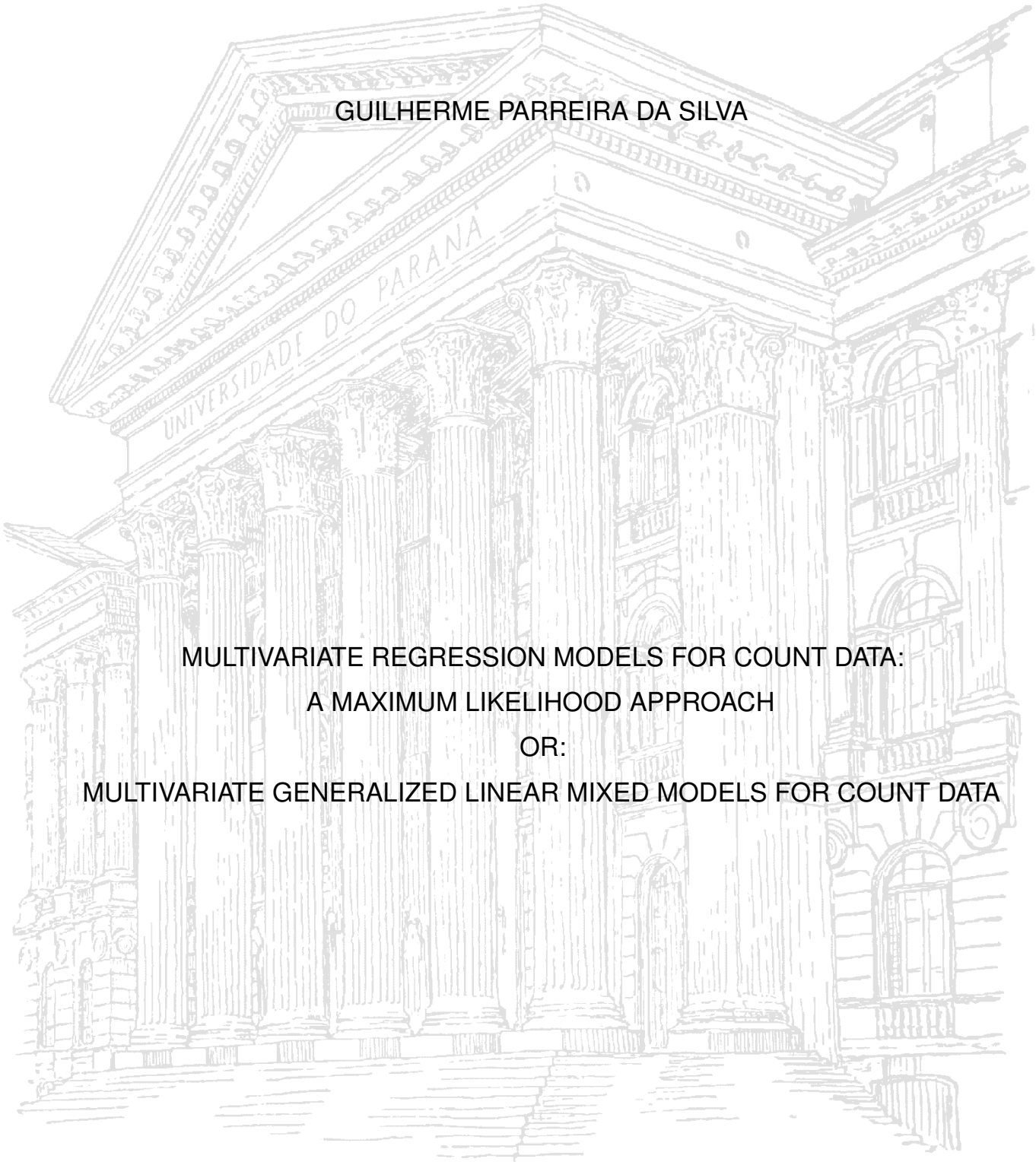
FEDERAL UNIVERSITY OF PARANA

GUILHERME PARREIRA DA SILVA



MULTIVARIATE REGRESSION MODELS FOR COUNT DATA: A MAXIMUM LIKELIHOOD APPROACH OR: MULTIVARIATE GENERALIZED LINEAR MIXED MODELS FOR COUNT DATA

CURITIBA

2021

GUILHERME PARREIRA DA SILVA

MULTIVARIATE REGRESSION MODELS FOR COUNT DATA: A MAXIMUM LIKELIHOOD APPROACH OR: MULTIVARIATE GENERALIZED LINEAR MIXED MODELS FOR COUNT DATA

Master thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Supervisor: Prof. PhD Wagner Hugo Bonat

Co-supervisor: Prof. PhD Paulo Justiniano Ribeiro Júnior

CURITIBA

2021

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

P499m Petterle, Ricardo Rasmussen
Modelo de regressão quase-beta multivariado [recurso eletrônico] / Ricardo Rasmussen Petterle. – Curitiba, 2018.

Dissertação - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia de Produção, 2018.

Orientador: Cassius Tadeu Scarpin – Coorientador: Wagner Hugo Bonat.

1. Análise de regressão. 2. Estatística. 3. Métodos de redes múltiplas (Análise numérica). 4. Métodos de simulação. 5. Algorítmos. I. Universidade Federal do Paraná. II. Scarpin, Cassius Tadeu. III. Bonat, Wagner Hugo. IV. Título.

CDD: 519.536

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

GUILHERME PARREIRA DA SILVA

MULTIVARIATE REGRESSION MODELS FOR COUNT DATA: A MAXIMUM LIKELIHOOD APPROACH OR: MULTIVARIATE GENERALIZED LINEAR MIXED MODELS FOR COUNT DATA

Master thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Master thesis presented in Curitiba, 01 April 2021 .

Prof. PhD Wagner Hugo Bonat
Orientador

Prof. Dr. Paulo Justiniano Ribeiro Júnior
Coorientador

Prof. Dr. Gustavo Valentim Loch
Examinador interno - PPGEP

Prof. Dr. Marcos Augusto Mendes Marques
Examinador interno - PPGEP

Prof. Dr. José Luiz Padilha da Silva
Examinador externo - DEST UFPR

2021

Aos meus pais,
pelo apoio e incentivo.

