Journal of Social Sciences and Humanities
Vol. 5, No. 3, 2019, pp. 308-315
http://www.aiscience.org/journal/jssh
ISSN: 2381-7763 (Print); ISSN: 2381-7771 (Online)

Multilevel Study of Global Status of Road Traffic

Ramesha Jayasinghe, Roshini Sooriyarachchi*

Department of Statistics, University of Colombo, Colombo, Sri Lanka

Abstract

The field of modelling, multilevel data is a new approach. This research study examines the emerging role of modelling multilevel data in the context of analysing the factors associated with number of deaths due to road traffic accidents and type of road user which has the highest death rate. One of the objectives of this project is to perform a missing value imputation in the context of multilevel data. It was successfully obtained by performing multiple imputation using 'jomo' package in R statistical software. Generalized linear mixed models (GLMM) within the 'Glimmix' procedure of 'SAS' software was used to model the number of road deaths response and type of road user which has the highest death rate response. The study was based on data which were retrieved from the "GLOBAL STATUS REPORT ON ROAD SAFETY 2015" which was published by World Health Organization. It consists of worldwide data related to socioeconomic, health and law variables in 180 United Nations countries in six regions. This study showed that the modelling of the number of road deaths and type of road user which has the highest death rate could be adequately done using a GLMM with a Negative Binomial model and Multinomial model respectively. A cluster effect was assumed within regions. The internal and external validation showed that the model predicts well.

Keywords

Generalized Linear Mixed Model, Negative Binomial Distribution, Multinomial Distribution, SAS

Received: April 15, 2019 / Accepted: June 10, 2019 / Published online: June 24, 2019

1. Introduction

1.1. Background of Road Accidents

Globally, road traffic deaths are a major concern, because road traffic injuries have been a leading cause of mortality for many years. Currently it is the ninth leading cause of death across all age groups and is predicted to become the seventh leading cause of death by 2030. [1]. Deaths that happened due to road accidents and the vehicle types that likely to give highest death rates are quite important factors on determining the global status on road safety of a particular country. Various external factors such as socioeconomic, health and law might cause a huge impact on road accidents. According to a study that has conducted on Public Health Perspective of Road Traffic Accidents which was focused on India, the highest burden of injuries and fatalities is experienced by poor people, as they are mostly pedestrians, cyclists, and passengers of buses. [2]. Mohammadi, 2009 has conducted a study to identify the influence of age, seatbelt, time of the day and type of vehicles on road accidents in Kerman city in Iran. It was found that most of the male drivers did not use seat belts and they had a higher chance of being involved in road accidents. The author mentioned that in order to increase usage rates of seatbelts, the level of enforcement should be increased. Not only Asian countries, European countries are also more focused on proper identification of factors that would navigate towards the reduction of road traffic accidents. National and Regional Analysis of Road Accidents in Spain is a study that attempts to analyse road accidents in Spain and its provinces in time and space during 1998–2009. [3]. They mentioned that

* Corresponding author
E-mail address: roshini@stat.cmb.ac.lk (R. Sooriyarachchi)

"further progress requires efficient road safety policy based on an optimal set of measures and targets selected by the corresponding authorities, who must be committed to their compliance and achievement". Thereby, it is salient to identify the factors associated with the road deaths and the type of road user which has the highest death rate. So that necessary actions can be taken by the authorities in order to reduce the road accidents.

1.2. Objectives of the Study

This study was conducted to achieve two main objectives,

1. Missing value imputation for clustered data.
2. Determine the variables associated with the two responses. (Number of Road deaths and type of road user which has the highest death rate).

1.3. About Data

Data for this study was obtained from "GLOBAL STATUS REPORT ON ROAD SAFETY 2015" which was published by the WHO in the year 2015. The dataset consists of data from 180 countries out of a total of 195 WHO Member States, covering 6.97 billion people or 97% of the world's population.

Data was spread across two main levels. Level 1 units consists of countries and they are believed to be similar within region but vary across regions, thus it is envisaged to group the countries into regions geographically which make the level 2 units.

The initial dataset consists of 180 countries with more than 60 explanatory variables. Some of these explanatory variables contain estimated values and some of them have quite a lot of missing values. Those were removed from the dataset. There was one observation, Federal State of Micronesia, which has no records on more than half of the variables. Therefore, that observation also removed from the dataset.

