

*Corresponding author

cannot keep anyone from being aware of the public health crisis. There is an urgent need that the government must appear strong strategies and actions to stop or reduce the threat of COVID-19, as stated by the OECD, the talking COVID-19 report (2020) entitled *"Territorial Impact of COVID-19: Managing crisis across levels of government"*. As a result, the strategy adopted by the Center for Disease Control and Prevention are more likely to produce positive results as long as they are based on reliable statistical and epidemiological results.

Statistical models such as regression models enable to investigate the prognostic associated with the COVID19 spread worldwide [2]. Count regression modelings [3] may enable us to verify the effect of the WHO-region (as ecosystem) on the COVID-19 death tolls. In public heath, there are many evidences that geographical regions do have considerable effects in the development, spread, persistence of some outbreaks [4]. The relationship between built-environment features and COVID-19 transmission risk in Hong Kong specifically investigated in [5]. In many real world applications, count response exhibits an over-dispersion. The over-dispersion can be caused by many reasons such as unobserved heterogeneity, special features in the count outcome, etc. Counts Models with an over-dispersion parameter such as zero-inflated negative binomial models and hurdle Negative binomial models [6] are more appropriate to deal with the skewed data ([7], [8]). For instance, Negative binomial have been intensively applied to infectious diseases in the presence of over-dispersion ([9], [10]). Poisson and Negative binomial regression are used to model new daily cases of COVID-19 by [11]. However, linear regression was used to predict the number of COVID-19 death India by [12]. Nevertheless, count data is generally modelled via log-linear model [13]. Similarly, in [14] statistical models used to study the risk factors that are associated with COVID-19 in Georgia. This study aims to reveal in different count model frameworks some interesting findings that may be important for decision makers who enact strategies and measures to counter the COVID-19 pandemic adventure.

2. Methods

2.1. *Simple Count Regression Models*

The traditional and popular model for count is the regular Poisson (RP). However, in other situations where RP shows limitations, Quasi-Poisson (QP) and Negative binomial (NB) are needed. Let $Y = 0, 1, \cdots, n$ be a Poisson outcome variable with probability mass function (PMF) defined as

$$P(Y = y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}, \tag{2.1}$$

where λ is Poisson mean. The PMF of a NB distribution is given as:

$$f(y : \mu, \tau) = \frac{\Gamma(y+\tau)}{\Gamma(y+1)\Gamma(\tau)} \left(\frac{\mu}{\mu+\tau}\right)^y \left(\frac{\tau}{\mu+\tau}\right)^\tau, \tag{2.2}$$

where $\mu = \lambda$, Γ is gamma function and τ is the parameter to account for the over-dispersion. The major difference between (2.2) and (2.1) is that model (2.1) assumes that the mean ($\lambda = E(Y)$) and variance ($Var(Y)$) of the outcome random variable remains equal, whereas model (2.2) assumes that $E(Y) \neq Var(Y)$. Unlike RP model, NB model has an extra parameter τ that enables the model to account for the over-dispersion feature by allowing the mean and variance to be different. When the assumption in (2.1) is violated, the results based on RP model become questionable. Nonetheless, in the presence of a zero-inflation feature, the RP, QP and NB models may fail to handle excess zeros in Y the outcome count variable.

2.2. *Zero-Inflated Models*

Zero-inflated (ZI) models are mixture models designed to handle the excessive proportion of zeros and over-dispersion in the outcome count variable [15]. In all zero-inflated (ZI) models zeros are generated from two distinct processes. Regarding the ZI Poisson (ZIP) model, the first process generates zeros from a zero mass function whereas, the second one generates random zeros from a Poisson distribution. The ZIP model is:

$$\begin{aligned} P(Y=0) &= \pi + (1-\pi)e^{-\lambda}, \\ P(Y=y_i) &= (1-\pi)\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}, y_i = 1, 2, \cdots, \end{aligned} \tag{2.3}$$

where the outcome variable y_i takes any non-negative integer value, λ is the expected Poisson count for the i^{th} individual; π is the probability of extra zeros. The mean is $(1-\pi)\lambda$ and the variance is $\lambda(1-\pi)(1+\pi\lambda)$. In practice, π and λ are functions of risk factors or predictors such that $\pi = \pi(\mathbf{X})$ and $\lambda = \lambda(\mathbf{X})$, where $\mathbf{X}$ is the matrice of predictors. Moreover, we can model the log odds of success $[\pi/(1-\pi)]$ as follows.

$$\log \frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p, \tag{2.4}$$

where, the inverse logit function of is given as:

