Biostatistics & Epidemiology

Joint Modelling of Two Count Variables using a Shared Random Effect Model in the presence of Clusters for Complex Data

--Manuscript Draft--

Full Title:	Joint Modelling of Two Count Variables using a Shared Random Effect Model in the presence of Clusters for Complex Data
Manuscript Number:	TBEP-2020-0037R3
Article Type:	Research Article
Keywords:	Joint Model; Generalized Linear Mixed Model (GLMM); Cluster; Spike at zero; random effects; Covid 19
Abstract:	In epidemiology it is often the case that two or more correlated count response variables are encountered. Under this scenario it is more efficient to model the data using a joint model. In addition if one of these count variables have an excess of zeros (spike at zero) the log link cannot be used in general. The situation is more complicated when the data is grouped in to clusters. A Generalized Linear Mixed Model (GLMM) is used to accommodate this cluster covariance. The objective of this research is to develop a new modelling approach which can handle this situation. The method is illustrated on a global data set of Covid 19 patients. The important conclusions are that the new model was successfully implemented both in theory and practice. A plot of the residuals indicated a well-fitting model to the data.
Order of Authors:	Roshini Sooriyarachchi
Response to Reviewers:	Please find attached

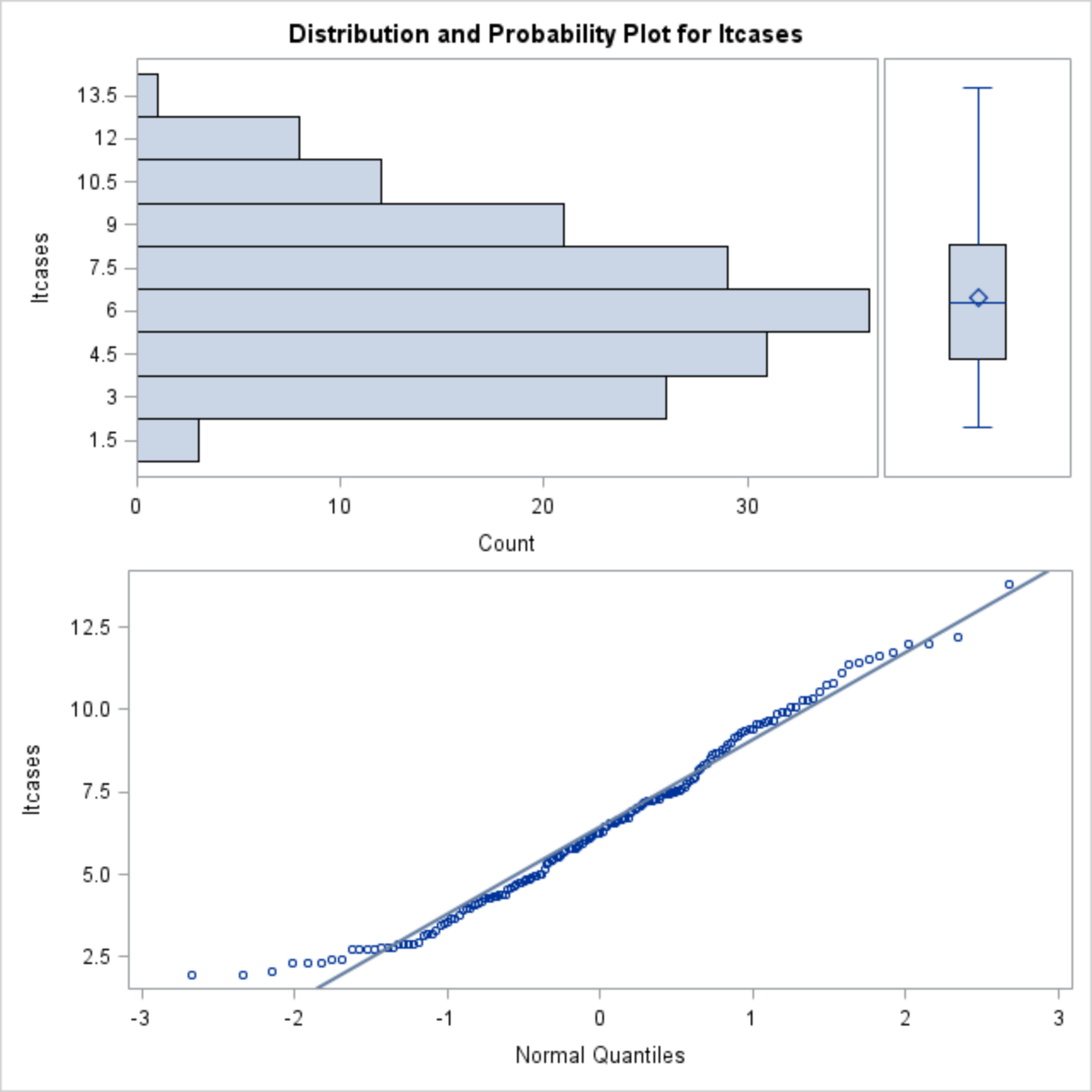

Fig 1(a) – Histogram of the log of total cases and Fig 1(b) – Normal Probability plot of the log of total cases.

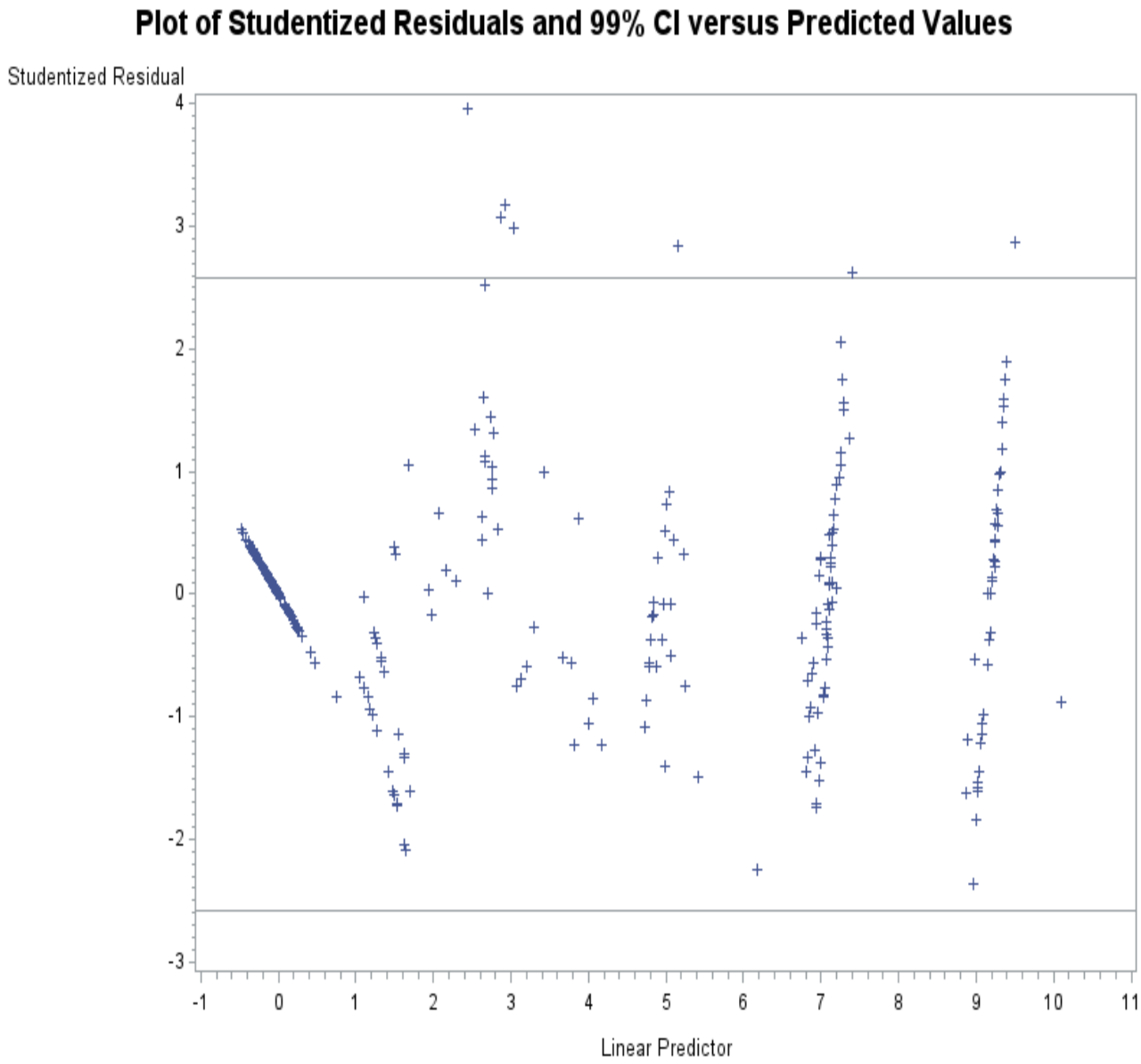

Figure 2 – Studentized Residuals versus Fitted Values

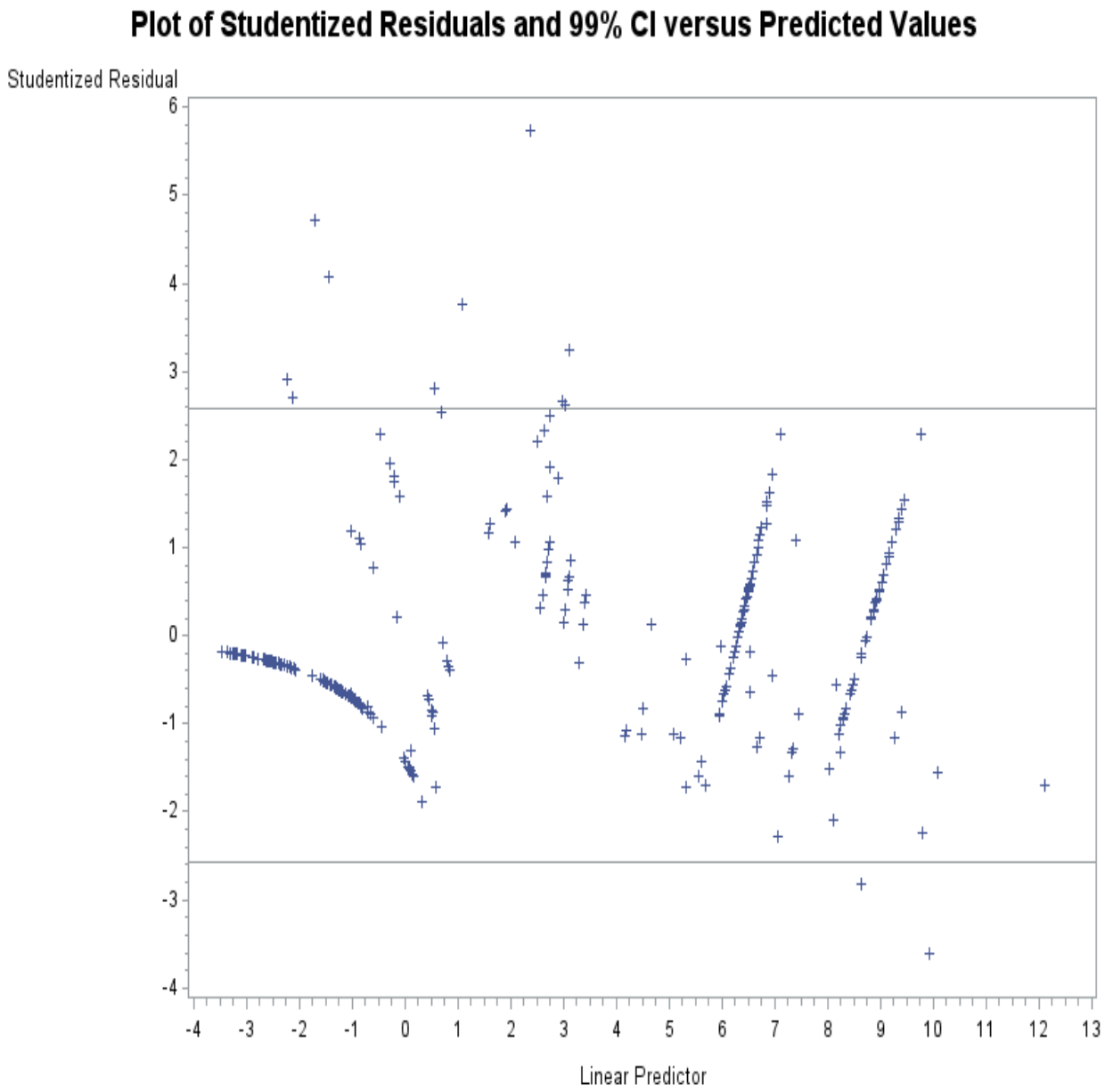

Figure 3 – Plot of Studentized Residuals versus Fitted values for the Traditional Model