One of the response variables is the type of road user according to highest death rate. The initial dataset consisted of different death percentages for four wheeled vehicles, two or three wheelers, cyclists, pedestrians and other or unspecified users. According to an article "Road accident fatalities – statistics by type of vehicles" which was published by eurostat commission, two or three wheelers, cyclists and other or unspecified users categories are merged as one category. [4]. Finally, there are only three categories, Four wheeled vehicles, pedestrians and other road users. After that the type of road user response variable was created by getting highest death category for the corresponding observations.

The final dataset consists of 179 observations with 18 explanatory variables 2 response variables and 1 cluster variable. It should be noted that this dataset consists of missing values for 8 variables as well. The detailed description of the data is given in table 1.

Table 1. Description of data.

Variable	Notation
Reported number of road deaths	Road_deaths
Type of road user which has highest death rate	Road_user
Region	Region
Income level	Inc
Universal access telephone number	tel
Emergency Training available for doctors	doc
Emergency Training available for nurses	nurse
Vital registration system exists	Vital_registration
There is National drink driving law	drink
There are Random breath testing or police check points	tests
There is National drug driving law	drug
There is National helmet law	Helmet
There is National seat-belt law	Seatbelt
Legislation on mobile phone use	Mobile
A lead agency is present	Agency
There are Policies that Promote walking & cycling	Walking & cycling
There are Policies that Promote investment in public transportation	Public_trans
There are Road audits	Audits
Population	Pop
Maximum speed on urban roads	Max speed
Registered number of vehicles	Vehicles

2. Methodology

Initially, descriptive analysis was done using mosaic plots. Mosaic plots can be obtained using R statistical software by importing 'vcd' package. [5]A preliminary analysis was then carried out to identify the association between categorical variables measured on clusters of observations. A modification of the Zhang and Boos test was used to assess the association between the categorical explanatory variables and the response variables in a univariate manner. [6]After that missing value imputation process was carried out to obtain a complete set of data. Out of different imputation techniques, Multiple Imputation (MI) is carried out and hierarchical nature of the data was also taken into consideration. [7]

An advanced analysis was done to model the number of Road deaths response and type of road user which has the highest death rate response in a univariate manner. Therefore, initially a Poisson distribution model was fitted for number of road deaths model. However the model displayed overdispersion. As a result a negative binomial distribution model was used, which displayed a better fit. A Multinomial distribution model was fitted for type of road user which has

the highest death rate response. The data were considered to be clustered within groups (Regions). The estimates were obtained using the PROC GLIMMIX procedure in SAS software.

2.1. Missing Value Imputation

The dataset consists of information regarding road accidents along with some health, law and socioeconomic variables. Therefore, it is natural to have incomplete records as some low- and middle-income countries do not have a proper information system to gather up the necessary data. In the current dataset the missing values are observed in the explanatory variables as well as in the response variables. To obtain a complete set of data multiple imputation was carried out using 'jomo' package for Multilevel Joint Modelling Multiple Imputation in R statistical software. Generalized Cochran Mantel Haenszel (GCMH) Test [8] was performed for imputed variables as well and it was identified that the significance of the variables at 20% significance level do not changed for most of the variables after doing imputation process. A liberal level of 20% was used as this considers each explanatory variable separately and variables that are slightly significant when taken separately may be quite significant when adjusted for the other variables [9].

2.2. Generalized Linear Mixed Models (GLMM)

The Generalized Linear Mixed models are an extension of generalized linear models. It includes both fixed and random effects (hence mixed models) as well as it allows response variables from different distributions.

The general linear mixed model is of the form:

$$y = X\beta + Z\gamma + \varepsilon \quad (1)$$

Where y is a (n x 1) column vector, the outcome variable; X is a (n x p) design matrix of the p predictor variables; β is a (p x 1) column vector of the fixed effects regression coefficients; Z is the (n x q) design matrix of the q random effects; γ is a (q x 1) vector of r random effects and ε is a (n x 1) column vector of errors (the part of y that is not explained by the model).

Generally, $\gamma \sim N(0, G)$ is assumed, where G is the variance-covariance matrix of the random effects. Also y is considered to have a normal distribution.

Generalized Linear Mixed models are obtained by extending this model to responses from any distribution of the exponential family. These models are of the form

$$E[Y|\gamma] = g^{-1}(X\beta + Z\gamma) \quad (2)$$

Where g (.) is a differentiable monotonic link function and g^{-1} (.) is its inverse.