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \tag{2.5}$$

Similarly, the count model is model by mean of log linear model such as:

$$\log(\lambda(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_q, \tag{2.6}$$

where $\lambda(\mathbf{X}) = \exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_p X_q)$. Let $d = p + q$, be dimension of $\boldsymbol{\theta}$, the regression vector of coefficient related to (2.3) becomes:

$$\boldsymbol{\theta}_{d\times 1} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T, \text{ where } \boldsymbol{\beta} = \beta_0, \beta_1, \cdots, \beta_p \text{ and } \boldsymbol{\gamma} = \gamma_0, \gamma_1, \cdots, \gamma_q.$$

Closely related to the ZIP model, but with the Zero-inflated negative binomial (ZINB) model, the second process is a negative binomial (NB) distribution. The ZINB model is:

$$\begin{aligned} P(Y=0) &= \pi + (1-\pi)\left(\frac{1}{1+\frac{\mu}{\tau}}\right)^{\tau}, \\ P(Y=y_i) &= (1-\pi)\frac{\Gamma(y_i+\tau)}{\Gamma(\tau)\Gamma(y_i+1)}\left(\frac{1}{1+\frac{\mu}{\tau}}\right)^{\tau}\left(\frac{\frac{\mu}{\tau}}{1+\frac{\mu}{\tau}}\right)^{y_i}, y_i = 1, 2, \cdots . \end{aligned} \tag{2.7}$$

When $\tau \longrightarrow \infty$ for a given mean μ, expressions (2.3) and (2.7) tend to become identical.

2.3. *Hurdle Models*

Closely related to the ZI models, but conceptually hurdle models differ from ZI models (see [15]). In particular, the hurdle models mingle a zero hurdle model $P_{zero}(Y = y)$ (right-censored at 1) and a count data model $P_{count}(Y = y)$ (left-truncated at 1), expressed as follows.

$$P_{hurdle}(Y = y) = \begin{cases} P_{zero}(Y = y), & \text{if } y = 0, \\ [1 - P_{zero}(Y = 0)]\dfrac{P_{count}(Y = y)}{1 - P_{count}(Y = 0)}, & \text{if } y > 0. \end{cases} \tag{2.8}$$

As long as one can realized conceptually realize how the excess zeros were generated, ZIs and hurdle models can serve as potential models for modelling a large proportion of zeros and the over-dispersion in outcome count data.

2.4. *Model Evaluation*

In the statistical model, we hope that the model is as simple as possible while having explanatory (parsimony). Therefore, in this part we use AIC criterion to evaluate the models. The criterion of AIC can be calculated by:

$$AIC = 2k - 2\ln(L), \tag{2.9}$$

where k is number of parameters, L is likelihood. This criterion help us to choose the model have optimal parsimony, or just the right amount of predictors needed to explain the model well. We consider the model with the smallest AIC as the best model.

3. Data

This research aimed to analyze factors affecting COVID-19 mortality. The data obtained from World Health Organization (WHO) COVID-19 global data are presented in Table 1. The dependent variable (Y) is the number of new deaths (reported in last 24 hours) 237 countries (observations), collected on January 11^{th}, 2021. The independent variables (X) consist of WHO region, Transmission Classification, and Total Cumulative Cases.

The number of new deaths (Y) and the total of cumulative cases (X_4) have a numeric (count) scale, it means that the value of these variables are positive integer. Meanwhile, WHO Region (X_2) and Transmission Classification (X_3) have a nominal (categorical) scale. WHO region consists of 6 categories, i.e., Africa, Americas, Eastern Mediteranian, Europe, South-East Asia, and Western Pacific. These regions were classified by the WHO, and we used Africa as the baseline for the further analysis.The data contain 1 missing value for the variable of Cumulative Cases population. Furthermore, we remove this missing value, so that there is 236 remaining observations.

Table 1: Scale of Variables

No.	**Variable**	**Scale**	**Source**
1	New Deaths in last 24 hours (Y)	Numeric (Count)	WHO COVID-19 global data
2	WHO Region (X_1)	Categorical (Nominal)	WHO COVID-19 global data
		Africa (baseline)	
		Americas	
		Eastern Mediterranean	
		Europe	
		South-East Asia	
		Western Pacific	
3	Transmission Classification (X_2)	Categorical (Nominal)	WHO COVID-19 global data
		Cluster of cases	
		Community-Transmission	
		Other(baseline)	
4	Total Cumulative Cases (X_3)	Numeric (Count)	WHO COVID-19 global data

4. Descriptive Statistics

4.1. *Dependent Variable Description*

The dependent variable in this analysis is New Deaths reported in last 24 hours. Figure 1 shows the distribution of New Deaths. It can be seen that the data seems follow Poisson distribution. There are so many zero values in this variable. The percentage of zero values also can be shown in Table 2.