Table 1 – Details of the data

Class	Levels	Values
country	144	Afghanis Albania Algeria Andorra Angola Antigua Argentin Armenia Australi Austria Azerbaij Bahamas Bahrain Barbados Belarus Belgium Belize Benin Bermuda Bhutan Bolivia Bosnia_H Botswana Brazil Brunei Burkina_ Burundi CAR Cabo_Ver Cambodia Cameroon Canada Chad Chile China Colombia Congo Croatia Cuba Cyprus Djibouti Dominica Ecuador Egypt Eswatini Ethiopia Fiji France French_P Gabon Gambia Georgia Ghana Gibralta Greece Grenada Guam Guatemal Guinea Guinea-B Guyana Haiti Honduras Hungary Iceland India Indonesi Iran Iraq Ireland Italy Jamaica Japan Jordan Kenya Korea Kosovo Kuwait Lao Latvia Lebanon Libya Lithuani Macedoni Madagasc Malawi Malaysia Maldives Mali Mauritan Mauritiu Mexico Mongolia Monteneg Morocco Mozambiq Myanmar Namibia Nepal Netherla New_Zeal Niger Nigeria Oman PNG Panama Paraguay Peru Philippi Poland Portugal Romania Rwanda Saudi Senegal Seychell Sierra_L Singapor Slovakia Slovenia Somalia Sudan Suriname Sweden Switzerl Tanzania Togo Trinidad Tunisia Turkey UK USA Uganda Ukraine Uruguay Uzbekist Venezuel Zambia Zimbabwe
region	5	AFR(African region) AR (American region) ER (European region) SEA (South East Asian Region) WPR (Western Pacific Region)
dist	2	Normal Poisson
trans_type	3	Clusters Community Sporadic

Table 2 - Comparison of Models

	Joint Model	Poisson Model	Normal Model
AIC	1043.34	749.37	538.84
Z statistic of Variance parameter (region) (Wald Test)	0.34	1	1.19
Z statistic of Variance parameter country(region) (Wald Test)	8.04	6.56	3.14

Bio Data – Marina Roshini Sooriyarachchi

Marina Roshini Sooriyarachchi, is a senior professor in the Department of Statistics, University of Colombo, Sri Lanka. She obtained her basic degree in Mathematics and Statistics (1985) and Postgraduate Diploma in Applied Statistics (1987) from the University of Colombo. She obtained her M.Sc. in Biometry (1989) and Ph.D. in Applied Statistics (1994) from the University of Reading, UK. She has over 100 research papers in peer reviewed indexed journals and conferences. She also has over 350 citations to her credit. She has won many research awards at international, national and local levels. She has 1 Ph.D. student, 1 M.Phil student and two students in the pipeline.

```
Data Covid_19;
input country$ tccases nccases tdeaths ndeaths trans_type$ dslrcase region$;
if region=0 or region=3 then delete;
if region='cases' then delete;
if region='EMR' then region='ER';
srcase=dslrcase;
if srcase=0 then X=0;
if srcase > 0 then X=1;
ltcases=log(tccases);
srcasen=srcase+1;
lsrcasen=log(srcasen);
cards;
China 84369 22 4643 0 Clusters 0                        WPR
Singapore 15222 799 14 0 Clusters 0                      WPR
Japan 13852 276 389 13 Clusters 0                        WPR
Korea 10761 9 246 2 Clusters 0                           WPR
Philippines 7958 181 530 19 Clusters 0                           WPR
Australia 6738 13 88 4 Clusters 0                        WPR
Malaysia 5851 31 100 1 Clusters 0                        WPR
New_Zealand 1126 2 19 0 Clusters 0                       WPR
VietNam 270 0 0 0 Clusters 4                     WPR
Brunei 138 0 1 0 Sporadic 9                      WPR
Cambodia 122 0 0 0 Sporadic 17                           WPR
Mongolia 38 0 0 0 Sporadic  2                    WPR
Lao 19 0 0 0 Sporadic 16                         WPR
Fiji 18 0 0 0 Sporadic 8                         WPR
PNG 8 0 0 0 Sporadic 6                   WPR
Guam 140 2 5 0 Clusters 0                        WPR
French_Polynesia 58 0 0 0 Sporadic 1                             WPR
New_Caledonia 18 0 0 0 Sporadic 26                       WPR
Italy 201505 2091 27359 382 Community 0                          ER
UK 161149 3996 21678 586 Community 0                             ER
Germany 157641 1304 6115 202 Community 0                         ER
France 125464 0 23627 366 Community 0                            ER
Turkey 114653 2392 2992 92 Community 0                           ER
Russia 99399 5841 972 105 Clusters 0                             ER
Belgium 47334 647 7331 124 Community 0                           ER
Netherlands 38416 171 4566 48 Community 0                        ER
Switzerland 29181 100 1379 27 Community 0                        ER
Portugal 24322 295 948 20 Community 0                            ER
Ireland 19877 229 1159 57 Community 0                            ER
Sweden 19621 695 2355 81 Community 0                             ER
Austria 15314 58 569 20 Community 0                      ER
Poland 12218 316 596 34 Community 0                      ER
Belarus 12208 0 79 0 Clusters 1                          ER
Romania 11616 277 650 19 Community 0                             ER
Ukraine 9866 456 250 11 Community 0                      ER
Czechia 7504 55 227 4 Community 0                        ER
Hungary 2727 78 300 9 Clusters 0                         ER
Greece 2534 0 136 0 Community 1                          ER
Croatia 2047 8 63 4 Community 0                          ER
Uzbekistan 1955 31 8 0 Clusters 0                        ER
Armenia 1932 65 30 0 Clusters 0                          ER
Iceland 1795 3 10 0 Community 0                          ER
Azerbaijan 1717 39 22 0 Clusters 0                       ER
Bosnia_Herzegovina 1588 24 62 2 Community 0                                 ER
Lithuania 1449 0 44 3 Community 2                        ER
```

Macedonia	1421	22	71	6	Clusters	0		ER	
Slovenia	1408	1	86	3	Community	0		ER	
Slovakia	1384	3	20	2	Clusters	0		ER	
Cyprus	837	15	20	0	Clusters	0	ER		
Latvia	836	18	13	0	Community	0		ER	
Albania	766	30	30	2	Clusters	0		ER	
Andorra	753	5	41	1	Community	0		ER	
Marino	553	15	41	0	Community	0		ER	
Georgia	517	6	6	0	Community	0	ER		
Montenegro	321	0	7	0	Clusters	2		ER	
Holy_See	10	1	0	0	Sporadic	0	ER		
Kosovo	790	10	22	0	Community	0		ER	
Guernsey	247	0	13	0	Community	1		ER	
Gibraltar	141	0	0	0	Clusters	2		ER	
India	31332	1897	1007	73	Clusters	0		SEA	
Indonesia	9511	415	773	8	Community	0			SEA
Sri Lanka	619	96	7	0	Clusters	0		SEA	
Maldives	245	31	0	0	Clusters	0		SEA	
Myanmar	150	4	5	0	Clusters	0	SEA		
Nepal	54	2	0	0	Sporadic	0	SEA		
Timor	24	0	0	0	Clusters	5	SEA		
Bhutan	7	0	0	0	Sporadic	6	SEA		
Iran	92584	1112	5877	71	Community	0		EMR	
Saudi	20077	1266	152	8	Clusters	0		EMR	
Egypt	5042	260	359	22	Clusters	0		EMR	
Morocco	4252	132	165	3	Clusters	0		EMR	
Kuwait	3440	152	23	1	Clusters	0		EMR	
Bahrain	2811	88	8	0	Clusters	0		EMR	
Oman	2274	143	10	0	Clusters	0	EMR		
Iraq	1928	81	90	2	Clusters	0	EMR		
Afghanistan	1827	124	60	0	Clusters	0			EMR
Djibouti	1072	37	2	0	Clusters	0		EMR	
Tunisia	975	8	40	1	Community	0		EMR	
Lebanon	717	7	24	0	Clusters	0	EMR		
Somalia	528	48	28	2	Sporadic	0		EMR	
Jordan	449	0	8	1	Clusters	1	EMR		
Sudan	318	43	25	3	Sporadic	0	EMR		
Libya	61	0	2	0	Clusters	4	EMR		
Syrian	43	0	3	0	Community	1	EMR		
Palestinian	343	1	2	0	Clusters	0		EMR	
USA	983457	22541	50492	1322	Community	0			AR
Brazil	66501	4613	4543	338	Community	0			AR
Canada	49014	1698	2766	149	Community	0			AR
Peru	28699	1182	782	54	Community	0		AR	
Ecuador	24258	1018	871	208	Community	0			AR
Mexico	15529	852	1434	83	Community	0			AR
Chile	14365	552	207	9	Community	0		AR	
Dominican_Republic	6416	123	286	4	Community	0			AR
Panama	6021	242	167	2	Community	0		AR	
Colombia	5597	218	253	9	Community	0		AR	
Argentina	4019	127	197	5	Community	0			AR
Cuba	1437	48	58	2	Clusters	0	AR		
Bolivia	1014	64	53	3	Clusters	0		AR	
Honduras	702	41	64	3	Clusters	0		AR	
Costa Rica	697	2	6	0	Clusters	0		AR	
Uruguay	620	14	15	0	Clusters	0		AR	
Guatemala	530	30	15	0	Clusters	0		AR	