ACKNOWLEDGEMENTS

XXX

"A simplicidade é o último grau de sofisticação".
(Leonardo da Vinci)

ABSTRACT

Researchers are often interested to understand the relationship between a set of covariates and set of response variables. In order to solve this issue, the use of regression analysis, either linear or generalized linear models is largely applied. However, such models only allow users to specify one response variable at a time. Moreover, it is not possible to directly calculate from the regression model a correlation measure between the response variables. In this master thesis, we propose the Multivariate Generalized Linear Mixed Models, which allows the specification of a set of response variables and to calculate the correlation between them by means of a random effect structure that follows a multivariate normal distribution. We use the maximum likelihood estimation method to estimate all parameters using Laplace approximation to integrate out the random effects. The derivatives are provided by automatic differentiation. The outer maximization is made using general purpose algorithm such as `PORT` and `BFGS`. We delimited this problem studying only count response variables with the following distributions: Poisson, negative binomial (NB) and COM-Poisson. While the first distribution can model only equidispersed data, the second models equi and overdispersed, and the third model all types of dispersion. The model was implemented on software `R` with package `TMB`, based on `C`. In order to evaluate the estimator properties we conducted a simulation study considering four different sample sizes and three different correlation values for each distribution. Unbiased and consistent estimators were found for Poisson and NB distributions; for COM-Poisson, estimators were consistent, but they were biased for dispersion, variance and correlation parameters specially. This model was also applied on three datasets. The first one is from the The National Health and Nutrition Examination Survey, a trivariate underdispersed response variable with 1281 participants. The second is from 30 different sites in Australia that the number of 41 different ant species were registered. The third is from the Australia Health Survey with 5 response variables and 5190 respondents. The last two datasets can be considered as overdispersed by the generalized dispersion index. The COM-Poisson model was the best among the other two competitors, estimating parameters with smaller standard error, and a greater number of significant correlation coefficients. Therefore, the proposed model is capable of dealing with multivariate count response and to measure the correlation between them.

Keywords: COM-Poisson. Simulation Study. Correlation. Template Model Builder. Optimization. Laplace Approximation.

RESUMO

XXX

Palavras-chave: Múltiplas variáveis respostas limitadas. Dados correlacionados. Intervalo unitário. Dados longitudinais. Estudo de simulação. Algoritmo NORTA.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Researchers are often interested to understand the relationship between variables in their different areas of study. For example, nurses are interested to know whether polypharmacy is related to complications after a surgery; veterinarians may be interested to know whether animal welfare is related to meat quality, milk production, among others; administrators may be interested know whether the usage of new policy has improved social indicators. When we have a set of covariates and one specific objective (response variable), this situations can be addressed in the statistical literature by regression models. Certainly, one of the most known and widely applied models is the linear regression (LM) (GALTON, 1886).

LM is widely used (and many times misused) due to its simplicity and the general ordinary least squares estimation procedure, that it is covered in different textbooks in different areas (business, numerical optimization, agronomy, among others). To correct apply it, it is necessary to verify whether the residuals are independent, not autocorrelated with homogeneous variance. These assumptions can be too restrictive depending on the context.

As an alternative, the generalized linear model (GLM) was introduced by Nelder e Wedderburn (1972) as a more flexible approach. GLM is a class of model that generalizes LM by supporting response variables that belongs to the exponential family (contrary to the LM, that can only fit a continuous variable). Moreover, it is built upon a link function that connects the linear predictor to the response variable, and the variance of data can be related to the mean. As members of GLM, we mention the binomial distribution that can fit discrete proportions, gamma and inverse Gaussian distributions to fit asymmetric positive continuous data, normal distribution for symmetric continuous data and Poisson distribution for counting data.

A counting data represents the number of times that an event occurs in a fixed interval, such as, time, space, distance, area, among others. Therefore, it is finite and non-negative. One example of such data, it is the number of times of ear, body posture and head orientation changes in ewes after brushing (it is a treatment proposed to increase welfare in animals) (TAMIOSO et al., 2018). The Poisson distribution is widely used for this purpose but relies on the fact that the variance of the data is equal to its mean, which is known as equidispersion. However, this assumption is too restrictive, and different mean variance relationship can be found, such as overdispersion and underdispersion. Overdispersion occurs when the variance of the data is greater than the mean, and it is often found in practice. It usually happens due to excess of zeros, heavy-tailed distribution, or absence of a covariate to model the data (GRUNWALD

et al., 2011). On the other hand, underdispersion occurs when the variance is smaller than the mean. The reason for this behaviour is still under study and not well defined; however, a underdispersed random variable is characterized by a smaller range of possible value of the continuous variable compared to a overdispersed random variable with same mean.

Different distributions have been proposed to model counting data. When the variance is equal to the mean, after considering the effects of all covariates, Poisson is the most obvious choice. When the data is overdispersed, negative binomial (NB) under the same framework of GLM (although, it is not a pure GLM because NB does not belong to the exponential family) is a good choice. The Extended Poisson Tweedie (BONAT et al., 2018) based on the Poisson Tweedie distribution (EL-SHAARAWI; ZHU; JOE, 2011; JØRGENSEN; KOKONENDJI, 2016), Conway-Maxwell-Poisson (COM-Poisson) (SHMUELI et al., 2005) and Gamma Count(ZEVIANI et al., 2014) distributions can be used to model either under-, equi- or overdispersed data. The drawback of them is that the probability mass function (pmf) does not have a closed form expression, making the process of inference time consuming for procedures that rely on the pmf, such as likelihood.

Beyond of different distributions, different modelling approaches can be used to model count data, specially when the data is zero inflated (RIDOUT; DEMÉTRIO; HINDE, 1998). Two widely know alternatives are hurdle and zero inflated models (ZEILEIS; KLEIBER; JACKMAN, 2008). A Hurdle model is a two-part model, whereas one part specifies the probability of the response assuming the value 0 (usually fitting a logistic regression), the other part determines how many times the event occurred truncated at 0 (common choices for probability distribution are Poisson and NB). Zero inflated models are based on zero-inflated probability distributions. The most common model is the zero inflated Poisson (ZIP), which is classified as a mixture model. Different from hurdle models, whereas by definition all zero counts are predicted just by the logistic regression (and it matches the number of zero in the observed data) and the non-zero counts are fitted by a zero-truncated, in zero-inflated models, there is a process that governs the zero excess, however the distribution for count data is not truncated (LOEYS et al., 2012).