The GLMM contains a linear mixed model inside the inverse link function. This model component is referred to as the linear predictor [10],

$$\eta = X\beta + Z\gamma \quad (3)$$

2.3. Fitting a Negative Binomial Model for Clustered Data

Number of road deaths is a non- negative integer value. (i.e. 0, 1, 2…). There is a possibility that the model produces negative predicted counts, if the distribution of the death count is taken as normal. Hence, count data is usually modelled using the Poisson distribution. In Poisson distribution, the mean and the variance of the response variable (Y_i) are equal. Thereby, E (Y_i) = Var (Y_i) =μ_i. However, when overdispersion is present, i.e. when E (Y_i) < Var (Y_i), the Poisson model will no longer be appropriate. The Negative Binomial model can be used in such scenarios. The Negative Binomial model is denoted by $Y_i \sim NB(\mu_i, \mu_i + \alpha \mu_i^2)$, where E ($Y_i$) = μ_i, V (Y_i) = $\mu_i + \alpha \mu_i^2$ and α controls for the overdispersion. [11, 12]

Initially, modelling was done using a Poisson model in this study. But, it was decided to use a Negative Binomial model because of the overdispersion signs of the data. [11]

2.4. Negative Binomial (NB) Regression Model

The Negative Binomial-P regression model (NB-P) is given by:

$$pr(y_i) = \left(\frac{\Gamma(y_i+\alpha^{-1}\mu_i^{2-p})}{y_i!\,\Gamma(\alpha^{-1}\mu_i^{2-p})}\right)\left(\frac{\alpha^{-1}\mu_i^{2-p}}{\alpha^{-1}\mu_i^{2-p}+\mu_i}\right)^{\alpha^{-1}\mu_i^{2-p}}\left(\frac{\mu_i}{\alpha^{-1}\mu_i^{2-p}+\mu_i}\right)^{y_i} \quad (4)$$

Where $\alpha = v_i^{-1}$ is the dispersion parameter. The mean and the variance of NB regression model are E (y_i) = μ_i and Var (y_i) = $v_i = (\mu_i + \alpha \mu_i^2)$.

Note that the most appropriate link function for the negative binomial distribution is the log link. [11]

2.5. Multinomial Regression Model

Suppose the number of categories for Y denoted by J. Let $\{\pi_1, \ldots \pi_J\}$ denote the response probabilities, satisfying $\Sigma_j \pi_j = 1$. With "n" independent observations, the probability distribution for the number of outcomes of the J types is the multinomial. It specifies the probability for each possible way the n observations can fall in the J categories.

Multinomial regression models simultaneously use all pairs of categories by specifying the odds of outcome in one category instead of another. The order of listing the categories is irrelevant, because the model treats the response

scale as nominal (unordered categories).

When the last category (J) is the baseline, the baseline category logits model with a predictor x is denoted by

$$log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x \quad (5)$$

Where j =1... J-1. The model has J-1 equations, with separate parameters for each. The effects vary according to the category paired with the baseline. [13]

3. Results from the Analysis

3.1. Best Univariate Model for Number of Road Deaths

The parameter estimates, standard errors of the estimates, degrees of freedom, t-value and the associated p-value of the final model are given in table 2. In the modelling procedure the logarithm of population was used instead of population to avoid convergence issues.

Table 2. Parameter estimates of the best model for road deaths.

	Inc	Seatbelt	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
Intercept			-9.5495	0.4364	5	-21.88	<.0001
Log (pop)			0.9801	0.02438	169	40.20	<.0001
Inc	High		-0.09825	0.1818	169	-0.54	0.5896
Inc	Middle		0.4242	0.1542	169	2.75	0.0066
Inc	Low		0				
Seatbelt		Yes	0.4717	0.1696	169	2.78	0.0060
Seatbelt		No	0				

Note that 'Low' level in Income variable and 'No' level in seatbelt variable were considered to be the reference levels. Estimation technique used was Laplace maximum-likelihood estimation method. [14]

3.2. Interpretation of the Parameter Estimates of the Best Univariate Model for Road Deaths

The fitted model is given in equation (6).

$$Log(\mu_{ij}) = -9.5495 + 0.9801\ (\log(Pop))_i - 0.09825(Inc_{High})_i + 0.4242(Inc_{middle})_i + 0.4717(seatbelt_{yes})_i \quad (6)$$

μ_{ij}= Expected number of road deaths of ith country in jth region.

The parameter estimates give the contribution of the explanatory variables to the log of the expected number of road deaths in each region. The parameter estimates which are significant at the 5 % level of significance are interpreted below.