Jamaica 364 59 7 0 Clusters 0 AR
El Salvador 345 22 8 0 Clusters 0 AR
Venezuela 329 4 10 0 Clusters 0 AR
Paraguay 230 2 9 0 Community 0 AR
Trinidad 116 0 8 0 Sporadic 1 AR
Bahamas 80 0 11 0 Clusters 1 AR
Barbados 80 1 6 0 Clusters 0 AR
Haiti 76 2 6 0 Clusters 0 AR
Guyana 74 0 8 0 Clusters 2 AR
Antigua 24 0 3 0 Clusters 6 AR
Belize 18 0 2 0 Sporadic 14 AR
Grenada 18 0 0 0 Clusters 2 AR
Dominica 16 0 0 0 Clusters 18 AR
Saint_Kitts 15 0 0 0 Sporadic 8 AR
Saint_Lucia 15 0 0 0 Sporadic 17 AR
Saint_Vincent 15 0 0 0 Sporadic 1 AR
Suriname 10 0 1 0 Sporadic 25 AR
Puerto_Rico 1400 11 54 1 Clusters 0 AR
Martinique 175 0 14 0 Clusters 2 AR
Guadeloupe 149 0 11 1 Clusters 4 AR
French Guiana 124 13 1 0 Clusters 0 AR
Bermuda 110 1 6 0 Clusters 0 AR
Aruba 100 0 2 0 6 Clusters 0 AR
Cayman Islands 70 0 1 0 Clusters 3 AR
Virgin_Islands 59 0 4 0 Clusters 1 AR
Curaçao 16 0 1 0 Sporadic 1 AR
Algeria 3649 132 437 5 Community 0 AFR
Cameroon 1705 84 58 2 Clusters 0 AFR
Ghana 1671 121 16 5 Clusters 0 AFR
Nigeria 1337 0 40 0 Community 1 AFR
Guinea 1240 77 7 0 Community 0 AFR
Ivoire 1183 19 14 0 Clusters 0 AFR
Senegal 823 88 9 0 Clusters 0 AFR
Niger 709 8 31 2 Clusters 0 AFR
Burkina_Faso 638 6 42 0 Community 0 AFR
Congo 491 20 30 0 Clusters 0 AFR
Mali 424 16 24 1 Clusters 0 AFR
Kenya 374 11 14 0 Clusters 0 AFR
Mauritius 332 0 10 1 Community 2 AFR
Guinea 315 57 1 0 Clusters 0 AFR
Tanzania 300 0 10 0 Clusters 4 AFR
Gabon 238 62 3 0 Clusters 0 AFR
Rwanda 212 5 0 0 Clusters 0 AFR
Congo 207 0 8 0 Clusters 1 AFR
Liberia 141 8 16 0 Clusters of cases 0 AFR
Madagascar 128 0 0 0 Clusters 2 AFR
Ethiopia 126 2 3 0 Clusters 0 AFR
Cabo_Verde 113 7 1 0 Sporadic 0 AFR
Sierra_Leone 104 5 5 1 Clusters 0 AFR
Togo 99 0 6 0 Clusters 1 AFR
Zambia 95 6 3 0 Sporadic 0 AFR
Uganda 79 0 0 0 Sporadic 1 AFR
Mozambique 76 0 0 0 Sporadic 2 AFR
Guinea-Bissau 73 0 1 0 Sporadic 1 AFR
Eswatini 71 6 1 0 Sporadic 0 AFR
Benin 64 0 1 0 Sporadic 2 AFR
Chad 52 6 2 2 Sporadic 0 AFR

```
CAR 50 8 0 0 Sporadic 0                    AFR
Eritrea 39 0 0 0 Sporadic 10                    AFR
Malawi 36 0 3 0 Sporadic 1                    AFR
South_Sudan 34 28 0 0 Sporadic 0                        AFR
Zimbabwe 32 1 4 0 Sporadic 0                    AFR
Angola 27 0 2 0 Sporadic 1                    AFR
Botswana 23 1 1 0 Sporadic  0                    AFR
Namibia 16 0 0 0 Sporadic 23                    AFR
Burundi 15 0 1 0 Sporadic  2                    AFR
São_Tomé 11 3 0 0 Sporadic  0                    AFR
Seychelles 11 0 0 0 Sporadic  22                        AFR
Gambia 10 0 1 0 Sporadic  8                    AFR
Mauritania 7 0 1 0 Sporadic 18                        AFR
Mayotte 460 27 4 0 Clusters 0                    AFR
Réunion 418 0 0 0 Clusters 1                    AFR
run;
proc Univariate Plots;
var ltcases;
run;

proc sort;
by country;
run;
data covid2;
input Country$ Pop;
cards;
Aruba 105845
Afghanistan 37172386
Angola      30809762
Albania     2866376
Andorra     77006
UAE   9630959
Argentina   44494502
Armenia     2951776
Samoa 55465
Antigua     96286
Australia   24982688
Austria     8840521
Azerbaijan  9939800
Burundi     11175378
Belgium     11433256
Benin 11485048
Burkina_Faso      19751535
Bangladesh  161356039
Bulgaria    7025037
Bahrain     1569439
Bahamas     385640
Bosnia_H    3323929
Belarus     9483499
Belize      383071
Bermuda     63973
Bolivia     11353142
Brazil      209469333
Barbados    286641
Brunei      428962
Bhutan      754394
Botswana    2254126
```

CAR 4666377
Canada 37057765
Switzerland 8513227
Chile 18729160
China 1392730000
Cote_Ivoire 25069229
Cameroon 25216237
Congo 84068091
Colombia 49648685
Comoros 832322
Cabo_Verde 543767
Costa_Rica 4999441
Cuba 11338138
Curacao 159800
Cyprus 1189265
Czech10629928
Germany 82905782
Djibouti 958920
Dominica 71625
Denmark 5793636
Algeria 42228429
Ecuador 17084357
Egypt 98423595
Spain 46796540
Estonia 1321977
Ethiopia 109224559
Finland 5515525
Fiji 883483
France 66977107
Micronesia 112640
Gabon 2119275
UK 66460344
Georgia 3726549
Ghana 29767108
Gibraltar 33718
Guinea 12414318
Gambia 2280102
Guinea-Bissau 1874309
Equatorial_Guinea 1308974
Greece 10731726
Grenada 111454
Greenland 56025
Guatemala 17247807
Guam 165768
Guyana 779004
Honduras 9587522
Croatia 4087843
Haiti 11123176
Hungary 9775564
Indonesia 267663435
India 1352617328
Ireland 4867309
Iran 81800269
Iraq 38433600
Iceland 352721
Israel 8882800
Italy 60421760

Jamaica 2934855
Jordan 9956011
Japan 126529100
Kazakhstan 18272430
Kenya 51393010
Cambodia 16249798
Kiribati 115847
St_Kitts 52441
Korea 51606633
Kuwait 4137309
Lao 7061507
Lebanon 6848925
Liberia 4818977
Libya 6678567
St_Lucia 181889
Liechtenstein 37910
Sri_Lanka 21670000
Lesotho 2108132
Lithuania 2801543
Luxembourg 607950
Latvia 1927174
St_Martin 37264
Morocco 36029138
Monaco 38682
Moldova 2706049
Madagascar 26262368
Maldives 515696
Mexico 126190788
Macedonia 2082958
Mali 19077690
Malta 484630
Myanmar 53708395
Montenegro 622227
Mongolia 3170208
Mozambique 29495962
Mauritania 4403319
Mauritius 1265303
Malawi 18143315
Malaysia 31528585
USA 364290258
Namibia 2448255
Caledonia 284060
Niger 22442948
Nigeria 195874740
Nicaragua 6465513
Netherlands 17231624
Norway 5311916
Nepal 28087871
Nauru 12704
New_Zealand 484100
Oman 4829483
Pakistan 212215030
Panama 4176873
Peru 31989256
Philippines 106651922
Palau 17907
PNG 8606316

```
Poland      37974750
Puerto Rico 3195153
Korea 25549819
Portugal    10283822
Paraguay    6956071
French_Polynesia  277679
Qatar 2781677
Romania     19466145
Russian Federation      144478050
Rwanda      12301939
Saudi 33699947
Sudan 41801533
Senegal     15854360
Singapore   5638676
Solomon_Islands   652858
Sierra_Leone      7650154
El_Salvador 6420744
San _Marino 33785
Somalia     15008154
Serbia      6982604
Sudan 10975920
Sao_Tome 211028
Suriname    575991
Slovakia    5446771
Slovenia    2073894
Sweden      10175214
Eswatini    1136191
Seychelles  96762
Syria 16906283
Chad  15477751
Togo  7889094
Thailand    69428524
Tajikistan  9100837
Turkmenistan      5850908
Timor-Leste 1267972
Tonga 103197
Trinidad 1389858
Tunisia     11565204
Turkey      82319724
Tuvalu      11508
Tanzania    56318348
Uganda      42723139
Ukraine     44622516
Uruguay     3449299
Uzbekistan  32955400
St_Vincent  110210
Venezuela   28870195 (U.S.)   106977
Vietnam     95540395
Vanuatu     292680
Samoa 196130
Kosovo      1845300
Yemen 28498687
South_Africa      57779622
Zambia      17351822
Zimbabwe    14439018
run;
proc sort;
```

```
by country;
run;
data covid_new;
merge covid_19 covid2;
by country;
lpop=log(pop);
rpop=pop**0.5;
if lpop='.' then delete;
run;
proc corr;
var srcasen tccases lsrcasen ltcases;
run;
proc glimmix data=covid_new method=laplace;
   class  country region trans_type;
   model srcasen =trans_type  X  /
                     s dist=Poisson offset=lpop;
random int/subject=region type=vc;
random int/subject=country(region) type=vc;
covtest / wald;
run;

proc glimmix data=covid_new method=laplace;
   class  country region trans_type;
   model ltcases =trans_type X  /
                     s dist=Normal;
random int/subject=region type=vc;
random int/subject=country(region) type=vc;
covtest / wald;
run;
data covid19;
   length dist $7;
   set covid_new;
   response = srcasen;
   dist     = "Poisson";
   offset=lpop;
   output;
   response = ltcases;
   dist     = "Normal";
   output;
run;
proc glimmix data=covid19 method=Laplace;
   class  country region dist trans_type;
   model response = dist dist*trans_type dist*X  /
                    noint s dist=byobs(dist);
random int/subject=region type=cs;
random int/subject=country(region) type=cs;
   output out=stat student=r pred=p;
   covtest / wald;
run;
proc gplot;
plot r*p/vref=2.58 vref=-2.58;
Title 'Plot of Studentized Residuals and 99% CI versus Predicted Values';
run;
```

Joint Modeling of Two Count Variables using a Shared Random Effect Model in the presence of Clusters for Complex Data

Abstract

In epidemiology, it is often the case that two or more correlated count response variables are encountered. Under this scenario, it is more efficient to model the data using a joint model. Besides if one of these count variables has an excess of zeros (spike at zero) the log link cannot be used in general. The situation is more complicated when the data is grouped into clusters. A Generalized Linear Mixed Model (GLMM) is used to accommodate this cluster covariance. The objective of this research is to develop a new modeling approach that can handle this situation. The method is illustrated on a global data set of Covid 19 patients. The important conclusions are that the new model was successfully implemented both in theory and practice. A plot of the residuals indicated a well-fitting model to the data.

Keywords: Joint Model, Generalized Linear Mixed Model (GLMM), Cluster, Spike at zero, random effects, Covid 19

1. Introduction

1.1 Background

In epidemiological and health studies often several correlated count responses are encountered [4] [6]. This type of data is often found to occur within groups (clusters). In this case, it is more efficient to model these count responses jointly rather than model each response separately. Fernando and Sooriyarachchi [4] use Generalized Linear Mixed Modeling (GLMM) with random cluster effects for this scenario which is now quite well developed. In this research, a more difficult

problem involving four complexities that can be encountered with count data is considered. The first is the presence of a huge number of zero counts resulting in an enormous spike at zero [15]. The zero-inflated Poisson is an option for modeling such data but in the context of joint models could have convergence problems [2]. The second problem is when the counts are non-zero but huge. Neither the Poisson nor the Negative Binomial converges especially in the case of joint models [9]. The third situation is the presence of a negative correlation between responses [19]. Even though there are indirect methods developed in the literature to handle this, these have been developed only for binary, ordinal, and continuous data [19]. The fourth problem is related to cluster-specific covariates which occur at different levels of the hierarchy [5]. Rizopoulos [18] has dealt with this for survival and normal responses but not for two count variables. In the case of the first problem, the literature provides a solution for univariate models [15]. In the case of the second problem log transforming the counts and modeling the transformed values as Normal often works [13, 16]. In the situation of the third problem GLMM with common random effects has been seen to work for survival and count joint responses [19]. In the final problem multilevel modeling for high dimensional problems has been seen to work for bivariate binary problems [5].

1.2 Objectives

The primary objective of this research is to develop a new model for the scenario described in section 1.1. The secondary objective is to apply the model to a suitable set of data.

1.3 Brief description of Methods

The joint model is developed for a Poisson-Normal joint distribution. A Generalized Linear Mixed Model with Maximum Likelihood Estimation and Laplace approximation for the marginal log-likelihood was used for this purpose [7]. Two random effects to incorporate two different

cluster effects were used. These random effects are used to join the two responses. The covariance structure used was compound symmetry. It was assumed that both responses have the same random variance.

1.4 Data for the example

The data for the example is related to Covid 19 and the relevant details were obtained from the website of the Epidemiology Unit of Sri Lanka [10]. The two, count responses pertained to the expected number of days elapsed after the last corona case and the total number of corona cases. The former had a spike at zero and the latter had very large counts. There was one explanatory variable, namely, the type of transmission of the cases. There were 144 countries and this database was merged with the population size database given by the UN website [11].

1.5 Structure of the Paper

Section 1 consists of an introduction to the problem, objectives, a brief description of methods, and an explanation of the data. Section 2 is made up of a literature review. In section 3 the new model is developed. Section 4 gives the example and the discussion consists of section 5 followed by references.