All the strategies pointed out here consider that we have available only one response variable. However, it is not difficult to find in the literature datasets where the researches possess more than 1 response variable for the same study. Usually, the analysis is made by each response individually due to the lack of alternatives in statistical packages. Nevertheless, there is an increasing interest in the literature to develop models or distributions that can handle multivariate responses, that is, when there is more than one response variable (BONAT; JØRGENSEN, 2016).

One approach to model more than one response variable at the same time is to construct multivariate distributions for count data. Inouye et al. (2017) present three alternatives to model multivariate count data. The first assumes that the marginal distribution is Poisson, and a multivariate distribution is build under copulas or multivariate distribution theory (CAMPBELL, 1934). The second uses mixture of independent Poisson. The third method generalizes the first one, whereas the conditional distributions are also Poisson. However, none of them deals with under-dispersed data. Famoye (2015) proposes a multivariate generalized Poisson regression model based on the multivariate generalized Poisson distribution (MGPD) that can deal with equi-, under- or overdispersed data, the correlation estimates can either be positive or negative, and the estimation is made via maximum likelihood (ML) method. Muñoz-Pichardo et al. (2021) proposed a multivariate conditional Poisson regression model, where the dependence between response variables is conditional on the other response variables and is measured into the linear predictor by a regression coefficient.

Winkelmann (2008) provides an overview of different distributions/models for count data. It presents the multivariate NB model (MNBM) and the multivariate Poisson-gamma mixture model (MPGM), which allow only for overdispersion and nonnegative correlation. It also presents the multivariate Poisson-lognormal regression (MPLR) model and the latent Poisson-normal regression model, which allows positive and negative correlation, however, it is suitable only for overdispersed data.

Copula is a general framework to build multivariate distributions from different marginal distributions based on copula functions, Nikoloulopoulos e Karlis (2009) use copula for multivariate count data. However, it is still difficult to choose a Copula to model negative dependence among many random variables. Bonat (2016) proposed the Multivariate Covariance Generalized Linear Models (MCGLM), which it is a class of models based on quasi-likelihood (WEDDERBURN, 1974) and generalized estimating equations (GEE) (LIANG; ZEGER, 1986), that allows to fit multivariate models using only second order moments assumptions with correlated data. Bayesian Regression Models using Stan - brms package (BüRKNER, 2018) and MCMC Generalised Linear Mixed Models - MCMCglmm package (HADFIELD, 2010) provide a framework to model multivariate models via Bayesian inference (DEMPSTER, 1968).

Another alternative is to model correlation between response variables for the same individual using the class of hierarchical GLM (LEE; NELDER, 1996). This class allows to model correlated variables or individuals via a random effect, an unobserved variable, that can follow any distribution. When the distribution of the random effect is Gaussian, we have the Generalized Linear Mixed Models (GLMM). However, GLMM is widely known and used to model correlation between sample units, not for response variables, such methodology is implemented in consolidated packages in software R (R

Core Team, 2020), such as, glmmTMB (BROOKS et al., 2017), lme4 (BATES et al., 2015) and nlme (PINHEIRO et al., 2017).

In this thesis we propose to model multivariate count data under the framework of GLMM in order to accommodate correlation between random variables. The estimation of this model is based on the ML method (ALDRICH et al., 1997), integration of the random effect is made via Laplace approximation (LA) (TIERNEY; KADANE, 1986), maximization of the marginal log-likelihood (result from Laplace integration), is made with traditional optimizers, such as `PORT` algorithm (GAY, 1990) which is available under the `nlminb` function, and BFGS (NOCEDAL; WRIGHT, 2006), using `optim` function. Standard errors (SEs) of the estimates are obtained via delta method (THYGESEN et al., 2017). This framework is implemented using the Template Model Builder (`TMB`) package (KRISTENSEN et al., 2016) in `R`. Three different distributions of the response variables are specified: Poisson, NB and COM-Poisson. Even though NB and COM-Poisson do not belong to the exponential family, we will refer this study under the GLMM framework, once the estimation process remain the same regardless of the distribution being used.

1.1 OBJECTIVES

1.1.1 General Objectives

To propose the multivariate generalized linear mixed model (MGLMM) class based on the GLMM and study it for count data.

1.1.2 Specific Objectives

1. Apply simulation studies to verify the property of the estimators.
2. Computational implementation of the model.
3. Apply the model in three datasets.
4. Estimate and interpret the model parameters.

1.2 JUSTIFICATION

Many studies apply regression models for each model separately, i.e., without accounting for the correlation between random variables. Even tough there are some approaches that can deal with this, none of them is built under the GLMM framework for multivariate data.

1.3 LIMITATIONS

This thesis explores only transversal multivariate count data under the MGLMM framework using only one distribution per study. However, this framework is easily extended to allow mix of different probability distributions, i.e., the first random variable is Poisson and the second is COM-Poisson. Moreover, the MGLMM proposed can also be used to accommodate correlation structure within sample units and different type of data, such as proportion, continuous, among others.

Another possible limitation that it was not studied here, it is the ability of the MGLMM to model an expressive number of response variables simultaneously, such as 200, 300, 500, and more.

1.4 THESIS WORKFLOW

This thesis contains six chapters including this introduction. Chapter two describes three datasets that will be used as an model example of application. Chapter three presents an literature review of the distributions used and the model which MGLMM are based on. Chapter 4 proposes the MGLMM model along with the estimation procedure. Chapter 5 presents the results of the model applied to the data from chapter 2. Finally, chapter 6 discusses the main contribution of this work and future work are also pointed.

2 MOTIVATIONAL DATASETS

This chapter describes three datasets that were used as an application example of the proposed regression model described in chapter 4. The order of them is according to the computational time taken to estimate them. The first consists of a dataset from The National Health and Nutrition Examination Survey (NHANES). The second comprises a dataset from abundances of ant species in south-eastern Australia. Lastly, but not least, a dataset from the Australian Health Survey (AHS).