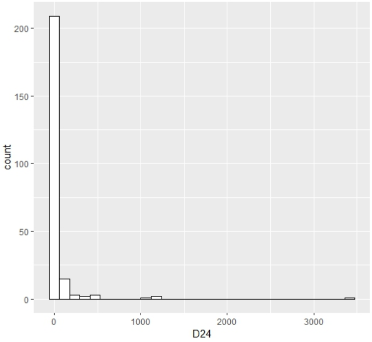


Figure. 1: Histogram of New Deaths in 24 hours

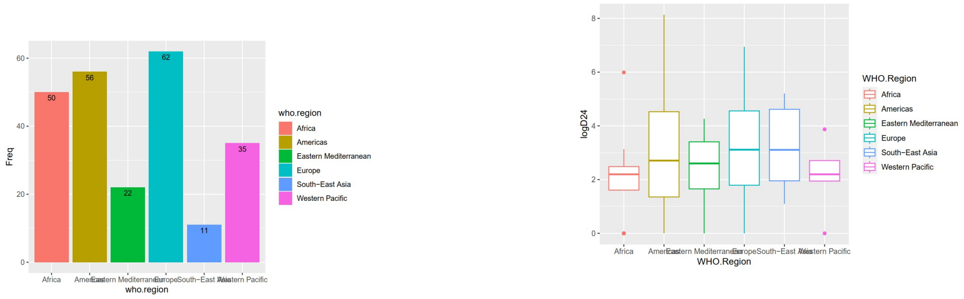
Table 2: Percentage of Zero Values

Variable	Total Zero	Total Non-Zero	Total	% Zero	% Non-Zero
New Deaths in 24 hours	123	113	236	57.12%	47.88%

According to the Table 2, we can see that there are about 57.12% of zero values in the COVID-19 daily new deaths. Figure 1 visualize the distribution of daily new deaths variable which shows that zero values exist in high frequency followed by the larger number. It can be indicated that mean and and variance of daily new deaths are not equal.

4.2. *Independent Variable Description*

- **WHO Region** (X_1)
  The first independent variable is WHO Region. The distribution of WHO Region is presented in Figure 2(a) and the relationship between WHO Region and new deaths is presented in Figure 2(b).



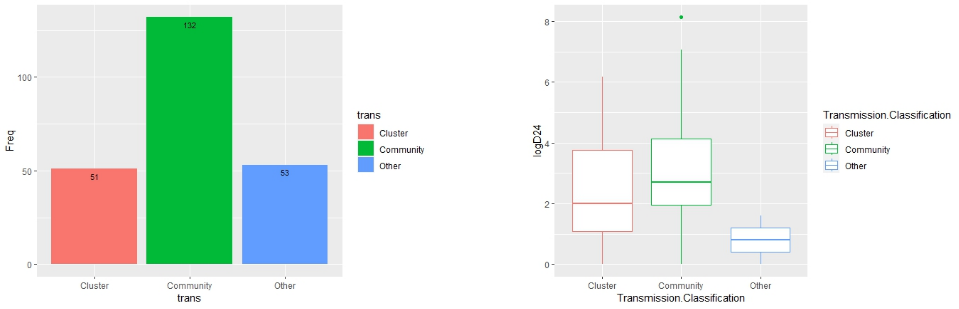
(a) Barchart of WHO region)

(b) Relationship Between WHO region and New Deaths

Figure. 2: Descriptive of WHO Region

Figure 2(a) shows that the highest frequency in the WHO region variable is Europe region (62 observations) and the lowest is south-east asia region (11 observations). The relationship between WHO region and new deaths is shown by boxplot in Figure 2(b). We apply log transformation in new deaths variable in order to get more visible boxplot and better resolution rather than using the original value. We can see that Africa region seems has the lowest distribution of new deaths almost similar with Western Pacific compared with the other regions. Therefore, we use Africa as the baseline in the next analysis.

- **Transmission Classification** (X_2)
  The second independent variable is Transmission Classification. The distribution of Transmission Classification is presented in Figure 3(a) and the relationship between Transmission Classification and new deaths is presented in Figure 3(b).



(a) Barchart of Transmission Classification)

(b) Relationship Between Transmission Classification and New Deaths

Figure. 3: Descriptive of Transmission Classification

Figure 3(a) shows that the highest frequency in the Transmission Classification variable is Community transmission (132 observations) and the lowest is cluster transmission (51 observations). The relationship between Transmission Classification and new deaths is shown by boxplot in Figure 3(b). We apply log transformation in new deaths variable in order to get more visible boxplot and better resolution rather than using the original value. We can see that 'other' category seems has the lowest distribution of new deaths compared with the other categories. Therefore, we use 'other' as the baseline in the next analysis.