2. Literature Review

2.1 Joint modeling of two count variables

In epidemiology, often two correlated count variables are encountered, such as the incidence of the disease and the platelet count in dengue, the incidence of the disease and white blood cell counts in Japanese encephalitis, the incidence of Leptospirosis and the count of serovar-specific

antibodies to name a few examples. As most of these diseases also depend on the climate and thus on the geographical region the region happens to be a cluster variable. Many zero counts are possible in regions where the weather is not conducive to the disease. Thus resulting in a SAZ. The second variable is usually related to huge counts and the weather parameters are cluster level variables. In the first example given the correlation between the two counts is negative with high dengue counts being related to low platelet counts. Under this epidemiological scenario the major characteristics are two count variables, cluster variation, one count variable with a SAZ and the other count variable with huge counts, correlated counts with the more challenging being that the counts are negatively correlated. Thus each of these characteristics are reviewed in the next sections.

Kochelerkota and Kochelerkota [14] and Ophem [17] give a detailed literature review of this situation. Gurmu and Elder [6], discuss the joint modeling of two count variables when these variables are negatively correlated. They mention that in this scenario the bivariate Poisson and the bivariate negative binomial cannot be used to model the data. They consider a two-factor framework where dependence between the count variables is modeled using correlated unobserved heterogeneity components. Their article uses semi-parametric methods for the estimation of a mixture of count models that include negatively correlated counts. Aitchinson and Ho [1] particularly discuss the case of negatively correlated count variables where they use a Poisson-Lognormal mixture to model two count variables with a negative correlation.

However, their methods [1] [6] do not take in to account the adjustments for clustering, a spike at zero, huge counts, and cluster-level covariates.

Hapugoda and Sooriyarachchi [8] develop a joint model using a single shared random effect to model survival and count responses combining the discrete time hazard model and Poisson model.

This method does not take in to account high dimensional data and has only one random effect. The model [8] though considering clustered data is a joint model for binary and count data and does not accommodate an excess of zeros nor huge counts. It can, however, handle a negative correlation.

Sunethra and Sooriyarachchi [19] develop joint models using two separate random effects to model survival and count data which are negatively correlated. They use the lognormal distribution to model the survival data and the Poisson distribution to model the count data. However, they [19] do not consider the case of excess zeros nor huge counts.

Wickremarachchi (unpublished B.Sc. thesis, 2017) [20] develop a bivariate binomial model in the presence of clusters. This is modeled using multilevel modeling. Here a different technique to Hapugoda and Sooriyarachchi [8] and Sunethra and Sooriyarachchi [19] is used to model the correlation within clusters. It is another option for the cluster scenario. This technique is multilevel modeling. This method [20] does not accommodate negative correlation, excess zeros, high dimensional data nor huge counts.

2.2 Use of Random Effect Models for Joint Model Development

A pioneer of random effect models for joint model development is Rizopoulos [18]. He developed joint models for survival and repeated measures responses considered to be normally distributed. Sunethra and Sooriyarachchi [19] give a detailed review of this situation. They consider both the case of positive, and negative correlation between two response variables. For the case of positive correlation, they discuss shared random effect models and for negatively correlated responses they consider separate random effect models. However, his work [18] is purely for joint survival and

normal longitudinal models. He has not considered count responses and so his work does not discuss an excess of zeros and huge counts.

Similar to Sunethra and Sooriyarachchi [19], in this paper, the author examined the joint modeling of two count variables using shared random effects with two random effects at two different levels. The method used in this research is a technique that combines the methods of Sunethra and Sooriyarachchi [19] with that of Aitchinson and Ho [1]; Lorenz, Jenkner et al. [15], and G. Fernando and Sooriyarachchi [5].

This current approach uses two shared random effects as in [19], uses appropriate methods to incorporate negative correlation for Poisson-Lognormal mixtures as in [2], adjusts for a spike at zero as in [15], and extends the problem to 3-dimensional data as in [5]. Apart from combining these methods, further extensions have been made by looking at significance tests for the random effects, extensions were also made to [19] and [5] where the survival and count joint model and the univariate binary model respectively were changed to a joint count model. Finally [15] was extended from a univariate model to a bivariate model.

2.3 The case of a peak (spike) at zero

When in addition to negative correlation if one of the count variables has a spike at zero, neither the zero-inflated Poisson nor the zero-inflated negative binomial usually converge [2]. Lorenz, Jenkner et al. [15] introduced four methods to handle this case. They discussed the case occurring often in Epidemiological and Clinical Research where variables are often semi-continuous with several patients often having exposure zero and a continuous distribution among those exposed. This is referred to as a spike at zero (SAZ). They illustrated their procedures on a German Breast Cancer Study Group data (GBSG). Their method involves dichotomizing the SAZ variable into a

binary variable (X) with the two levels relating to the zeros and non-zeros. Then the binary information is combined with the positive continuous variables. This information is combined into one variable in the standard technique to give by default the linear component. Using the approach of Lorenz, Jenker et al. [15] in our study we take the first count variable (srcasen) split into two, one a binary variable for zeros and non-zeros and the other a continuous srcasen variable and the two srcasen variables are treated as one prognostic factor in the model. Both variables are tested jointly in the model.

Expectation [Response] (combined variable) = Exp ($\beta Z + \gamma X$) where Z consists of the other explanatory variables and the intercept (for intercept $\beta=1$) and X consists of the binary explanatory variable. There is no intercept in this model. If the Response is a count it can be taken as having a Poisson distribution with a log link. Here β and γ are the unknown coefficients of Z and X respectively.

The authors of this paper [15] do not consider hierarchical data in the form of clusters nor huge counts. Also, they only consider the univariate case.

The authors [15] have also mentioned about some discrepancies in the method and the way of getting over these discrepancies. According to them, "Modeling such SAZ variables is challenging and there are both statistical problems and problems conserning interpretation arising from this situation. Readers are referred to paper [15] for more information on the overcoming of these problems.

2.4 The case of cluster effects

When the data is hierarchical we refer to this as multilevel data. We consider here the case where there are three levels. The third level is a large cluster within which lies a small cluster referred to

as the second level within which lies two correlated count responses referred to as the first level. If the correlation within a cluster is significant then the model cannot be fitted using standard models. This correlation has to be taken in to account. [5]

2.5 Methods used in this research

Here a shared random effects joint model is used to model two negatively correlated counts using the Poisson-Normal mixture. A spike at zero is taken into consideration also. The model is developed within the framework of hierarchical models. The method used in this research is a technique that combines the methods of Sunethra and Sooriyarachchi [19] with that of Aitchinson and Ho [1]; Lorenz, Jenkner et al. [15], and G. Fernando and Sooriyarachchi [5]. This combination is not found in the literature and therefore, is a novel development.

3. Theory

Consider the method of Hapugoda and Sooriyarachchi [8] where the Procedure Glimmix in SAS 9.4 is used to fit a shared parameter joint model to a survival and count response. The model fitted is a Generalized Linear Mixed Model (GLMM) with one random effect representing a single cluster. This method will be modified for this situation. Here we use a Poisson model with an adjustment for a spike at zero for one count variable and the other count variable is log-transformed and modeled as a Normal response variable [1][16]. The correlation matrix is modeled as of type compound symmetry. The method of estimation used was the Maximum Likelihood with Laplace Approximation of the marginal log-likelihood [7]. In this research, a shared parameter joint model

for joining the two count responses is fitted. The model fitted is a Generalized Linear Mixed Model (GLMM) with two random effects representing two sets of clusters.

3.1 Poisson Regression Model for clustered data

Suppose y_{ij1} is the first observed count for the ith small cluster in level 2 within the jth third level cluster where y_{ij1}~ Poisson (μ_{ij1}) and μ_{ij1} is the mean of the Poisson distribution for the 1st observation of level 1 within the ith 2nd level unit within the jth third level unit. E_{ij1} is the Expected count or offset [12]. The Z_{ij} are the predictors and $\beta_{0ij}= \beta_0+u_{0ij} + v_{0j}$ is the random intercept where β_0 is a fixed component and u_{0ij} is a random component for cluster-level 2 (intercept) and v_{0j} is a random component for cluster-level 3. Let X_{ij} be the binary variable to adjust for the spike at zero. Then a three-level random intercept Poisson Regression model can be given by

$\log(\mu_{ij1}) = \log(E_{ij1}) + \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$. (1)

To classify a count as either zero or not, a binary variable X is added to the model. It is assessed in a two-stage procedure to determine whether the binary variable and/or the continuous function for the positive part is required for a suitable fit [15].

3.2 The Normal model for log count clustered data

Let y_{ij2} be the log-transformed second count variable for the ith small cluster in level 2 within the jth third level cluster where y_{ij2} ~ Normal (μ_{ij2}, σ_{ij2}^2) where μ_{ij2} is the mean of the Normal distribution for the 2nd observation of level 1 within the ith 2nd level unit within the jth third level

unit and σ_{ij2}^2 is the variance of the Normal distribution for the 2nd observation of level 1 within the i[th] second level unit within the j[th] third level unit. The Z_{ij} are the predictors and $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ where u_{0ij} and v_{0j} are as in section 3.1. Then a three-level random intercept Normal Regression model can be given by

$\mu_{ij2} = \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$

(2)

3.3 The joint model for clustered data

The responses of analysis are Yij_1 (Poisson – Count 1) and Yij_2 (Normal – Log transformed count 2). The suffixes i and j, are as defined before. Variables that impact $Y = (Y_1, Y_2)$ are the explanatory variables (X_{ij} and Z_{ij}) as defined before i=1,2,…,I where I is the number of small clusters and j=1,2,…., J where J is the number of large clusters. To formulate a joint model, Generalized Linear Model (GLM) can be used to form marginal models for each response by considering mean $E(Y_{ijk}/X_{ij}, Z_{ij})$ and variance $Var(Y_{ijk}/X_{ij}, Z_{ij})$ where k=1,2. The approach to link the responses is by structuring a covariance matrix $Var(Y_{ijk}/X_{ij}, Z_{ij})$ to include potential correlations.[16]. The random effects are assumed to be the same for both responses so this is a shared random-effects model.

In GLM $l_k\left(E\left(Y_{ijk}/X_{ij}, Zij\right)\right) = X_{ijk}'\beta_k + Z_{ijk}'\gamma_k$, k=1,2 where i,j denotes each record from each i[th] small cluster within each j[th] large cluster and l_k is the link function. Here, $l_1(u)$ is the log link and $l_2(u)$ is the identity link. GLIMMIX is used to estimate two marginal models jointly.

A structural formulation of the model is given as:

$l_1(Y_{ij1}') = \text{Log}(\mu_{ij1}) = \log(E_{ij1}) + \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$. (3)

and

$l_2(Y'_{ij2}) = (\mu_{ij2}) = \vartheta_{0ij} + \vartheta X_{ij} + \delta X_{ij}$ where $\vartheta_{0ij} = \vartheta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$ (4)

Here k=1,2 and i=1(1) n_j and j=1(1)m where n_j is the number of small clusters within big cluster j and m is the number of big clusters.

For simplicity, we assume that both sets of random effects are the same (u_{0ij} and v_{0j}) and have the same variance(σ_u^2 *and* σ_v^2 respectively). The joint model variance-covariance matrix, Var-Cov (Y_{ij1}, Y_{ij2}) is of the form $\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ where:

$\sigma_2^2 = \sigma_u^2 + \sigma_v^2 + \sigma^2$ and σ^2 is the variance of the error term in the regression.

Here σ_1^2 can be derived using the methods of Sunethra et al. (2020) [19]. The correlation between the Y_{ij1} and Y_{ij2} is taken to be ρ_{12}. It is assumed that the u_{0ij}'s are independent of the v_{0j}'s. GLIMMIX will structure the variance-covariance matrix of Y = (Y_1, Y_2) as in Hapugoda et al. [7]. The development of the joint log-likelihood and thereby the joint model is given in detail in Sunethra and Sooriyarachchi [19].

4. Example

4.1 Description of the data set

The data was extracted from the website of the Epidemiology Unit of Sri Lanka (http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=225&lang=en) [9] and is related to Covid 19. The data is global and consists of 144 countries. The population

sizes of each country were obtained from a United Nations (UN) database (https://population.un.org/wpp/Download/Standard/Population/)[10]. The data from the Epidemiology unit was made up of the country, the geographical region, the days elapsed from the last Covid 19 case in the country, the total number of Covid 19 cases in the country, and the type of spread of the virus. There were three types of spread, namely, clusters, community, and sporadic. The variable, days lapsed from the last Covid 19 case in the country is a count response variable with a spike at zero. Therefore, a binary variable X was created to differentiate the zeros from the non-zeros. The variable, the total number of Covid 19 cases in the country was another count response with huge numbers and no zeros. The two explanatory variables were X and the type of spread of the virus (Z). There were two cluster variables, namely country, and region. Table 1 gives details of the data.