2.1 DATASET I: NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

NHANES is a program who studies the health and nutrition status of adults and children in the United States. This survey is being conducted every year since the early 1960s. Nowadays, it examines a nationally representative sample of about 5,000 persons each year (NHANES. . . , 2007). NHANES collects different type of data, such as, demographic, air quality, alcohol use, blood pressure & cholesterol, oral health, sexual behaviour. A full list of datasets can be found in National. . . (2007). The main objective of this study was to investigate whether demographical variables influence the sexual behaviour.

In this dissertation we obtained the data using the same procedure as described in Famoye (2015). However, it was not possible to obtain the same final sample that the authors did. After an e-mail contact with Felix Famoye, he could not longer assist with this topic. Even so, we decided to continue to use the dataset with the same filters and procedures used, except that we did not filter respondents by age.

From the list of datasets, it was used the sexual behaviour and demographic data. From the first dataset, it was selected three response variables Nmsp (Number of male sex partners in the past year), Nmosp (Number of male oral sex partners in the past year) and Nspfy (Number of sex partners who are five years older in the past year). From the second dataset, it was used the covariates race (1 = White, 0 = Others), education level (range from 1 = Less Than 9^{th} Grade to 5 = College Graduate or above) and martial status (1 = Married, 0 = Others). After deleting those respondents who had missing data, the sample consists of 1281 women, with age ranging from 18 to 80, 57% were married, 43% white and 31% had some college or AA degree.

Clearly, one way to analyse this dataset is via regression models. Once all three response variables are counting, we need to choose a suiltable pmf for this data. In this thesis, we compared the Poisson, NB and Compoisson probability distributions. As

this is a cross-sectional study, we do not have responses correlated over time; however, we may have correlated responses variables, once they were measured in the same individual. The model proposed in chapter 4 can accommodate this situation.

Table 1 presents the mean, variance, Fisher Dispersion Index (DI) (FISHER, 1934) for every response variable and the generalized dispersion index (GDI) (KOKONENDJI; PUIG, 2018) for the dataset. The main reason for choosing those response variables is the fact that Nmsp and Nmosp can be considered marginally underdispersed, once the sample variance is smaller than the sample mean. Moreover, Nspfy is only a little overdispersed, as it is shown in Table 1. It is easily seen from the DI, which is calculate dividing the variance of the variable by its mean. A variable with DI > 1 can be said as overdispersed, DI = 1 as equidispersed and DI < 1 underdispersed.

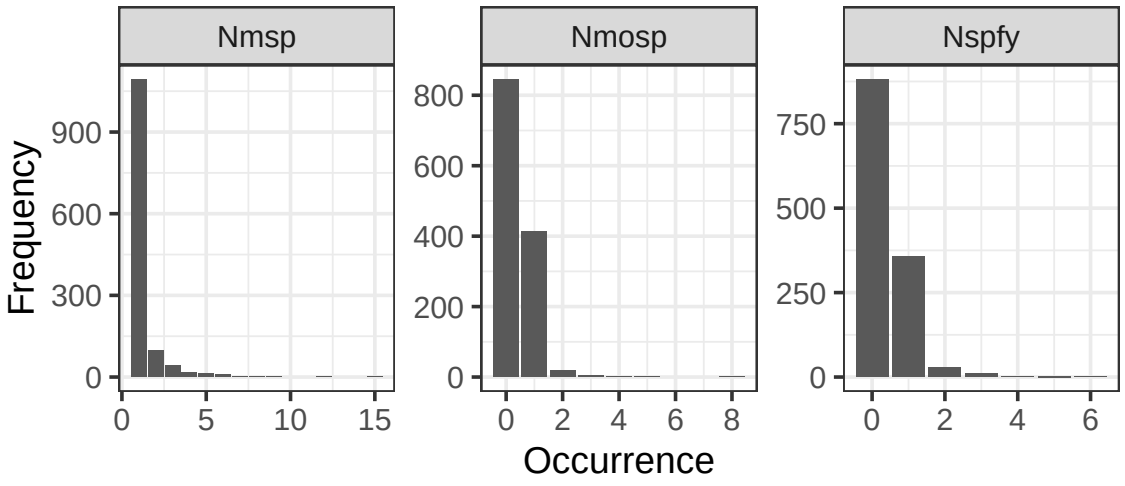
Moreover, the GDI is a recently proposed multivariate dispersion index. It is calculated based on the expectation of each variable and the covariance among them. When the number of variables is equal to 1, it is just the classical Fisher DI. A GDI > 1 classifies the multivariate responses as overdispersed, GDI = 1 as equidispersed and GDI < 1 underdispersed. Standard error (SE) for GDI are calculated using the asymptotic behaviour of the estimator. According to this index, the dataset is overdispersed according to the pontual estimate, but a 95% confidence interval based on SE suggests that this dataset is equidispersed as 1 is included on it.

TABLE 1 – DESCRIPTIVE MEASUREMENTS FOR NHANES RESPONSE VARIABLES

	Spearman Correlation ρ			Mean	Variance	DI	GDI(SE)
	Nmsp	Nmosp	Nspfy				
Nmsp		0.033	0.222	1.313	1.084	0.826	
Nmosp			0.038	0.372	0.350	0.939	1.092(.22)
Nspfy				0.368	0.411	1.117	

The three response variable shows no or small (Nmsp and Nspfy) correlation between them. Figure 1 shows the barplot of each response variable. We can see that there is a higher frequency for non occurrence of events, rather than the occurrency.

FIGURE 1 – BARPLOT FOR EACH RESPONSE VARIABLE FROM NHANES DATA



2.2 DATASET II: ABUNDANCE EPIGAEIC ANT SPECIES

This data was obtained from a study conducted in south-eastern Australia by Gibb et al. (2015), and it is available in the software R (R Core Team, 2020) throughout the `mvabund` package (WANG et al., 2020). The study consisted to count the number of 41 different ant species that fell into a pitfall traps during 18-day sessions in 30 different sites in south-eastern Australia in November 2007 and April 2008. The main interest with this dataset here is to investigate the relationship between the environmental variables with the occurrence of different ant species, and the co-occurrence of ant species.

The response variables considered here are the occurrence of each of the 41 species in the 30 sites, while the covariates comprises 5 environmental variables and the their full description is given in Table 2. The name of the response variables starts with an index number (1,...,41) followed by their abbreviated name.