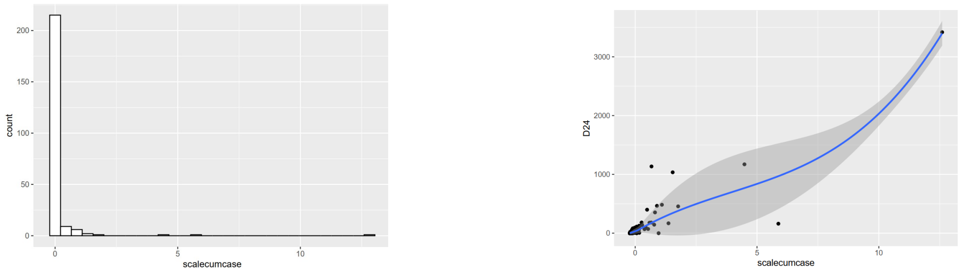
- **Total Cumulative Cases** (X_3)

  For the next analysis, we transform cumulative cases variable using scale transformation. Hence, we describe the distribution of total cumulative cases in the scale form in Figure 4(a) and the relationship between scale cumulative cases and new deaths in last 24 hours is presented in Figure 4(b).

  Figure 4(a) shows that the zero values of scale cumulative cases distribute in high frequency and (excess zero). The relationship between scale cumulative cases and daily new deaths is visualized in Figure 4(b). The scatterplot shows that more scale cumulative cases cause more new deaths (positive relationship).

4.3. *Multiple Correspondence Analysis*

The correspondence analysis (CA) is a graphical method for analyzing the associations among factors in multivariate data. It enables to uncover the pattern of associations between two, three or more categorical variables by using a contingency table. In practice, when there are more than two categories, the CA becomes multiple correspondence analysis (MCA). This study investigates the pattern and strength of the association between 3 categories including WHO region (A), transmission



(a) Histogram of Scale (Cumulative Cases)

(b) Relationship Between Scale (Cumulative Cases) and New Deaths

Figure. 4: Descriptive of Cumulative Cases

classification (B), and the number of deaths (C).

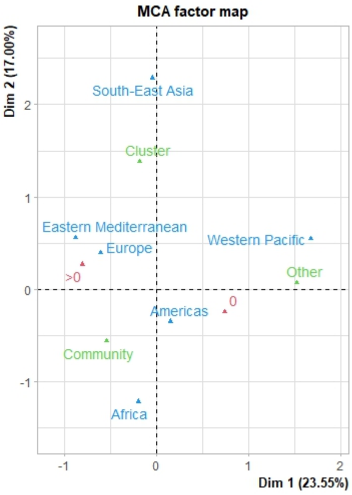


Figure. 5: Multiple Correspondence Analysis

From Figure 5, we can see that the percentage of variance explained by the first and the second principal axes are 23.55% and 17.00%, respectively. The details of numerical results obtained by performing MCA are summarized in Table 3. The inertia (squared eigenvalue) of each dimension is also provided. This result reveals that we need 6 dimensions to have a minimum 80% cumulative variance explained. It means that the variables do not have a very high correlation with each other. The association patterns from the plot of MCA are:

- People who come from the Eastern Mediterranean and Europe are more likely to have some deaths in 24 hours because of Covid 19.

- People who come from the Americas are more likely to have no deaths in 24 hours because of Covid 19.
- People who come from cluster and community transmission of Covid 19 are more likely to have some deaths in 24 hours than other transmission.

Table 3: Result of MCA

Dimension	Eigen Value	% of Variance	Cumulative % of Variance
dim1	0.6281	23.5535	23.5535
dim2	0.4533	16.9990	40.5524
dim3	0.3546	13.2987	53.8512
dim4	0.3333	12.5000	66.3512
dim5	0.3333	12.5000	78.8512
dim6	0.2466	9.2475	88.0987
dim7	0.1967	7.3750	95.4736
dim8	0.1207	4.5264	100.0000