4.2 Preliminaries for Modeling

Before fitting models, the distribution of the responses and their correlation needs to be determined. For the first response related to the days elapsed from the last Covid 19 case in the country which has a spike at zero, based on Lorenz, Jenkner et al. [15] a Poisson model with an adjustment for the spike at zero was selected. For the total number of Covid 19 cases in the country, another Poisson Model could not be used as the two response variables were negatively correlated. Based on Aitchinson and Ho [1] a Lognormal model was used for the second response with a view to joint modeling. As the counts in the second response were extremely large and therefore, to avoid convergence problems the second response was log-transformed and a Normal model was fitted to impose a lognormal model. Figure 1(a) gives a histogram of the log-transformed second response and Figure 1(b) gives a Normal probability plot of the transformed second response.

Table 1 should come here.

Fig 1(a) and Fig 1(b) should come here.

Figure 1 (a) shows a symmetric histogram close to a Normal distribution while figure 1(b) is close to a straight line except at the lower extreme. Based on these figures a Normal model is selected as the Normal distribution is usually quite robust to small departures from Normality [16].

The two responses are labeled srcase (response 1) and ltcases (response 2) and the correlation between these two variables is -0.4894 and is significant at 0.01%. Before modeling according to Lorenz, Jenkner et al. [15] a value of one is added to srcase in order to attain convergence. This new variable is labeled srcasen. As the correlation is a large negative value srcasen is modelled as Poisson after adjusting for the zeros as in Lorenz, Jenkner et al. [15] and following Aitchinson and Ho [1] the log-transformed second response (ltcases) is modelled as a Normal.

4.3 Univariate Modeling

4.3.1. Modeling srcasen using a univariate random effect model with two random effects

Here we take srcasen to have a Poisson response and the explanatory variables are taken to be the type and X. The random effects are taken to be region and country nested within the region. The link is taken as a log and the offset is taken to be the log of the population size. The type of correlation structure used is variance components. The method of estimation is the maximum likelihood with Laplace approximation of the marginal likelihood. In model fitting, the type variable and X are both significant. The parameter estimates can be interpreted as follows. When type=community, the expected number of days elapsed after the last case, decreases by a ratio of 0.162 compared to type=sporadic. The type=clusters is not significant. When srcase is non-zero the expected number of days elapsed after the last case increases by a ratio of 17.63 compared to

when srcase is zero. The variance parameter estimate of the region random effect is 0.3805. The variance parameter estimate of the country (region) random effect is 2.9052. While the country (region) variance parameter is significant the region variance parameter is not significant. The AIC of the fitted model is 749.37 and the Z value given by the Wald test of the two variance parameters is 1 and 6.56 respectively resulting in p-values of 0.1586 and <0.0001 respectively.

4.3.2. Modeling ltcase using a univariate random effect model with two random effects

Here we take ltcase the log-transformed response 2 to have a normal distribution. As before the explanatory variables are taken to be type and X. The random effects are taken to be region and country nested within the region. The link is taken as identity. The type of correlation structure used is variance components. The method of estimation is the maximum likelihood with Laplace approximation of the marginal log-likelihood. Both the type variable and X are significant in this model. The parameter estimates can be interpreted as follows. When type=clusters the ratio of the total number of cases increases by 6.63 compared to sporadic type. When type=community the ratio of the total number of cases increases by 58.99 compared to sporadic type. When X is non-zero the ratio of the total number of cases decreases by 0.1617 compared to the case when X is zero. The variance parameter estimate of region random effect is 0.3364. The variance parameter estimate of the country (region) is 1.5817. Here the country (region) random effect is significant while the region random effect is not significant. The AIC of the fitted model is 538.84 and the Z value given by the Wald test of the two variance components is 1.19 and 3.14 resulting in p-values of 0.117 and 0.0008 respectively.

4.4 Joint Modeling of srcasen and ltcases using Poisson and Normal distributions respectively

Here we build a joint model taking srcasen to have a Poisson distribution and ltcases to have a normal distribution. A random-effects model with shared random effects is used to fit this joint model. Two random effects one to represent the correlation between countries within regions and the other to represent the correlation between responses within countries were used. The covariance structure used was compound symmetry. The method of estimation was maximum likelihood with Laplace approximation of the marginal log-likelihood. The parameter estimates can be interpreted as follows. Both type and X were significantly associated with both responses srcasen and ltcases. For srcasen the ratio of the number of days after the last case for type=clusters reduces by 0.5612 compared to type=sporadic and for type=community it decreases by 0.4982 compared to type=sporadic. When X is non zero the rate increases by 4.833 compared to when X is zero. For ltcases the ratio of the total number of cases increases by 7.54 for type=cluster compared to type=sporadic and increases by 59.81 for type=community compared to type=sporadic. When X is non zero the total number of cases reduces by 0.1572 compared to when X=0. The variance-covariance matrix of the region is

$$\begin{bmatrix} 0.0081 & 0.0074 \\ 0.0074 & 0.0081 \end{bmatrix}$$

And the Variance-Covariance matrix of the country (region) is

$$\begin{bmatrix} 0.4101 & -0.2082 \\ -0.2082 & 0.4101 \end{bmatrix}$$

As in the univariate case, the region random effect is not significant while the country (region) random effect is highly significant. As seen before the correlation between responses within

countries is negative. The AIC of this model is 1043.34 and the Z value given by the Wald test of the two variance components is 0.34 and 8.04 resulting in p-values 0.3667 and <0.0001 respectively.

4.5 Comparison of the Univariate and Joint Models

Table 2 gives the fit statistics of the two univariate and the joint models.

Table 2 should come here.

By comparing the estimates given in table 2, it is evident that the joint model has a better performance as its AIC was lower (1043.34) than the sum of the AICs of the univariate models (1288.21) and the Z value given by the Wald test of the variance parameter of the Country (Region) effect of the joint model was higher than those of the univariate models.

The parameter estimates for the normal component of the joint model are close to the parameter estimates of the corresponding normal univariate model. However, the Poisson components are very different. The difference in the parameter estimates of the univariate and joint model is due to the joint model taking account of the correlation between responses while the univariate models are unadjusted for correlation.

4.6 Examining the fit of the joint model

To examine how good the fitted joint model is the students' residuals were plotted against the predicted values. The 99% horizontal confidence bands were also superimposed on the same plot at y=-2.58 and y=2.58. Figure 2 gives this plot. Of the 288 observations all but 6 observations lay within the 99% confidence bands. Even the 6 observations outside the bands were small outliers.

Also, there is no other pattern in the plot except some lines. The covariates in this model, Dist (pertaining to distribution) and X (Pertaining to the SAZ variable) are binary/dichotomous and Z (pertaining to type) is categorical. Collett [3] explains that wherever there are binary/dichotomous independent predictors in the model these linear patterns are a usual occurrence. This indicates a satisfactory fit of the model.

Figure 2 should come here.

4.7 Comparing the developed joint model with the traditional joint model

Here we compare the newly developed joint model with the traditional joint model ignoring the methods developed for the excess of zeros and huge counts. In the traditional model count response, 1 is modeled as a Poisson variable while count response 2 is modeled as a lognormal variable. The AIC is smaller (1026.70) in the traditional model compared to the newly developed model (1043.34). However, these two AIC values cannot be directly compared as these are based on two slightly different data sets due to the newly developed model including the additional binary variable. The Z statistic given by the Wald test for the Country (Region) variance component is much less for the traditional model (5.35) compared to the newly developed model (8.04). The Studentized residual plot for the traditional model shows 11 points outside the 99% confidence bands while the newly developed model shows only 6 points outside these bands. Also, the width of the residual plot on the vertical axis is much wider for the traditional model compared to the newly developed model indicating that the traditional model has bigger outliers. The only patterns here are linear and curvy linear. The covariates in this model, are Dist (pertaining to distribution) is binary/dichotomous and Z (pertaining to type) is categorical. Collett [3] explains that wherever there are binary/dichotomous independent predictors in the model then these patterns are a usual

occurrence and therefore, it is no surprise that the plot contains some linear and curvy-linear patterns [3]. This is shown in Figure 3.

Figure 3 should come here.

There are three instances where the newly developed model is better than the traditional model. So overall the newly developed model is superior to the traditional model.

5. Discussion

5.1 Important Conclusions

When there are two counts and one has a spike at zero and the other has very large non zero counts the former variable can be modeled as a Poisson random variable with log link using the technique of Lorenz, Jenkner et al. [15] successfully. The other variable can be log-transformed and modeled as a Normal response with an identity link.

When these two variables are highly negatively correlated these cannot be jointly modeled using a bivariate Poisson or bivariate negative binomial distribution. Thus transforming one variable is the only option. Aitchinson and Ho [1] suggest a way around this situation and they have modeled one count using the lognormal distribution. As there is a problem of heterogeneity in our second variable our joint model gives a better fit when log-transformed and modeled as a Normal response. A Generalized Linear Mixed Model (GLMM) in the form of Hapugoda and Sooriyarachchi [8] can be used to fit this joint model with Maximum Likelihood Estimation (MLE) and Laplace Approximation of the marginal log-likelihood.

This procedure was illustrated on an example related to a covid 19 data set. Two random effects of which one was a nested effect were used in the joint model [19]. The type of covariance matrix

used was compound symmetry [8]. The explanatory variable Type indicated that the expected number of days elapsed after the last covid 19 cases was significantly less for cluster and community type of spreading of the disease compared to a sporadic type of spreading. For the log count of the total number of covid 19 cases, there were significantly more cases when the spread type was cluster and community compared to sporadic type.

When the joint model was compared to two univariate models, the AIC of the joint model was nearly 250 less than the AIC of the sum of the two univariate models. In addition, the standard error of the variance parameters of the variance-covariance matrix was very much lower in the joint model compared to the two univariate models [8]. The explanatory variables X and type were both significant for both responses in the joint model. Of the two random effects, only the nested effect country (region) was significant and the random effect region was not significant.

The plot of the studentized residuals versus the predicted value was drawn to examine the goodness of fit of the joint model. Of 288 observations all except 6 observations were within the 99% confidence bands. Even these 6 observations resulted in small outliers. The complex scenario was successfully modeled using the model proposed.

5.2 Comparing this research with what is known in the literature

Comparing our method with Lorenz, Jenkner et al. [15] our method was as successful as theirs for a much more complicated scenario. When comparing our research with Fernando and Sooriyarachchi [4] they had a positive correlation which was modeled by a bivariate negative binomial distribution. However, as our responses were negatively correlated that research could not be followed. When comparing our research with Hapugoda and Sooriyarachchi [8] their research modeled survival and count data and the count variable did not have a spike at zero and

had only one random effect. When comparing our research with Sunethra and Sooriyarachchi [19] they too developed a joint model for survival and count variables where they did not have a spike at zero. They used a separate random effect model while we used a shared random effect model with the transformation of one variable. When comparing the newly developed model with the traditional model, overall the newly developed model was superior based on the example used.

Here in this research, it should be mentioned that log transforming the second count variable and modeling it as normal is the same as modeling the untransformed second count variable using the lognormal distribution.

The methods used here build upon a combination of ideas from the literature and these ideas have never been put together before as explained in the literature review. Therefore, the methods developed here and the example analyzed is a novel technique for the scenario considered.

5.3 Limitations of the study

In the example, there was only one explanatory variable in the study. The offset variable was the log of the population size. For some countries there were no values for the population so these countries had to be dropped from the analysis. Most distribution combinations for the two responses did not converge. Both zero-inflated Poisson and Zero-inflated Negative Binomial did not work for the count response with a spike at zero.

5.4 Further Work

On the methodology side, one could write a computer program to incorporate the zeros in the likelihood function without adding unity to the count data with zeros. However, this is not straight

forward as this is a joint model. More covariates and interactions could be implemented using another example. The deviance could be calculated to provide an objective goodness of fit statistic than the residual plot which is subjective.