TABLE 2 – COVARIATES COLLECTED IN THE ANT STUDY IN SOUTH-EASTERN AUSTRALIA

Name	Description
Bare ground	Percent cover of bare ground, as estimated from ten 1x1 metre quadrat
Canopy cover	Percent canopy cover, as estimated from two 20x20m transects
Shrub.cover	Percent canopy cover, as estimated from two 20x20m transects
Volume.lying.CWD	Estimated volume of Coarse Woody Debris in two 20x20m transects, including all debris greater than 5cm diameter
Feral.mammal.dung	Proportion of quadrat including mammal dung, out of ten 1x1m quadrat

Source: (WANG et al., 2020).

Figure 2 presents the barplot for each response variable from ANT data. Among the 41 different species it is seen that there is a high variability among the responses. Some species were only seen 1 time in a single site, such as 35.Polyrhachis and 38.Solenopsis, while 15.Iridomyrmex and 31.Pheidole were seen 20 times in more than one site. In other cases, 8.Cardiocondyla, 9.Crematogaster and 30.Paraparatrechina species were not seen the most of sites, but they had an appearance in a single site with 8, 10 and 10 ants. It is clear that 35^{th} and 38^{th} response variables do not have the behaviour of a count data (only the definition), but of a binary data. Even tough, we are going to try to estimate the count model with these two response variable.

FIGURE 2 – BARPLOT FOR EACH RESPONSE VARIABLE FROM ANT DATA

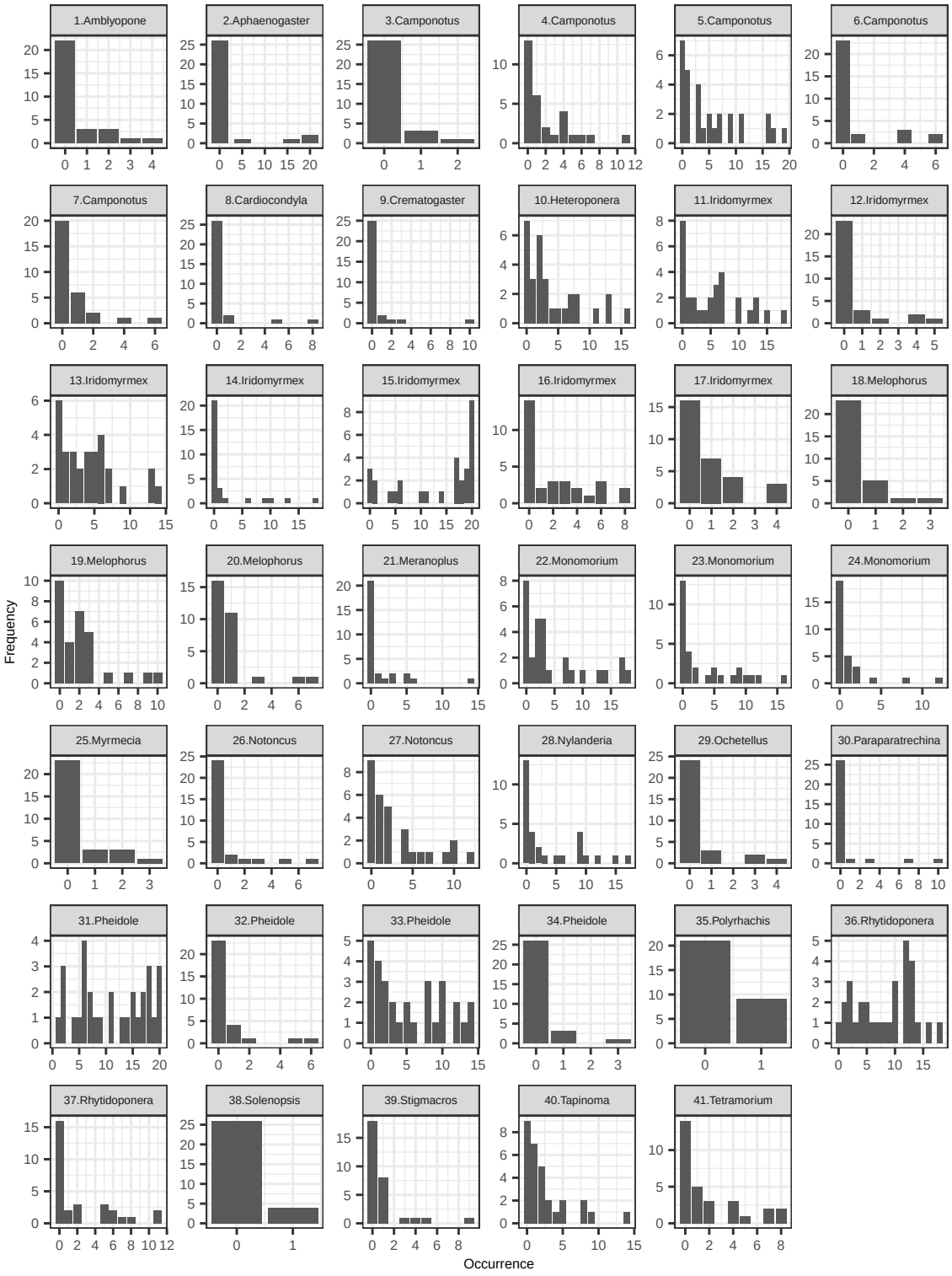


Table 3 presents the mean, variance, Fisher Dispersion Index (DI) (FISHER, 1934) for every response variable and the generalized dispersion index (GDI) (KOKONENDJI; PUIG, 2018) for the data set. The variance is higher than the mean almost for all variables, except for 35.Polyrhachis and 38.Solenopsis. It is easily seen from the DI. This dataset is clearly classified as overdispersed based on the GDI as the inferior limit of a 95%