5. Count Model

5.1. *One Part Model*

One part model consists of Regular Poisson (RP) model, Quasi-Poisson (QP) model, and Negative-Binomial (NB) model. In the one part models, we regress daily new deaths as the dependent variable on two predictors including cumulative cases which is a count variable and WHO regions which is a categorical variable that has six levels as shown in Table 1. We use Africa as reference category. When Poisson model is used, the regression model can be written as:

$$\eta_i(\mathbf{X}_i) = \log(E_i|\mathbf{X}_i) = \boldsymbol{\beta}\boldsymbol{X} = \beta_0 + \beta_1\text{CumulativeCases} + \boldsymbol{\beta_2}\textbf{WHOregions}, \quad (5.1)$$

where $\eta_i(\mathbf{X}_i)$ is the linear predictor, $\boldsymbol{X} = (1, \text{CumulativeCases}, \textbf{WHOregions})$ is the design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta_2})$ are regression coefficients (parameters). Note that β_0 is the intercept while β_1 and $\boldsymbol{\beta_2}$ are the regression slopes, with $\boldsymbol{\beta_2} = (\beta_{2,1}, \beta_{2,2}, \cdots, \beta_{2,5})$ is a parameter vector. The coefficient estimates are $\hat{\beta} = \hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\beta}}_{\mathbf{2}}$, where $\hat{\beta}_1$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$ quantify the contributions of Cumulative Cases and WHO Regions effects in (5.1). These contributions are quantified as $\exp(\hat{\beta}_1)$ and $\exp(\hat{\boldsymbol{\beta}}_{\mathbf{2}})$ for $\hat{\beta}_1$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$, respectively.

The overall result of RP, QP and NB model fits shown in Table 5 are almost similar for the coefficents but for the standard error and significance are different. The RP, QP model fits have identical regression coefficients, but they differ in terms of their standard errors. In particular, both RP and QP model fits differ from the NB model fit. In the RP and QP models, Western Pacific has the negative sign coefficient, whereas it has the possitive sign coefficient by using the NB model. However it is not significant for all models. This also indicated by Figure 2(b) which show that Western Pacific has almost similar distribution of the new deaths compared with Africa (baseline). Eastern Mediteranian has a positive and significant coefficient by using RP and NB models. For instance, based on NB model fit compared with Africa, Eastern Mediteranian is 6.7798 times more likely to be associated with COVID-19 new deaths. The predictor cumulative cases (in scale-transformed) has a positive and significant association with the number of new deaths in RP and NB models. For instance by using NB model, when the cumulative cases (in scale-transformed) increase of 1 unit, the new deaths will increase 21.3464 times. The dispersion parameter (τ) of QP and NB model are very significant. It means that the data has an overpersion issue and NB model is considerable as better model. It also supported by the smallest AIC shown in Table 5.

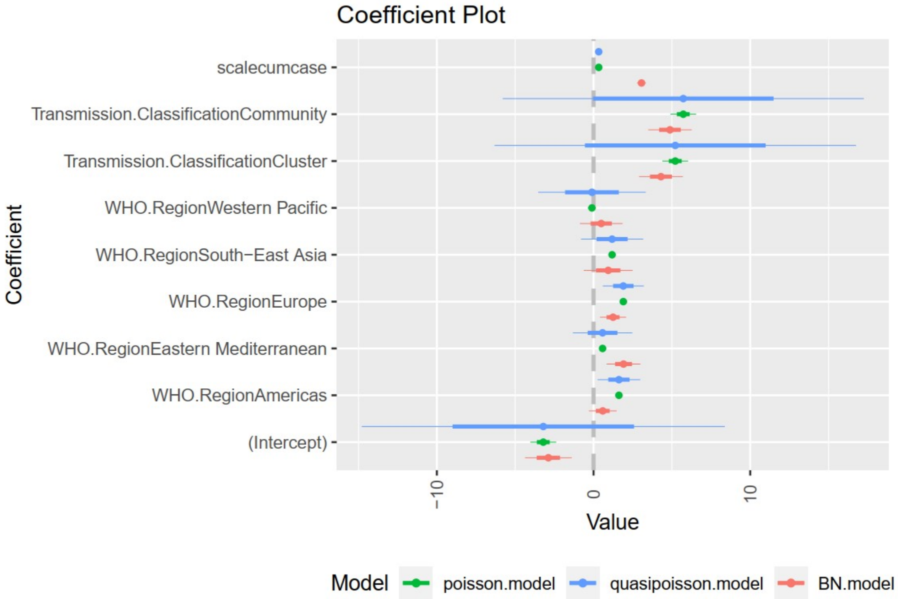


Figure. 6: Coefficient Plot of One Part Models

Figure 6 visualizes the interval coefficient estimate of the three models. We can see that the coefficient estimate is very close among the three models. The interval estimate which means the standard error of QP model is the largest, followed by BN model, and the smallest is RP model.

5.2. *Two Parts Models*

To account for excess of zeros, we apply two parts model. Two parts model consists of Zero-Inflated Poisson (ZIP) model, Zero-Inflated Negative Binomial (ZINB) model, Hurdle Poisson (HP) model, and, Hurdle Negative Binomial (HNB) model.