References

1. Aitchinson J. and Ho C.H. (1989). The Multivariate Log-Normal Distribution. *Biometrika* 76 643 - 653
2. Allison, D. (2012) Do we really need inflated models? https://statisticalhorizons.com/zero-inflated-models retrieved on 15th May, 2020.
3. Collett, D. (1991). Modeling Binary Data. Chapman & Hall/CRC Texts in Statistical Science, USA.
4. Shenali Maryse Fernando, Marina Roshini Sooriyarachchi (2018). Bivariate Negative Binomial Modeling of Epidemiological Data. *Open Science Journal of Statistics and Application* 5(3) : 47-57
5. Gayara Fernando and Roshini Sooriyarachchi (2020): The development of a goodness-of-fit test for high level binary multilevel models, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1700275
6. Gurumu, S. and Elder, J. (2012). Flexible Bivariate Count Data Regression Model. *Journal of Business and Economic Statistics* 30(2) : 265 – 274.
7. Hapugoda, J. C., Sooriyarachchi, M. R., Kalupahana, R. S. and Satharasinghe, D. A. "Joint Modeling of Mixed Responses: An Application to Poultry Data". Proceedings of 5th

Annual International Conference on Operations Research and Statistics (ORS), Singapore, pp. 182-185, 2017.

8. Hapugoda J. and Sooriyarachchi, M.R.(2018). Joint Modeling of Discrete Time Hazard Model with Poisson Regression Model: A Simulation Study. Proceedings of the JUICE conference, University of Jaffna.
9. http://personal.lse.ac.uk/tenreyro/poisson.pdf. Retrieved on 15th May, 2020
10. http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=225&lang=en, retrieved on 25th April 2020
11. https://www.un.org/en/development/desa/population/publications/manual/projection/index.asp. Retrieved on 17th May, 2020
12. Jayanetti, W., and Sooriyarachchi, R. "A multilevel study of dengue Epidemiology in Sri Lanka: modeling survival of dengue patients", International Journal of Mosquito Research, 2 (3), pp. 114-121, 2015.
13. Karunarathna G.H.S. and Sooriyarachchi M.R. (2019). Joint Multilevel Model for Analyzing Length of Stay through Competing End points in Dengue Epidemiology. *Sri Lankan Journal of Applied Statistics* 19(1) : 45 – 60.
14. Kochelerkota S. and Kochelakota K. (1992). Discrete Distributions New York. *Marcel Deckker.*
15. Lorenz, E. Jenkner, C. Sauerbrei, W. Becher, H. (2019). Modeling Exposures with a Spike at Zero: Simulation Study and a Practical Application to Survival Data. *Biostatistics and Epidemiology* 3 (1) : 23 – 37.
16. McDonald, J.H. (2014) Handbook of Biological Statistics (3rd ed.). *Sparky House Publishing*, Baltimore, Maryland pages 140-144.

17. Ophem, H. van (1999) A General Method to Estimate Correlated Discrete Random Variables. *Econometric Theory* 15: 228 -237.
18. Rizopoulos D. (2012).Joint models for longitudinal and time to event data with applications in R. *India CRC Press*.
19. Sunethra A. A. and Sooriyarachchi M. R. (2020). A Novel Method for Joint Modeling of Survival Data and Count Data for both Simple Randomized and Cluster Randomized Data. *Communications in Statistics. Theory and Methods*. Published online.
20. D.Wickramarachchi A Goodness of Fit (GOF) Test for Bivariate Binary Multilevel Logistic Model (unpublished B.Sc. thesis, 2017). University of Colombo, Sri Lanka.

Joint Modeling of Two Count Variables using a Shared Random Effect Model in the presence of Clusters for Complex Data

Marina Roshini Sooriyarachchi

Department of Statistics, University of Colombo, Colombo 3, Sri Lanka

Prof. M.R.Sooriyarachchi

Corresponding Author

Department of Statistics,

University of Colombo

Colombo 3

Sri Lanka

Telephone: 009411-2590111

Email:roshini@stat.cmb.ac.lk

Abstract

In epidemiology, it is often the case that two or more correlated count response variables are encountered. Under this scenario, it is more efficient to model the data using a joint model. Besides if one of these count variables has an excess of zeros (spike at zero) the log link cannot be used in general. The situation is more complicated when the data is grouped into clusters. A Generalized Linear Mixed Model (GLMM) is used to accommodate this cluster covariance. The objective of this research is to develop a new modeling approach that can handle this situation. The method is illustrated on a global data set of Covid 19 patients. The important conclusions are that the new model was successfully implemented both in theory and practice. A plot of the residuals indicated a well-fitting model to the data.

Keywords: Joint Model, Generalized Linear Mixed Model (GLMM), Cluster, Spike at zero, random effects, Covid 19

1. Introduction

1.1 Background

In epidemiological and health studies often several correlated count responses are encountered [4] [6]. This type of data is often found to occur within groups (clusters). In this case, it is more efficient to model these count responses jointly rather than model each response separately. Fernando and Sooriyarachchi [4] use Generalized Linear Mixed Modeling (GLMM) with random cluster effects for this scenario which is now quite well developed. In this research, a more difficult problem involving four complexities that can be encountered with count data is considered. The first is the presence of a huge number of zero counts resulting in an enormous spike at zero [15].

The zero-inflated Poisson is an option for modeling such data but in the context of joint models could have convergence problems [2]. The second problem is when the counts are non-zero but huge. Neither the Poisson nor the Negative Binomial converges especially in the case of joint models [9]. The third situation is the presence of a negative correlation between responses [19]. Even though there are indirect methods developed in the literature to handle this, these have been developed only for binary, ordinal, and continuous data [19]. The fourth problem is related to cluster-specific covariates which occur at different levels of the hierarchy [5]. Rizopoulos [18] has dealt with this for survival and normal responses but not for two count variables. In the case of the first problem, the literature provides a solution for univariate models [15]. In the case of the second problem log transforming the counts and modeling the transformed values as Normal often works [13, 16]. In the situation of the third problem GLMM with common random effects has been seen to work for survival and count joint responses [19]. In the final problem multilevel modeling for high dimensional problems has been seen to work for bivariate binary problems [5].

1.2 Objectives

The primary objective of this research is to develop a new model for the scenario described in section 1.1. The secondary objective is to apply the model to a suitable set of data.

1.3 Brief description of Methods

The joint model is developed for a Poisson-Normal joint distribution. A Generalized Linear Mixed Model with Maximum Likelihood Estimation and Laplace approximation for the marginal log-likelihood was used for this purpose [7]. Two random effects to incorporate two different cluster effects were used. These random effects are used to join the two responses. The covariance

structure used was compound symmetry. It was assumed that both responses have the same random variance.

1.4 Data for the example

The data for the example is related to Covid 19 and the relevant details were obtained from the website of the Epidemiology Unit of Sri Lanka [10]. The two, count responses pertained to the expected number of days elapsed after the last corona case and the total number of corona cases. The former had a spike at zero and the latter had very large counts. There was one explanatory variable, namely, the type of transmission of the cases. There were 144 countries and this database was merged with the population size database given by the UN website [11].

1.5 Structure of the Paper

Section 1 consists of an introduction to the problem, objectives, a brief description of methods, and an explanation of the data. Section 2 is made up of a literature review. In section 3 the new model is developed. Section 4 gives the example and the discussion consists of section 5 followed by references.

2. Literature Review

2.1 Joint modeling of two count variables

In epidemiology, often two correlated count variables are encountered, such as the incidence of the disease and the platelet count in dengue, the incidence of the disease and white blood cell counts in Japanese encephalitis, the incidence of Leptospirosis and the count of serovar-specific antibodies to name a few examples. As most of these diseases also depend on the climate and thus

on the geographical region the region happens to be a cluster variable. Many zero counts are possible in regions where the weather is not conducive to the disease. Thus resulting in a SAZ. The second variable is usually related to huge counts and the weather parameters are cluster level variables. In the first example given the correlation between the two counts is negative with high dengue counts being related to low platelet counts. Under this epidemiological scenario the major characteristics are two count variables, cluster variation, one count variable with a SAZ and the other count variable with huge counts, correlated counts with the more challenging being that the counts are negatively correlated. Thus each of these characteristics are reviewed in the next sections.

Kochelerkota and Kochelerkota [14] and Ophem [17] give a detailed literature review of this situation. Gurmu and Elder [6], discuss the joint modeling of two count variables when these variables are negatively correlated. They mention that in this scenario the bivariate Poisson and the bivariate negative binomial cannot be used to model the data. They consider a two-factor framework where dependence between the count variables is modeled using correlated unobserved heterogeneity components. Their article uses semi-parametric methods for the estimation of a mixture of count models that include negatively correlated counts. Aitchinson and Ho [1] particularly discuss the case of negatively correlated count variables where they use a Poisson-Lognormal mixture to model two count variables with a negative correlation.

However, their methods [1] [6] do not take in to account the adjustments for clustering, a spike at zero, huge counts, and cluster-level covariates.

Hapugoda and Sooriyarachchi [8] develop a joint model using a single shared random effect to model survival and count responses combining the discrete time hazard model and Poisson model. This method does not take in to account high dimensional data and has only one random effect.

The model [8] though considering clustered data is a joint model for binary and count data and does not accommodate an excess of zeros nor huge counts. It can, however, handle a negative correlation.

Sunethra and Sooriyarachchi [19] develop joint models using two separate random effects to model survival and count data which are negatively correlated. They use the lognormal distribution to model the survival data and the Poisson distribution to model the count data. However, they [19] do not consider the case of excess zeros nor huge counts.

Wickremarachchi (unpublished B.Sc. thesis, 2017) [20] develop a bivariate binomial model in the presence of clusters. This is modeled using multilevel modeling. Here a different technique to Hapugoda and Sooriyarachchi [8] and Sunethra and Sooriyarachchi [19] is used to model the correlation within clusters. It is another option for the cluster scenario. This technique is multilevel modeling. This method [20] does not accommodate negative correlation, excess zeros, high dimensional data nor huge counts.

2.2 Use of Random Effect Models for Joint Model Development

A pioneer of random effect models for joint model development is Rizopoulos [18]. He developed joint models for survival and repeated measures responses considered to be normally distributed. Sunethra and Sooriyarachchi [19] give a detailed review of this situation. They consider both the case of positive, and negative correlation between two response variables. For the case of positive correlation, they discuss shared random effect models and for negatively correlated responses they consider separate random effect models. However, his work [18] is purely for joint survival and normal longitudinal models. He has not considered count responses and so his work does not discuss an excess of zeros and huge counts.

Similar to Sunethra and Sooriyarachchi [19], in this paper, the author examined the joint modeling of two count variables using shared random effects with two random effects at two different levels. The method used in this research is a technique that combines the methods of Sunethra and Sooriyarachchi [19] with that of Aitchinson and Ho [1]; Lorenz, Jenkner et al. [15], and G. Fernando and Sooriyarachchi [5].

This current approach uses two shared random effects as in [19], uses appropriate methods to incorporate negative correlation for Poisson-Lognormal mixtures as in [2], adjusts for a spike at zero as in [15], and extends the problem to 3-dimensional data as in [5]. Apart from combining these methods, further extensions have been made by looking at significance tests for the random effects, extensions were also made to [19] and [5] where the survival and count joint model and the univariate binary model respectively were changed to a joint count model. Finally [15] was extended from a univariate model to a bivariate model.

2.3 The case of a peak (spike) at zero

When in addition to negative correlation if one of the count variables has a spike at zero, neither the zero-inflated Poisson nor the zero-inflated negative binomial usually converge [2]. Lorenz, Jenkner et al. [15] introduced four methods to handle this case. They discussed the case occurring often in Epidemiological and Clinical Research where variables are often semi-continuous with several patients often having exposure zero and a continuous distribution among those exposed. This is referred to as a spike at zero (SAZ). They illustrated their procedures on a German Breast Cancer Study Group data (GBSG). Their method involves dichotomizing the SAZ variable into a binary variable (X) with the two levels relating to the zeros and non-zeros. Then the binary information is combined with the positive continuous variables. This information is combined into one variable in the standard technique to give by default the linear component. Using the approach

of Lorenz, Jenker et al. [15] in our study we take the first count variable (srcasen) split into two, one a binary variable for zeros and non-zeros and the other a continuous srcasen variable and the two srcasen variables are treated as one prognostic factor in the model. Both variables are tested jointly in the model.