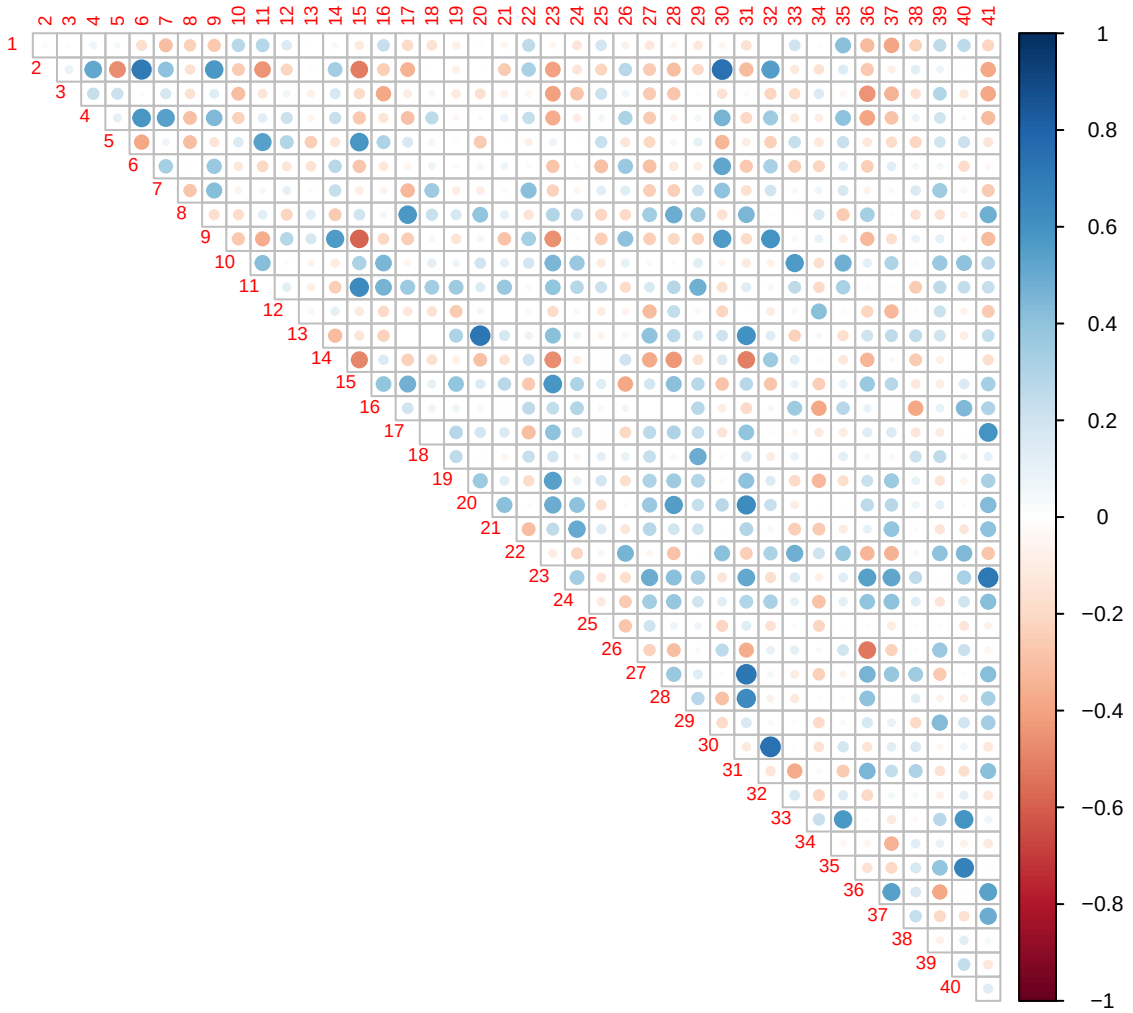
confidence interval does not contain and it is greater than 1.

TABLE 3 – DESCRIPTIVE MEASUREMENTS FROM ANT DATA

	Mean	Variance	DI	GDI(SE)
1.Amblyopone.australis	0.533	1.085	2.034	
2.Aphaenogaster.longiceps	2.033	32.999	16.229	
3.Camponotus.cinereus.amperei	0.167	0.213	1.276	
4.Camponotus.claripes	1.933	7.099	3.672	
5.Camponotus.consobrinus	5.300	33.252	6.274	
6.Camponotus.nigriceps	0.867	3.430	3.958	
7.Camponotus.nigroaeneus	0.667	1.816	2.724	
8.Cardiocondyla.nuda.atalanta	0.500	2.879	5.759	
9.Crematogaster.sp..A	0.567	3.633	6.412	
10.Heteroponera.sp..A	4.067	19.857	4.883	
11.Iridomyrmex.bicknelli	5.333	26.437	4.957	
12.Iridomyrmex.dromus	0.600	1.834	3.057	
13.Iridomyrmex.mjobergi	4.300	15.803	3.675	
14.Iridomyrmex.purpureus	2.033	20.447	10.056	
15.Iridomyrmex.rufoniger	13.300	59.045	4.439	
16.Iridomyrmex.suchieri	2.133	6.809	3.192	
17.Iridomyrmex.suchieroides	0.900	1.610	1.789	
18.Melophorus.sp..E	0.333	0.506	1.517	
19.Melophorus.sp..F	2.133	6.740	3.159	
20.Melophorus.sp..H	0.900	2.783	3.092	11.543(.92)
21.Meranoplus.sp..A	1.333	8.713	6.534	
22.Monomorium.leae	4.733	32.409	6.847	
23.Monomorium.rothsteini	3.433	20.944	6.100	
24.Monomorium.sydneyense	1.167	6.902	5.916	
25.Myrmecia.pilosula.complex	0.400	0.662	1.655	
26.Notoncus.capitatus	0.633	2.654	4.191	
27.Notoncus.ectatommoides	2.900	12.300	4.241	
28.Nylanderia.sp..A	3.733	25.720	6.889	
29.Ochetellus.glaber	0.433	1.082	2.496	
30.Paraparatrechina.sp..B	0.700	4.976	7.108	
31.Pheidole.sp..A	11.100	40.300	3.631	
32.Pheidole.sp..B	0.567	2.047	3.613	
33.Pheidole.sp..E	5.467	22.809	4.172	
34.Pheidole.sp..J	0.200	0.372	1.862	
35.Polyrhachis.sp..A	0.300	0.217	0.724	
36.Rhytidoponera.metallica.sp..A	8.300	25.528	3.076	
37.Rhytidoponera.sp..B	2.400	11.834	4.931	
38.Solenopsis.sp..A	0.133	0.120	0.897	
39.Stigmacros.sp..A	0.967	3.826	3.958	
40.Tapinoma.sp..A	2.533	11.154	4.403	
41.Tetramorium.sp..A	1.933	7.030	3.636	

Figure 3 explores how the occurrence of different species are related to each other visualizing the Spearman correlation in a correlogram. The marginal correlation ranges from -.58 up to .74, which is well distributed along all possible values of the correlation parameter $\rho = [-1,1]$ and it may not difficult the estimation once it is not close to the boundary of the parameter.