The ZIP model is

$$\begin{aligned} P(Y=y) &= \pi(\boldsymbol{X})I_{(y=0)} + (1-\pi(\boldsymbol{X}))I_{(y>0)}f(\lambda(\boldsymbol{X})), \\ logit(\pi(\boldsymbol{X})) &= \exp(\boldsymbol{\beta^T}\mathcal{X}_1), \text{and}\lambda(\boldsymbol{X}) = \exp(\boldsymbol{\gamma^T}\mathcal{X}_2), \end{aligned} \tag{5.2}$$

where $logit(\pi(\boldsymbol{X}))$ and $\lambda(\boldsymbol{X})$a re given in (2.5) and (2.6), respectively. We consider that $\mathcal{X}_1 = (1_n, \text{CumulativeCases}, \text{WHOregion}$ and $\mathcal{X}_2 = (1_n, \text{CumulativeCases})$ are the corresponding design matrices. Expression (5.2) enables to find $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, which is the vector of estimated regression coefficients and τ is the over-dispersion parameter. The estimation parameters of two parts model model are presented in Table 6. Table 6 reveals that the four models reached very similar results in nature. All regression coefficients in count part and zero part are very significant.

For ZIP and HP models, Americas, Europe, and South-East Asia regions have a significant positive effects on the number of new deaths in last 24 hours for all of count part models, but in ZINB and HNB models are positive and not significant. Eastern Mediteranean and Western Pacific have a significant negative effect to the new deaths by using ZIP and HP models, but possitive and not significant in ZINB and HNB models. For ZINB model fits (see Table 5) for instance, regression coefficients in $\exp\hat{\theta}$ of Europe is 1.9747. It means that living in Europe is 1.9747 times more contribute to the number of new deaths than living in Africa.

Transmission classification has a positive and significant effect to the new deaths in all of count part models. For instance, by ZINB model, regression coefficients in $\exp\hat{\theta}$ of Community Transmission is 29.8683. It means that community transmission is 29.8683 times more likely to increase the number of new deaths than other transmission.

By the four models, cumulative cases (in scale-transformed) have a positive-significant effect on the number of new deaths. For ZINB count part model (see Table 5) for instance, regression coefficients in $\exp\hat{\theta}$ of scale (cumulative cases) is 7.7853. It means that when the scale (cumulative cases) increases by 1 unit, the new deaths will increase by 7.7853 times. Meanwhile, for zero part, scale (cumulative cases) has a possitive-significant effect in HP and HNB models, whereas in ZIP and ZINB models has a negative-significant effect. Zero-Inflated and Hurdle model have different in zero model's computation. Hurdle model is compute that probability of non-zero deaths but zero-inflated model is compute probability of zero deaths. For instance, by the ZINB model, the regression coefficients in $\exp\hat{\theta}$ of scale (cumulative cases) in zero-part is 0.0000. It means that when the scale (cumulative cases) increases by 1 unit, the probability of zero deaths is 0.0000 times lower than non-zero deaths. In other words, it almost impossible to be zero deaths when the increasing of cumulative cases. By the HNB model, the regression coefficients in $\exp\hat{\theta}$ of scale (cumulative cases) in zero-part is 205369.1. It means that when the scale (cumulative cases) increases by 1 unit, the probability of non-zero deaths is 205369.1 times higher than zero deaths.

The dispersion coefficient (log Theta) is very significant, it shows that the negative binomial model should more suitable to model this data. Table 5 also provides AIC values which the minimum is for ZINB model.

6. Model Evaluation and Conclusion

Model evaluation using AIC criteria is presented in Table 4. ZINB model has the smallest AIC followed by NB, HNB, HP, ZIP, RP. The QP model does not have AIC value, however, it overcome over-dispersion issue in the model. We can see that the model using negative binomial distribution has smaller AIC than the model using poisson distribution. It obviously because the data has over-dispersion issue that the mean is different to the variance. Two part model is better than one part model, which means that in this case, it is important to model and explain the zero part in addition to modeling the count part. Based on ZINB (the best model), WHO region doesn't have significant effect to the new deaths, except europe region with a postive contribution to the new deaths. Transmission classification has a positive and singnificant effect to the new seaths. Both of community and cluster transmission are more likely contribute to the new deaths than other transmission. Total cumulative cases also has a possitive and significant effect to the number of new deaths in last 24 hours. Meanwhile, in zero part, it has a ngeative and significant effect with a very small coefficient in $\exp(\hat{\theta})$. It indicates that the probability of zero deaths is almost 0 by the increasing of cumulative cases.