Expectation [Response] (combined variable) = Exp ($\beta Z + \gamma X$) where Z consists of the other explanatory variables and the intercept (for intercept $\beta=1$) and X consists of the binary explanatory variable. There is no intercept in this model. If the Response is a count it can be taken as having a Poisson distribution with a log link. Here β and γ are the unknown coefficients of Z and X respectively.

The authors of this paper [15] do not consider hierarchical data in the form of clusters nor huge counts. Also, they only consider the univariate case.

The authors [15] have also mentioned about some discrepancies in the method and the way of getting over these discrepancies. According to them, "Modeling such SAZ variables is challenging and there are both statistical problems and problems conserning interpretation arising from this situation. Readers are referred to paper [15] for more information on the overcoming of these problems.

2.4 The case of cluster effects

When the data is hierarchical we refer to this as multilevel data. We consider here the case where there are three levels. The third level is a large cluster within which lies a small cluster referred to as the second level within which lies two correlated count responses referred to as the first level. If the correlation within a cluster is significant then the model cannot be fitted using standard models. This correlation has to be taken in to account. [5]

2.5 Methods used in this research

Here a shared random effects joint model is used to model two negatively correlated counts using the Poisson-Normal mixture. A spike at zero is taken into consideration also. The model is developed within the framework of hierarchical models. The method used in this research is a technique that combines the methods of Sunethra and Sooriyarachchi [19] with that of Aitchinson and Ho [1]; Lorenz, Jenkner et al. [15], and G. Fernando and Sooriyarachchi [5]. This combination is not found in the literature and therefore, is a novel development.

3. Theory

Consider the method of Hapugoda and Sooriyarachchi [8] where the Procedure Glimmix in SAS 9.4 is used to fit a shared parameter joint model to a survival and count response. The model fitted is a Generalized Linear Mixed Model (GLMM) with one random effect representing a single cluster. This method will be modified for this situation. Here we use a Poisson model with an adjustment for a spike at zero for one count variable and the other count variable is log-transformed and modeled as a Normal response variable [1][16]. The correlation matrix is modeled as of type compound symmetry. The method of estimation used was the Maximum Likelihood with Laplace Approximation of the marginal log-likelihood [7]. In this research, a shared parameter joint model for joining the two count responses is fitted. The model fitted is a Generalized Linear Mixed Model (GLMM) with two random effects representing two sets of clusters.

3.1 Poisson Regression Model for clustered data

Suppose y_{ij1} is the first observed count for the ith small cluster in level 2 within the jth third level cluster where y_{ij1}~ Poisson (μ_{ij1}) and μ_{ij1} is the mean of the Poisson distribution for the 1st observation of level 1 within the ith 2nd level unit within the jth third level unit. E_{ij1} is the Expected count or offset [12]. The Z_{ij} are the predictors and $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ is the random intercept where β_0 is a fixed component and u_{0ij} is a random component for cluster-level 2 (intercept) and v_{0j} is a random component for cluster-level 3. Let X_{ij} be the binary variable to adjust for the spike at zero. Then a three-level random intercept Poisson Regression model can be given by

$\log(\mu_{ij1}) = \log(E_{ij1}) + \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$. (1)

To classify a count as either zero or not, a binary variable X is added to the model. It is assessed in a two-stage procedure to determine whether the binary variable and/or the continuous function for the positive part is required for a suitable fit [15].

3.2 The Normal model for log count clustered data

Let y_{ij2} be the log-transformed second count variable for the ith small cluster in level 2 within the jth third level cluster where y_{ij2} ~ Normal (μ_{ij2}, σ_{ij2}^2) where μ_{ij2} is the mean of the Normal distribution for the 2nd observation of level 1 within the ith 2nd level unit within the jth third level unit and σ_{ij2}^2 is the variance of the Normal distribution for the 2nd observation of level 1 within the ith second level unit within the jth third level unit. The Z_{ij} are the predictors and $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ where u_{0ij} and v_{0j} are as in section 3.1. Then a three-level random intercept Normal Regression

model can be given by

$\mu_{ij2} = \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$

(2)

3.3 The joint model for clustered data

The responses of analysis are Yij_1 (Poisson – Count 1) and Yij_2 (Normal – Log transformed count 2). The suffixes i and j, are as defined before. Variables that impact $Y = (Y_1, Y_2)$ are the explanatory variables (X_{ij} and Z_{ij}) as defined before i=1,2,…,I where I is the number of small clusters and j=1,2,…., J where J is the number of large clusters. To formulate a joint model, Generalized Linear Model (GLM) can be used to form marginal models for each response by considering mean $E(Y_{ijk}/X_{ij,} Z_{ij})$ and variance $Var(Y_{ijk}/X_{ij,} Z_{ij})$ where k=1,2. The approach to link the responses is by structuring a covariance matrix $Var(Y_{ijk}/X_{ij,} Z_{ij})$ to include potential correlations.[16]. The random effects are assumed to be the same for both responses so this is a shared random-effects model.

In GLM $l_k\left(E\left(Y_{ijk}/X_{ij}, Zij\right)\right) = X_{ijk}'\beta_k + Z_{ijk}'\gamma_k$, k=1,2 where i,j denotes each record from each i^{th} small cluster within each j^{th} large cluster and l_k is the link function. Here, $l_1(u)$ is the log link and $l_2(u)$ is the identity link. GLIMMIX is used to estimate two marginal models jointly.

A structural formulation of the model is given as:

$l_1(Y'_{ij1}) = \text{Log}(\mu_{ij1}) = \log(E_{ij1}) + \beta_{0ij} + \beta Z_{ij} + \gamma X_{ij}$ where $\beta_{0ij} = \beta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$. (3)

and

$l_2(Y'_{ij2}) = (\mu_{ij2}) = \vartheta_{0ij} + \vartheta X_{ij} + \delta X_{ij}$ where $\vartheta_{0ij} = \vartheta_0 + u_{0ij} + v_{0j}$ and $u_{0ij} \sim N(0, \sigma_u^2)$ and $v_{0j} \sim N(0, \sigma_v^2)$ (4)

Here k=1,2 and i=1(1) n_j and j=1(1)m where n_j is the number of small clusters within big cluster j and m is the number of big clusters.

For simplicity, we assume that both sets of random effects are the same (u_{0ij} and v_{0j}) and have the same variance(σ_u^2 *and* σ_v^2 respectively). The joint model variance-covariance matrix, Var-Cov (Y_{ij1}, Y_{ij2}) is of the form $\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ where:

$\sigma_2^2 = \sigma_u^2 + \sigma_v^2 + \sigma^2$ and σ^2 is the variance of the error term in the regression.

Here σ_1^2 can be derived using the methods of Sunethra et al. (2020) [19]. The correlation between the Y_{ij1} and Y_{ij2} is taken to be ρ_{12}. It is assumed that the u_{0ij}'s are independent of the v_{0j}'s. GLIMMIX will structure the variance-covariance matrix of Y = (Y_1, Y_2) as in Hapugoda et al. [7]. The development of the joint log-likelihood and thereby the joint model is given in detail in Sunethra and Sooriyarachchi [19].

4. Example

4.1 Description of the data set

The data was extracted from the website of the Epidemiology Unit of Sri Lanka (http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=225&lang=en) [9] and is related to Covid 19. The data is global and consists of 144 countries. The population sizes of each country were obtained from a United Nations (UN) database

(https://population.un.org/wpp/Download/Standard/Population/)[10]. The data from the Epidemiology unit was made up of the country, the geographical region, the days elapsed from the last Covid 19 case in the country, the total number of Covid 19 cases in the country, and the type of spread of the virus. There were three types of spread, namely, clusters, community, and sporadic. The variable, days lapsed from the last Covid 19 case in the country is a count response variable with a spike at zero. Therefore, a binary variable X was created to differentiate the zeros from the non-zeros. The variable, the total number of Covid 19 cases in the country was another count response with huge numbers and no zeros. The two explanatory variables were X and the type of spread of the virus (Z). There were two cluster variables, namely country, and region. Table 1 gives details of the data.

4.2 Preliminaries for Modeling

Before fitting models, the distribution of the responses and their correlation needs to be determined. For the first response related to the days elapsed from the last Covid 19 case in the country which has a spike at zero, based on Lorenz, Jenkner et al. [15] a Poisson model with an adjustment for the spike at zero was selected. For the total number of Covid 19 cases in the country, another Poisson Model could not be used as the two response variables were negatively correlated. Based on Aitchinson and Ho [1] a Lognormal model was used for the second response with a view to joint modeling. As the counts in the second response were extremely large and therefore, to avoid convergence problems the second response was log-transformed and a Normal model was fitted to impose a lognormal model. Figure 1(a) gives a histogram of the log-transformed second response and Figure 1(b) gives a Normal probability plot of the transformed second response.

Table 1 should come here.

Fig 1(a) and Fig 1(b) should come here.

Figure 1 (a) shows a symmetric histogram close to a Normal distribution while figure 1(b) is close to a straight line except at the lower extreme. Based on these figures a Normal model is selected as the Normal distribution is usually quite robust to small departures from Normality [16].

The two responses are labeled srcase (response 1) and ltcases (response 2) and the correlation between these two variables is -0.4894 and is significant at 0.01%. Before modeling according to Lorenz, Jenkner et al. [15] a value of one is added to srcase in order to attain convergence. This new variable is labeled srcasen. As the correlation is a large negative value srcasen is modelled as Poisson after adjusting for the zeros as in Lorenz, Jenkner et al. [15] and following Aitchinson and Ho [1] the log-transformed second response (ltcases) is modelled as a Normal.

4.3 Univariate Modeling

4.3.1. Modeling srcasen using a univariate random effect model with two random effects

Here we take srcasen to have a Poisson response and the explanatory variables are taken to be the type and X. The random effects are taken to be region and country nested within the region. The link is taken as a log and the offset is taken to be the log of the population size. The type of correlation structure used is variance components. The method of estimation is the maximum likelihood with Laplace approximation of the marginal likelihood. In model fitting, the type variable and X are both significant. The parameter estimates can be interpreted as follows. When type=community, the expected number of days elapsed after the last case, decreases by a ratio of 0.162 compared to type=sporadic. The type=clusters is not significant. When srcase is non-zero the expected number of days elapsed after the last case increases by a ratio of 17.63 compared to

when srcase is zero. The variance parameter estimate of the region random effect is 0.3805. The variance parameter estimate of the country (region) random effect is 2.9052. While the country (region) variance parameter is significant the region variance parameter is not significant. The AIC of the fitted model is 749.37 and the Z value given by the Wald test of the two variance parameters is 1 and 6.56 respectively resulting in p-values of 0.1586 and <0.0001 respectively.

4.3.2. Modeling ltcase using a univariate random effect model with two random effects

Here we take ltcase the log-transformed response 2 to have a normal distribution. As before the explanatory variables are taken to be type and X. The random effects are taken to be region and country nested within the region. The link is taken as identity. The type of correlation structure used is variance components. The method of estimation is the maximum likelihood with Laplace approximation of the marginal log-likelihood. Both the type variable and X are significant in this model. The parameter estimates can be interpreted as follows. When type=clusters the ratio of the total number of cases increases by 6.63 compared to sporadic type. When type=community the ratio of the total number of cases increases by 58.99 compared to sporadic type. When X is non-zero the ratio of the total number of cases decreases by 0.1617 compared to the case when X is zero. The variance parameter estimate of region random effect is 0.3364. The variance parameter estimate of the country (region) is 1.5817. Here the country (region) random effect is significant while the region random effect is not significant. The AIC of the fitted model is 538.84 and the Z value given by the Wald test of the two variance components is 1.19 and 3.14 resulting in p-values of 0.117 and 0.0008 respectively.