Log (Population) (Continuous variable)

The coefficient of Log (Population) is positive, indicating that increase of the Log (Population) will lead to a increase in the count of road deaths. Suppose that Log (Population) increases by 1 unit while all other effects remained constant and the expected number of road deaths before and after this increment are μ' and μ'' respectively.

$$Log(\mu') = \beta_0 + \beta_1\ (\log(Pop) + \beta_2(Inc_{High}) + \beta_3(Inc_{middle}) + \beta_4(seatbelt_{yes}) \quad (7)$$

$$Log(\mu'') = \beta_0 + \beta_1\ ((\log(Pop) + 1) + \beta_2(Inc_{High}) + \beta_3(Inc_{middle}) + \beta_4(seatbelt_{yes}) \quad (8)$$

(8) - (7)

$$\log\left(\frac{\mu''}{\mu'}\right) = \beta_1 = 0.9801$$

$$\left(\frac{\mu''}{\mu'}\right) = \exp(\beta_1) = \exp(0.9801) = 2.6647$$

$$\mu'' = 2.6647(\mu') \quad (9)$$

This result implies that the expected number of road deaths of a particular region increases by a ratio of approximately 2.66, as a result of 1 unit increment in the Log (population) where population is given in tens of thousands.

Similarly,

The expected number of road deaths of a particular region from the middle income level is 1.53 times higher than that in low income level. There is no significant difference between the high income level compared to the low income level.

The expected number of road deaths of a particular region having the availability of seatbelt laws is 1.60 times higher than that in non-availability of seatbelt laws.

3.3. Best Univariate Model for Type of Road User

The parameter estimates, standard errors of the estimates, degrees of freedom, t-value and the associated p-value of the final model are given in table 3.

Table 3. Parameter estimates of the best model for type of road user.

Effect	road user	doc	mobile	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
Intercept	Four wheel			4.6138	2.2094	10	2.09	0.0633
Intercept	Pedestrians			7.9947	2.1993	10	3.64	0.0046
Log (pop)	Four wheel			-0.1771	0.1241	161	-1.43	0.1556
Log (pop)	Pedestrians			-0.3906	0.1377	161	-2.84	0.0051
doc	Four wheel	Yes		0.7984	0.7442	161	1.07	0.2850
doc	Pedestrians	Yes		-1.9146	0.6385	161	-3.00	0.0031
doc	Four wheel	No		0	.		.	.
doc	Pedestrians	No		0	.		.	.
mobile	Four wheel		Yes	-1.7270	0.6799	161	-2.54	0.0120
mobile	Pedestrians		Yes	-0.3368	0.7337	161	-0.46	0.6468
mobile	Four wheel		No	0	.		.	.
mobile	Pedestrians		No	0	.		.	.

Note that 'Other' level in road user response variable, 'No' level in Doc variable and 'No' level in mobile variable were considered to be the reference levels. Estimation technique used was Laplace maximum-likelihood estimation method. [14]

3.4. Interpretation of the Parameter Estimates of the Best Univariate Model for Type of Road User

$$Log\left(\frac{R_0}{R_2}\right) = 4.6138 - 0.1771\,(\log(Pop)) + 0.7984(doc_{yes}) - 1.7270(mobile_{yes}) \quad (10)$$

$$Log\left(\frac{R_1}{R_2}\right) = 7.997 - 0.3906\,(\log(Pop)) - 1.9146(doc_{yes}) - 0.3368(mobile_{yes}) \quad (11)$$

R_0 = Probability of being in four wheeled road user category.

R_1= Probability of being in pedestrian road user category.

R_2 = Probability of being in other road user category.

Using a similar method to what was used in section 3.2 the parameter estimates that are significant at 5% level of significance were interpreted as follows,

For every one unit increase in Log (Population), the log odds of being in four wheeled road user category (versus other road user category) decreases by 0.1771. Similarly, the log odds of being in pedestrian road user category (versus other road user category) decreases by 0.3906.

The odds ratio of being in four wheeled road user category (versus other road user category) is 2.2219 higher for availability of emergency training in doctors compared to non-availability of emergency training in doctors. Similarly, the odds ratio of being in pedestrian road user category (versus other road user category) is 0.1474 lower for availability of emergency training in doctors compared to non-availability of emergency training in doctors.

The odds ratio of being in four wheeled road user category (versus other road user category) is 0.1778 lower for availability of legislation on mobile phone usage compared to non-availability of legislation on mobile phone usage. Similarly, the odds ratio of being in pedestrian road user category (versus other road user category) is 0.7141 lower for availability of legislation on mobile phone usage compared to non-availability of legislation on mobile phone usage.