Table 4: Model Evaluation

No.	**Model**	**AIC**
1	Poisson	22201.797
2	Quassi-Poisson	-
3	Zero-Inflated Poisson	16696.051
4	Hurdle Poisson	16706.758
5	Hurdle Negbin	1304.314
6	Negative-Binomial	1276.391
7	Zero-Inflated Negative Binomial	1220.633 *

'*' is the smallest value of AIC

Bibliography

[1] World Health Organization, 2020, Case definitions: suspected case; probable case; confirmed case, *Updated in Public Health Surveillance for COVID-19* Vol. **1**: 1

[2] Armitage, P., Berry, G. Mattews, J.N.S., 2020, *Statistical Methods in Medical Research*, 4th Ed., Blackwell Science

[3] Colin, C.A., Pravin, T., 2013, *Regression Analysis of Count Data*, Second Edition, Cambridge University Press

[4] Anyamba, A., Chretien, J.P., Britch, S.C., Soebiyanto, R.P., Small, J.L., Jepsen, R., Forshey, B. M., Sanchez, J.L., Smith, R.D., Harris, R., Tucker,

C.J., Karesh, W.B., Linthicum, K.J., 2019, Global Disease Outbreaks Associated with the 20152016 El Niño Event, *Scientific Reports* Vol. **9**(1): 1 – 14

[5] Huang, J., Kwan, M.P., Kan, Z., Wong, M.S., Kwok, C.Y.T., Yu, X., 2020, Investigating the relationship between the built environment and relative risk of COVID-19 in Hong Kong, *ISPRS International Journal of Geo-Information* Vol. **9**(11):

[6] Lee, J.H., Han, G., Fulp, W.J., Giuliano, A.R., 2012, Analysis of overdispersed count data: application to the human papillomavirus infection in men (HIM) Study, *Epidemiology Infection* Vol. **140**

[7] Endo, A., Abbott, S., Kucharski, A.J., Funk, S., 2020, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China, *Wellcome Open Research* Vol. **5**: 67

[8] Zhao, S., Shen, M., Musa, S.S., Guo, Z., Ran, J., Peng, Z., Zhao, Y., Chong, M.K.C., He, D., Wang, M.H., 2021, Inferencing superspreading potential using zero-truncated negative binomial model: exemplification with COVID-19, *BMC Medical Research Methodology* Vol. **21**(1): 1 – 8

[9] Lloyd-Smith, J.O., 2007, Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases, PLoS ONE Vol. **2**(2): 1 – 8

[10] Kim, T., Lieberman, B., Luta, G., Peña, E.A., 2021, Prediction Regions for Poisson and Over-Dispersed Poisson Regression Models with Applications in Forecasting the Number of Deaths during the COVID-19 Pandemic, *Open Statistics* Vol. **2**(1): 81 – 112

[11] Chan, S., Chu, J., Zhang, Y., Nadarajah, S., 2020, Count regression models for COVID-19, *Physica A* Vol. **2021** Feb 1

[12] Ghosal, S., Sengupta, S., Majumder, M., Sinha, B., 2020, Prediction of the number of deaths in India due to SARS-CoV-2 at 56 weeks, *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* Vol. **14**(4): 311 – 315

[13] Agresti, A., 2013, *An Introduction to Categorical Data Analysis*, John-Wiley and Sons, New York

[14] Khan, H., 2020, COVID-19 Epidemic Models: A Study from Georgia State in the USA, *American Journal of Biomedical Science and Research* Vol. **10**(3): 295 – 302

[15] Lukusa, T. M., Lee, S., Li, C., 2017, Review of Zero-Inflated Models with Missing Data, *Current Research in Biostatistics* Vol. **7**(1): 1 – 12