4.4 Joint Modeling of srcasen and ltcases using Poisson and Normal distributions respectively

Here we build a joint model taking srcasen to have a Poisson distribution and ltcases to have a normal distribution. A random-effects model with shared random effects is used to fit this joint model. Two random effects one to represent the correlation between countries within regions and the other to represent the correlation between responses within countries were used. The covariance structure used was compound symmetry. The method of estimation was maximum likelihood with Laplace approximation of the marginal log-likelihood. The parameter estimates can be interpreted as follows. Both type and X were significantly associated with both responses srcasen and ltcases. For srcasen the ratio of the number of days after the last case for type=clusters reduces by 0.5612 compared to type=sporadic and for type=community it decreases by 0.4982 compared to type=sporadic. When X is non zero the rate increases by 4.833 compared to when X is zero. For ltcases the ratio of the total number of cases increases by 7.54 for type=cluster compared to type=sporadic and increases by 59.81 for type=community compared to type=sporadic. When X is non zero the total number of cases reduces by 0.1572 compared to when X=0. The variance-covariance matrix of the region is

$$\begin{bmatrix} 0.0081 & 0.0074 \\ 0.0074 & 0.0081 \end{bmatrix}$$

And the Variance-Covariance matrix of the country (region) is

$$\begin{bmatrix} 0.4101 & -0.2082 \\ -0.2082 & 0.4101 \end{bmatrix}$$

As in the univariate case, the region random effect is not significant while the country (region) random effect is highly significant. As seen before the correlation between responses within

countries is negative. The AIC of this model is 1043.34 and the Z value given by the Wald test of the two variance components is 0.34 and 8.04 resulting in p-values 0.3667 and <0.0001 respectively.

4.5 Comparison of the Univariate and Joint Models

Table 2 gives the fit statistics of the two univariate and the joint models.

Table 2 should come here.

By comparing the estimates given in table 2, it is evident that the joint model has a better performance as its AIC was lower (1043.34) than the sum of the AICs of the univariate models (1288.21) and the Z value given by the Wald test of the variance parameter of the Country (Region) effect of the joint model was higher than those of the univariate models.

The parameter estimates for the normal component of the joint model are close to the parameter estimates of the corresponding normal univariate model. However, the Poisson components are very different. The difference in the parameter estimates of the univariate and joint model is due to the joint model taking account of the correlation between responses while the univariate models are unadjusted for correlation.

4.6 Examining the fit of the joint model

To examine how good the fitted joint model is the students' residuals were plotted against the predicted values. The 99% horizontal confidence bands were also superimposed on the same plot at y=-2.58 and y=2.58. Figure 2 gives this plot. Of the 288 observations all but 6 observations lay within the 99% confidence bands. Even the 6 observations outside the bands were small outliers.

Also, there is no other pattern in the plot except some lines. The covariates in this model, Dist (pertaining to distribution) and X (Pertaining to the SAZ variable) are binary/dichotomous and Z (pertaining to type) is categorical. Collett [3] explains that wherever there are binary/dichotomous independent predictors in the model these linear patterns are a usual occurrence. This indicates a satisfactory fit of the model.

Figure 2 should come here.

4.7 Comparing the developed joint model with the traditional joint model

Here we compare the newly developed joint model with the traditional joint model ignoring the methods developed for the excess of zeros and huge counts. In the traditional model count response, 1 is modeled as a Poisson variable while count response 2 is modeled as a lognormal variable. The AIC is smaller (1026.70) in the traditional model compared to the newly developed model (1043.34). However, these two AIC values cannot be directly compared as these are based on two slightly different data sets due to the newly developed model including the additional binary variable. The Z statistic given by the Wald test for the Country (Region) variance component is much less for the traditional model (5.35) compared to the newly developed model (8.04). The Studentized residual plot for the traditional model shows 11 points outside the 99% confidence bands while the newly developed model shows only 6 points outside these bands. Also, the width of the residual plot on the vertical axis is much wider for the traditional model compared to the newly developed model indicating that the traditional model has bigger outliers. The only patterns here are linear and curvy linear. The covariates in this model, are Dist (pertaining to distribution) is binary/dichotomous and Z (pertaining to type) is categorical. Collett [3] explains that wherever there are binary/dichotomous independent predictors in the model then these patterns are a usual

occurrence and therefore, it is no surprise that the plot contains some linear and curvy-linear patterns [3]. This is shown in Figure 3.

Figure 3 should come here.

There are three instances where the newly developed model is better than the traditional model. So overall the newly developed model is superior to the traditional model.

5. Discussion

5.1 Important Conclusions

When there are two counts and one has a spike at zero and the other has very large non zero counts the former variable can be modeled as a Poisson random variable with log link using the technique of Lorenz, Jenkner et al. [15] successfully. The other variable can be log-transformed and modeled as a Normal response with an identity link.

When these two variables are highly negatively correlated these cannot be jointly modeled using a bivariate Poisson or bivariate negative binomial distribution. Thus transforming one variable is the only option. Aitchinson and Ho [1] suggest a way around this situation and they have modeled one count using the lognormal distribution. As there is a problem of heterogeneity in our second variable our joint model gives a better fit when log-transformed and modeled as a Normal response. A Generalized Linear Mixed Model (GLMM) in the form of Hapugoda and Sooriyarachchi [8] can be used to fit this joint model with Maximum Likelihood Estimation (MLE) and Laplace Approximation of the marginal log-likelihood.

This procedure was illustrated on an example related to a covid 19 data set. Two random effects of which one was a nested effect were used in the joint model [19]. The type of covariance matrix

used was compound symmetry [8]. The explanatory variable Type indicated that the expected number of days elapsed after the last covid 19 cases was significantly less for cluster and community type of spreading of the disease compared to a sporadic type of spreading. For the log count of the total number of covid 19 cases, there were significantly more cases when the spread type was cluster and community compared to sporadic type.

When the joint model was compared to two univariate models, the AIC of the joint model was nearly 250 less than the AIC of the sum of the two univariate models. In addition, the standard error of the variance parameters of the variance-covariance matrix was very much lower in the joint model compared to the two univariate models [8]. The explanatory variables X and type were both significant for both responses in the joint model. Of the two random effects, only the nested effect country (region) was significant and the random effect region was not significant.

The plot of the studentized residuals versus the predicted value was drawn to examine the goodness of fit of the joint model. Of 288 observations all except 6 observations were within the 99% confidence bands. Even these 6 observations resulted in small outliers. The complex scenario was successfully modeled using the model proposed.

5.2 Comparing this research with what is known in the literature

Comparing our method with Lorenz, Jenkner et al. [15] our method was as successful as theirs for a much more complicated scenario. When comparing our research with Fernando and Sooriyarachchi [4] they had a positive correlation which was modeled by a bivariate negative binomial distribution. However, as our responses were negatively correlated that research could not be followed. When comparing our research with Hapugoda and Sooriyarachchi [8] their research modeled survival and count data and the count variable did not have a spike at zero and

had only one random effect. When comparing our research with Sunethra and Sooriyarachchi [19] they too developed a joint model for survival and count variables where they did not have a spike at zero. They used a separate random effect model while we used a shared random effect model with the transformation of one variable. When comparing the newly developed model with the traditional model, overall the newly developed model was superior based on the example used.

Here in this research, it should be mentioned that log transforming the second count variable and modeling it as normal is the same as modeling the untransformed second count variable using the lognormal distribution.

The methods used here build upon a combination of ideas from the literature and these ideas have never been put together before as explained in the literature review. Therefore, the methods developed here and the example analyzed is a novel technique for the scenario considered.

5.3 Limitations of the study

In the example, there was only one explanatory variable in the study. The offset variable was the log of the population size. For some countries there were no values for the population so these countries had to be dropped from the analysis. Most distribution combinations for the two responses did not converge. Both zero-inflated Poisson and Zero-inflated Negative Binomial did not work for the count response with a spike at zero.

5.4 Further Work

On the methodology side, one could write a computer program to incorporate the zeros in the likelihood function without adding unity to the count data with zeros. However, this is not straight

forward as this is a joint model. More covariates and interactions could be implemented using another example. The deviance could be calculated to provide an objective goodness of fit statistic than the residual plot which is subjective.

References

1. Aitchinson J. and Ho C.H. (1989). The Multivariate Log-Normal Distribution. *Biometrika* 76 643 - 653
2. Allison, D. (2012) Do we really need inflated models? https://statisticalhorizons.com/zero-inflated-models retrieved on 15th May, 2020.
3. Collett, D. (1991). Modeling Binary Data. Chapman & Hall/CRC Texts in Statistical Science, USA.
4. Shenali Maryse Fernando, Marina Roshini Sooriyarachchi (2018). Bivariate Negative Binomial Modeling of Epidemiological Data. *Open Science Journal of Statistics and Application* 5(3) : 47-57
5. Gayara Fernando and Roshini Sooriyarachchi (2020): The development of a goodness-of-fit test for high level binary multilevel models, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1700275
6. Gurumu, S. and Elder, J. (2012). Flexible Bivariate Count Data Regression Model. *Journal of Business and Economic Statistics* 30(2) : 265 – 274.
7. Hapugoda, J. C., Sooriyarachchi, M. R., Kalupahana, R. S. and Satharasinghe, D. A. "Joint Modeling of Mixed Responses: An Application to Poultry Data". Proceedings of 5th

Annual International Conference on Operations Research and Statistics (ORS), Singapore, pp. 182-185, 2017.

8. Hapugoda J. and Sooriyarachchi, M.R.(2018). Joint Modeling of Discrete Time Hazard Model with Poisson Regression Model: A Simulation Study. Proceedings of the JUICE conference, University of Jaffna.
9. http://personal.lse.ac.uk/tenreyro/poisson.pdf. Retrieved on 15th May, 2020
10. http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=225&lang=en, retrieved on 25th April 2020
11. https://www.un.org/en/development/desa/population/publications/manual/projection/index.asp. Retrieved on 17th May, 2020
12. Jayanetti, W., and Sooriyarachchi, R. "A multilevel study of dengue Epidemiology in Sri Lanka: modeling survival of dengue patients", International Journal of Mosquito Research, 2 (3), pp. 114-121, 2015.
13. Karunarathna G.H.S. and Sooriyarachchi M.R. (2019). Joint Multilevel Model for Analyzing Length of Stay through Competing End points in Dengue Epidemiology. *Sri Lankan Journal of Applied Statistics* 19(1) : 45 – 60.
14. Kochelerkota S. and Kochelakota K. (1992). Discrete Distributions New York. *Marcel Deckker.*
15. Lorenz, E. Jenkner, C. Sauerbrei, W. Becher, H. (2019). Modeling Exposures with a Spike at Zero: Simulation Study and a Practical Application to Survival Data. *Biostatistics and Epidemiology* 3 (1) : 23 – 37.
16. McDonald, J.H. (2014) Handbook of Biological Statistics (3rd ed.). *Sparky House Publishing*, Baltimore, Maryland pages 140-144.

17. Ophem, H. van (1999) A General Method to Estimate Correlated Discrete Random Variables. *Econometric Theory* 15: 228 -237.
18. Rizopoulos D. (2012).Joint models for longitudinal and time to event data with applications in R. *India CRC Press*.
19. Sunethra A. A. and Sooriyarachchi M. R. (2020). A Novel Method for Joint Modeling of Survival Data and Count Data for both Simple Randomized and Cluster Randomized Data. *Communications in Statistics. Theory and Methods*. Published online.
20. D.Wickramarachchi A Goodness of Fit (GOF) Test for Bivariate Binary Multilevel Logistic Model (unpublished B.Sc. thesis, 2017). University of Colombo, Sri Lanka.

Department of Statistics,
University of Colombo
Colombo 3.
Sri Lanka
19-02-2021

Editor,
Journal of Biostatistics and Epidemiology
Dear Sir / Madam,

Re-submission of Paper TBEP-2020-0037R2

I am re-submitting the paper titled '**Joint Modelling of Two Count Variables using a Shared Random Effect Model in the presence of Clusters for Complex Data'** to your journal under the Biostatistical methods and models section. Included are the full paper with the corrections suggested by the reviewer and Editor, together with tables and figures. The corrections are given in red. Also included are responses to the reviewer and Editor. Please acknowledge the receipt of this paper and Thank you.

Yours Faithfully,

Prof. Roshini Sooriyarachchi

Comments to Editor and Reviewers

1. On page 12, the formula for variance of Possion response is incorrect. An approximation of the variance can be obtained using delta method with exponential function as the transformation function. I would recommend removing the paragraph on the description of the variance and covariance matrix. But please provide a reference instead.

   The required material has been removed and references have been given. The corrections are given in red text.