This study considered the joint modelling of the two response variables however, there was no improvement in fit over using two univariate models. Thus more parsimonious univariate models were fitted.

4. Validation

4.1. Internal Validation

Predicted values and categories from the fitted models were computed in order to validate the fitted models using the existing data set. (2015 data).

Road death count model.

In preliminary analysis, number of road deaths variable was categorized according to the percentiles due to the lack of methods about handling continuous data in a hierarchical nature. Here also same categorization was considered for validation purpose.

Table 4. Number of road deaths categorization.

Road death count	Road death category
Intercept	4.6138
Intercept	7.9947

4.1.1. Internal Predictive Accuracy of the Road Deaths Model

Table 5. Internal predictive accuracy of the road deaths model.

		Actual			Total
		Low	Middle	High	
Predicted	Low	56	0	8	64
	Middle	0	44	5	49
	High	4	15	47	66
Total		60	59	60	179

$$Internal\ predictive\ accuracy\ of\ the\ model = \frac{(56+44+47)}{179} \text{X } 100 = 82.12\% \sim 82\% \quad (12)$$

Therefore, the predictive accuracy of the internal data for the road deaths model is 82%.

4.1.2. Internal Predictive Accuracy of the Type of Road User Model

Table 6. Internal predictive accuracy of the type of road user model.

		Actual			Total
		Four wheeled	Pedestrian	Other	
Predicted	Four wheeled	86	4	7	97
	Pedestrian	6	27	5	38
	Other	12	10	22	44
Total		104	41	34	179

$$Internal\ predictive\ accuracy\ of\ the\ model = \frac{(86+27+22)}{179} \text{X } 100 = 75.41\% \sim 75\% \quad (13)$$

Therefore, the predictive accuracy of the internal data for the type of road user model is 75%.

4.2. External Validation

It is important to test the predictive accuracy of the developed model using a new set of data (external data). Therefore, a new set of data which is obtained from Global Status Report on Road Safety 2013 was used for this purpose. This dataset consists of 124 observations. [15]

4.2.1. External Predictive Accuracy of the Road Deaths Model

Table 7. External predictive accuracy of the road deaths model.

		Actual			Total
		Low	Middle	High	
Predicted	Low	29	1	0	30
	Middle	8	41	4	53
	High	0	6	35	41
Total		37	48	39	124

$$External\ predictive\ accuracy\ of\ the\ model = \frac{(29+41+35)}{124} \text{X } 100 = 84.67\% \sim 85\% \quad (14)$$

Therefore, the predictive accuracy of the external data for the road deaths model is 85%.

4.2.2. External Predictive Accuracy of the Type of Road User Model

Table 8. External predictive accuracy of the type of road user model.

		Actual			Total
		Four wheeled	Pedestrian	Other	
Predicted	Four wheeled	78	0	0	78
	Pedestrian	10	9	0	19
	Other	22	0	5	27
Total		110	9	5	124

$$External\ predictive\ accuracy\ of\ the\ model = \frac{(78+9+5)}{124} X\ 100\ = 74.19\% \sim 74\% \quad (15)$$

Therefore, the predictive accuracy of the external data for the type of road user model is 74%.

5. Discussion

Number of road deaths response variable is a count variable. Therefore, initially a Poisson distribution model was fitted. However the model displayed overdispersion. As a result a negative binomial distribution model was used, which displayed a better fit.

The responses of interest within the regions are assumed to be more similar than the responses of interest between regions. Therefore, regions were considered to be the cluster variable. A random intercept was used to make the same adjustment to the observations from the same region.

The significance level used in the model selection process was 0.05. To avoid convergence issues the logarithm of the population was considered, in the model selection process.

The estimates were obtained using the PROC GLMMIX procedure in SAS software.

The forward selection technique was used for variable selection when developing the models. The significance of the parameter estimates added at each stage was assessed using the p value of the Wald statistic of that particular variable.

Univariate modelling of number of road deaths.

After fitting negative binomial regression model, it was found that Log (Population), Income level and Availability of seatbelt law variables have significant association with the number of road deaths response. There is an increment in the road deaths when Log (Population) increases. When the income level is high, number of road deaths get decreased compared to low income level countries but when the income level is middle, number of road deaths get increased compared to low income level countries. There is an increment in road deaths when there is seatbelt laws compared to no seatbelt laws. Further analysis should be carried out to study this nature of the variables.

Univariate modelling of type of road users.