Appendix

Table 5: One Part Models

Parameter	**RP**	**QP**	**NB**
Intercept	-3.2064 ***	-3.2064	-2.8818 ***
	(0.0405)	(0.0405)	(0.0560)
	[0.4109]	[5.7907]	[0.7451]
WHO region (Americas)	1.6209 ***	1.6209 *	0.5888
	(5.0576)	(5.0576)	(1.8019)
	[0.0481]	[0.6781]	[0.4391]
WHO region (Eastern Mediteranian)	0.5764 ***	0.5764	1.9139 ***
	(1.7796)	(1.7796)	(6.7798)
	[0.0673]	[0.9484]	[0.5396]
WHO region (Europe)	1.9011 ***	1.9011 **	1.2458 **
	(6.6932)	(6.6932)	(3.4757)
	[0.0463]	[0.6526]	[0.4144]
WHO region (South-East Asia)	1.1844 ***	1.1844	0.9358
	(3.2688)	(3.2688)	(2.5492)
	[0.0704]	[0.9918]	[0.7776]
WHO region (Western Pacific)	-0.1012	-0.1012	0.4902
	(0.9038)	(0.9038)	(1.6327)
	[0.1216]	[1.7144]	[0.6780]
TC (Cluster)	5.2161 ***	5.2161	4.3034 ***
	(184.2209)	(184.2209)	(73.9535)
	[0.4092]	[5.7672]	[0.6998]
TC (Community)	5.7285 ***	5.7285	4.8789 ***
	(307.5182)	(307.5182)	(131.4829)
	[0.4087]	[5.7598]	[0.6910]
Scale(CumulCases)	0.3317 ***	0.3317	3.0609 ***
	(1.3933)	(1.3933)	(21.3464)
	[0.0021]	[0.0290]	[0.1346]
Tau (τ)		198.6274 ***	0.2490 ***
Observations	236	236	236
AIC	22202	-	1276.4

Values in brackets '( )' are the exponential of estimate, values in square brackets '[ ]' are the standard errors

p-value$<$ 0.001 is '***', p-value$<$ 0.01 is '**', p-value$<$ 0.05 is '*', p-value$<$ 0.1 is '.'

Baseline 'AFRO' of factor WHO region.

Table 6: Two Parts Model

Parameter	**ZIP**	**ZINB**	**Hurdle Poisson**	**Hurdle NB**
Count Part				
Intercept	-1.4820 **	-0.7063	1.3073 **	-0.3244
	(0.2272)	(0.4934)	(3.6963)	(0.7229)
	[0.5528]	[0.9112]	[0.4555]	[1.3370]
WHO region (Americas)	1.1709 ***	0.6741	1.1716 ***	0.5572
	(3.2251)	(1.9623)	(3.2273)	(1.7457)
	[0.0481]	[0.4539]	[0.0481]	[0.5513]
WHO region (Eastern Mediteranean)	-0.2229 ***	0.7150	-0.2382 ***	0.4368
	(0.8002)	(2.0442)	(0.7880)	(1.5478)
	[0.0674]	[0.5487]	[0.0677]	[0.6261]
WHO region (Europe)	1.0983 ***	0.6804 .	1.0983 ***	0.6274
	(2.9991)	(1.9747)	(2.9991)	(1.8728)
	[0.0465]	[0.4110]	[0.0465]	[0.4960]
WHO region (South-East Asia)	0.4076 ***	0.6175	0.4082 ***	0.5159
	(1.5033)	(1.8542)	(1.5041)	(1.6752)
	[0.0700]	[0.6955]	[0.0700]	[0.8148]
WHO region (Western Pacific)	-0.4487 ***	0.1032	-0.4331 ***	0.3681
	(0.6384)	(1.1087)	(0.6485)	(1.4450)
	[0.1219]	[0.7135]	[0.1217]	[0.9418]
TC (Cluster)	4.6280 ***	3.0792 ***	1.8389 ***	2.4427 *
	(102.3123)	(21.7402)	(6.2898)	(11.5037)
	[0.5515]	[0.8411]	[0.4539]	[1.2422]
TC (Community)	4.9853 ***	3.3968 ***	2.1964 ***	2.8467 *
	(146.2470)	(29.8683)	(8.9925)	(17.2305)
	[0.5513]	[0.8552]	[0.4535]	[1.2480]
Scale (CumulCases)	0.2879 ***	2.0522 ***	0.2879 ***	2.3386 ***
	(1.3337)	(7.7853)	(1.3336)	(10.3664)
	[0.0021]	[0.3682]	[0.0021]	[0.5366]
Log(Theta)		-0.8435 ***		-1.2876 ***
		[0.1314]		[0.3493]
Zero Part				
Intercept	-1.5943 ***	-64.61 *	1.8657 ***	1.8657 ***
	(0.2030)	(0.0000)	(6.4606)	(6.4606)
	[0.4511]	[27.3800]	[0.5166]	[0.5166]
Scale (CumulCases)	-8.8452 ***	-304.31 *	12.2326 ***	12.2326***
	(0.0001)	(0.0000)	(205369.1)	(205369.1)
	[2.4553]	[126.9000]	[2.6710]	[2.6710]
Observations	189527	189527	189527	
AIC	16696.051	1220.633	16706.758	1304.314

Values in brackets '( )' are the exponential of estimate, values in square brackets '[ ]' are the standard errors

p-value< 0.001 is '***', p-value< 0.01 is '**', p-value< 0.05 is '*', p-value< 0.1 is '.'

Baseline 'AFRO' of factor WHO region.