FIGURE 3 – CORRELOGRAM OF ANT SPECIES OCCURENCE USING SPEARMAN CORRELATION



2.3 DATASET III: AUSTRALIAN HEALTH SURVEY

The AHS is the largest survey conducted in Australia concerning health issues. The data used here was collected during the years 1987-88 and it is available through the `mcglm` package (BONAT, 2018) in the `ahs` object. The full reference for this dataset can be found in Colin e Trivedi (1998). This dataset is a sample from the survey conducted in 1987-88, which comprises only individuals over 18 years old who have answered all questions (so, there is no missing data) and every respondent was measured only once (cross-sectional study).

The main objective of this study is to investigate whether more acess to health care services and demographic covariates (sex, age, income) are related to the number of times patients use health services. So, we have the same structure as the other datasets presented to analyse: a set of covariates that may influence a set of response variables, and the correlation between response variables.

The data has 5190 respondents and 15 variables, whereas 10 were considered as covariates and 5 counting response variables. The five response variables are Ndoc (Number of consultations with a doctor or specialist), Nndoc (Number of consultations with health professionals), Nadm (Number of admissions to a hospital, psychiatric hospital, nursing or convalescence home in the past 12 months), Nhosp (Number of nights in a hospital during the most recent admission) and Nmed (Total number of prescribed and non prescribed medications used in the past two days). Table 4 gives the description of each covariate:

TABLE 4 – COVARIATES COLLECTED IN THE AUSTRALIA HEALTH SURVEY (AHS)

Name	Description
sex	Factor with levels male and female
age	Respondent's age in years divided by 100
income	Respondent's annual income in Australian dollars divided by 1000
levyplus	Coded factor. If respondent is covered by private health insurance fund for private patients in public hospital with doctor of choice (1) or otherwise (0)
freepoor	Coded factor. If respondent is covered by government because low income, recent immigrant, unemployed (1) or otherwise (0)
freerepa	Coded factor. If respondent is covered free by government because of old
illnes	Number of illnesses in past 2 weeks, with 5 or more illnesses coded as 5
actdays	Number of days of reduced activity in the past two weeks due to illness or injury
hscore	Respondent's general health questionnaire score using Goldberg's method. High score indicates poor health
chcond	Factor with three levels. If respondent has chronic condition(s) and is limited in activity (limited), or if the respondent has chronic condition(s) but is not limited in activity (nonlimited) or otherwise (otherwise, reference level)

Source: (BONAT, 2018).

Figure 4 shows the barplot for each response variable. The figure suggests that all variables have some degree of overdispersion, due to the right heavy-tailed and excess of zeros. Table 5 presents some descriptive statistics for the 5 response variables. It is seen a positive (small in most of cases) correlation between the 5 responses variables, which it is acceptable, once they were measured in the same individual; except between

Nadm and Nhosp, where the correlation is .996 (close to the maximum value). Moreover, the mean and variance relationship between the variables shows overdispersion for all variables, being Nhosp the most overdispersed. Also, the GDI index also classify this dataset as overdispersed.

FIGURE 4 – BARPLOT FOR EACH RESPONSE VARIABLE FROM AHS DATA

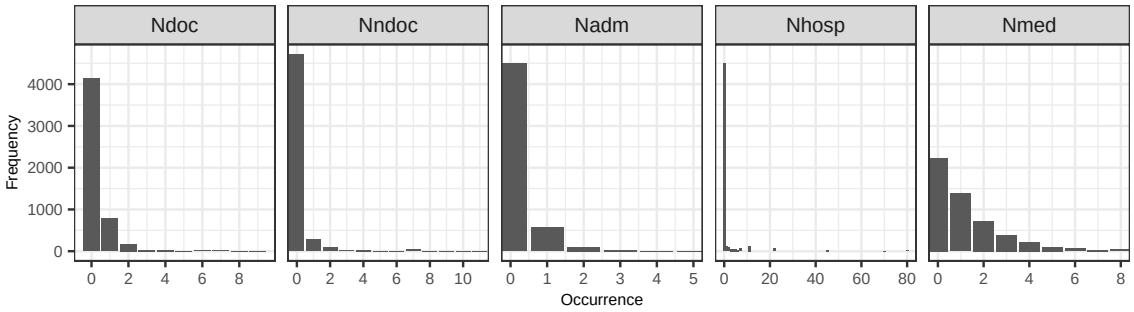


TABLE 5 – DESCRIPTIVE MEASUREMENTS FOR AHS RESPONSE VARIABLES

	Spearman Correlation ρ				Mean	Variance	DI	GDI(SE)
	Nndoc	Nadm	Nhosp	Nmed				
Ndoc	0.11	0.158	0.160	0.290	0.302	0.637	2.111	
Nndoc		0.107	0.113	0.165	0.215	0.932	4.341	
Nadm			0.996	0.135	0.174	0.258	1.484	17.944 (1.99)
Nhosp				0.139	1.334	37.455	28.083	
Nmed					1.218	2.423	1.989	

Although these descriptive analysis are marginal (without taking into account the covariates and model structure), they can provide insights from what to expect from the model.

3 LITERATURE REVIEW

This chapter presents a review of the pmfs used in section 3.1 and the models that serves as a basis of the proposed MGLMM model: GLM in section 3.2 and GLMM in section 3.3.

3.1 PROBABILITY MASS FUNCTION (PMF)

This section presents the three pmf used: Poisson in subsection 3.1.1, NB in subsection 3.1.2 and COM-Poisson in subsection 3.1.3.