Multinomial logistic regression was adapted to the nominal response variable, type of road user which has the highest death rate considering type 'Other' as base level. Log (Population), Emergency training available for doctors and Legislation on mobile phone usage while driving variables have significant association with type of road user with the highest death rate response variable. Those two explanatory variables have shown 5% significant in univariate analysis also.

5.1. Akaike Information Criteria (AIC) and Bayes Information Criterion (BIC) for the Road Deaths Model

AIC's and BIC's of all two models were calculated using Laplace maximum-likelihood estimation method. Table 9 gives these results respectively.

Table 9. AIC's and BIC's of the model.

Model	AIC	BIC
Road deaths model	2584.09	2582.63
Type of road user model	274.73	272.85

Generally lower AIC and BIC values were considered as better models. [16]

5.2. Validation

The internal and external validations were carried out for the two models. The results obtained are as follows:

1. The predictive accuracy of internal data for the road deaths model is 82%.
2. The predictive accuracy of internal data for the type of road user model is 75%.
3. The predictive accuracy of external data for the road deaths model is 85%.
4. The predictive accuracy of external data for the type of road user model is 74%.

6. Conclusions from the Study

Log (Population), Income level and Availability of seatbelt law variables have significant association with the number of road deaths response. Log (Population), Emergency training available for doctors and Legislation on mobile phone usage while driving variables have significant association with type of road user with the highest death rate response variable. Internal and external predictive accuracy of the two models are high. Resources could be allocated in a more effective way to reduce the number of road deaths by using developed models. In the analysis the two responses were initially taken to be a bivariate response. However, the bivariate model had worse fit than two univariate models indicating that there was no significant correlation between the two response, namely, number of road deaths and type of road user. All the variables that were important in our study have also been shown to be important in previous similar studies. However, the presence of drinking driving law has been shown to reduce the number of deaths significantly in some studies though it is not

significant in our study. This could be due to our adjustment in the model for the population size.

References

[1] WHO, "GLOBAL STATUS REPORT ON ROAD SAFETY," World Health Organization, 2015.

[2] S. Gopalakrishnan, "A Public Health Perspective of Road Traffic Accidents," Journal of family Medicine and Primary Care, vol. 1, no. 20, pp. 144-150, 2012.

[3] A. T. Becerra, X. L. Bravo and I. F. Parra, "National and Regional Analysis of Road Accidents in Spain," Traffic Injury Prevention, vol. 14, no. 5, pp. 486-495, 2013.

[4] Eurostat, "Road accidents fatalities-statistics by type of vehicle," 2015.

[5] D. Meyer, A. Zeileis and K. Hornik, "Residual-based Shading for Visualizing (Conditional) Independence," Journal of Computational and Graphical Statistics, vol. 16, no. 3, pp. 507-525, 2007.

[6] D. B. De Silva and M. R. Sooriyarachchi, "Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R function," Journal of the National Science Foundation of Sri Lanka, vol. 40, no. 2, pp. 137-148, 2012.

[7] J. R. Carpenter and M. G. Kenward, Multiple Imputation and its Application, A John Wiley & Sons.

[8] J. Zhang and D. D. Boos, "Generalized Cochran-Mantel-Haenszel Test Statistics for correlated categorical data," Communications in Statistics - Theory and Methods, pp. 1813-1837, 1997.

[9] D. Collett, Modelling Binary Data, London: Chapman & Hall, 1991.

[10] Institute for Digital Research and Education, "Introduction to Generalized Linear Mixed Models," 10 January 2018. [Online]. Available: https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/.

[11] S. M. Fernando and M. R. Sooriyarachchi, "Bivariate Negative Binomial Modelling of Epidemiological Data," Open Science Journal of Statistics and Application, pp. 47-57, 2018.

[12] D. D. B. Trinidade, R. Ospina and L. D. Amorim, "Choosing the right strategy to model longitudinal count data in Epidemiology: An application with CD4 cell counts," Epidemiology Biostatistics and Public Health, vol. 12, no. 4, 2015.

[13] A. Agresti, Categorical Data Analysis, Florida: A John Wiley & Sons, 2002.

[14] SAS Inc. Institute, SAS/STAT 9. 2 User's Guide, Second Edition, Cary: SAS Pub, 2009.

[15] WHO, "GLOBAL STATUS REPORT ON ROAD SAFETY," World Health Organization, 2013.

[16] H. Akaike, "A New Look at the Statistical Model Identification," IEEE Transaction on Automatic Control, pp. AC–19, 716–723., 1974.