3.1.1 Poisson distribution

The Poisson probability distribution is used to model non-negative integers count. It has only a single parameter. The probability mass function for a Poisson random variable is defined as:

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \tag{3.1}$$

and its mean and variance are:

$$\mathbb{E}[Y] = \mathbb{V}[Y] = \lambda.$$

So, the mean is equal to the variance of the distribution and both are defined by the lambda parameter, that control the shape of the pmf.

3.1.2 Negative binomial distribution

The NB distribution is defined as a Poisson-gamma mixture distribution. Therefore, it can be seen as a generalization of the Poisson distribution. The NB distribution is derived from the Poisson distribution with its mean parameter λ, and λ is gamma distributed. The idea of this definition is that there may be other sources of variability that are not modelled by the Poisson model, but can exists in the real world, such as unobserved covariates.

Consider again that y is a Poisson random distribution with probability mass function defined in Equation 3.1. Now, consider that λ is a gamma random variable with parameters ϕ and β. This NB distribution is given by the following hierarchical specification:

$$\begin{aligned} \mathrm{y} &\sim P(\lambda) \\ \lambda &\sim G(\phi, \beta), \end{aligned}$$

where the probability density function of a gamma random variable is:

$$f(\lambda; \phi, \beta) = \frac{\beta^{\phi}}{\Gamma(\phi)} \lambda^{\phi-1} e^{-\beta\lambda}.$$

What we want to know is the marginal distribution of Y, that can be found via:

$$[\mathrm{Y}] = \int_0^{\infty} [\mathrm{Y}/\lambda]\,[\lambda]\, d\lambda. \tag{3.2}$$

Equation 3.2 is solved analytically as follow:

$$\begin{aligned} [\mathrm{Y}] &= \int_0^{\infty} [\mathrm{Y}/\lambda]\,[\lambda]\, d\lambda \\ &= \int_0^{\infty} \frac{\lambda^{\mathrm{y}} e^{-\lambda}}{\mathrm{y}!} \frac{\beta^{\phi}}{\Gamma(\phi)} \lambda^{\phi-1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^{\phi}}{\mathrm{y}!\Gamma(\phi)} \int_0^{\infty} e^{-\lambda} \lambda^{\mathrm{y}} \lambda^{\phi-1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^{\phi}}{\mathrm{y}!\Gamma(\phi)} \int_0^{\infty} e^{-\lambda(\beta+1)} \lambda^{\mathrm{y}+\phi-1} d\lambda \\ &= \frac{\beta^{\phi}}{\mathrm{y}!\Gamma(\phi)} \frac{\Gamma(\mathrm{y}+\phi)}{(\beta+1)^{\mathrm{y}+\phi}} \int_0^{\infty} \frac{(\beta+1)^{\mathrm{y}+\phi}}{\Gamma(\mathrm{y}+\phi)} e^{-\lambda(\beta+1)} \lambda^{\mathrm{y}+\phi-1} d\lambda \\ &= \frac{\Gamma(\mathrm{y}+\phi)}{\mathrm{y}!\Gamma(\phi)} \frac{\beta^{\phi}}{(\beta+1)^{\mathrm{y}+\phi}} \\ &= \frac{\Gamma(\mathrm{y}+\phi)}{\Gamma(\mathrm{y}+1)\Gamma(\phi)} \left(\frac{\beta}{\beta+1}\right)^{\phi} \left(\frac{1}{\beta+1}\right)^{\mathrm{y}} \end{aligned}$$

where $\phi \geq 0$, $\beta \geq 0$ and $\mathrm{y} = 0, 1, 2, \ldots$. The mean and variance of Y are given by:

$$\mathbb{E}[\mathrm{Y}] = \phi\beta$$

and

$$\mathbb{V}[\mathrm{Y}] = \phi\beta(1+\beta) = \mathbb{E}[\mathrm{Y}](1+\beta)$$

therefore, the variance of a NB random variable will exceeds its mean, unless in the limit case when $\beta \to 0$, resulting in variance equals to the mean. In order to use the NB distribution in a regression model, it is usually necessary to use a mean reparametrization. So, we can define:

$$\lambda = \phi\beta$$

where λ is the expected value. Therefore, $\beta = \lambda/\phi$, and the variance takes the following form:

$$\mathbb{V}[Y] = \lambda(1 + \frac{\lambda}{\phi})$$

This model is called as "NB-II", and it is known as the quadratic parametrization of the variance (because of the quadratic relationship of the mean and variance. In this model, when $\phi \to \infty$ the NB-II model converges to a Poisson model (equidispersion), for all other values, it models overdispersion.

To summarize, $Y \sim \text{NB}(\lambda, \phi)$ with probability mass function given by:

$$f(y;\lambda,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\lambda}{\lambda+\phi}\right)^{\phi} \left(\frac{\phi}{\lambda+\phi}\right)^{y} \tag{3.3}$$

where ϕ can be interpreted as a dispersion parameter, controlling the variability of Y. The mean and variance are given by:

$$\mathbb{E}[Y] = \lambda$$

and

$$\mathbb{V}[Y] = \lambda + \frac{\lambda^2}{\phi}.$$

There are different parametrization of the NB distribution. One that deserves to mention is the number of times that a success occurs in a sequence of Bernoulli trials before a specified number of failure occurs. It can be obtained via definition, or reparametrization of Equation 3.3, setting $p = \phi/(\lambda+\phi)$ as the probability of success and ϕ as the number of failures. Both definitions generate the same distribution, while the one deduced here is the most natural in context of regression. Two books on this topic are Winkelmann (2008) and Hardin e Hilbe (2018).

3.1.3 COM-Poisson distribution

The Conway-Maxwell-Poisson (COM-Poisson) Distribution was originally proposed by Conway e Maxwell (1962) to model queuing systems and it is an extension to the Poisson distribution that can model both under and equidispersed data. Shmueli et al. (2005) reintroduced the CMP distribution into the statistical peer. Consider that Y follows a CMP distribution with rate parameter λ and dispersion parameter ν that has probability mass function (pmf) given by

$$\mathbb{P}(Y = y;\lambda,\nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda,\nu)}, y = 0,1,2,\ldots,$$

where $Z(\lambda,\nu) = \sum_{y=0}^{\infty} \lambda^y/(y!)^\nu$ is a normalising constant. As an alternative to the original model, Huang (2017) proposed a mean parametrized CMP model, which it is

useful in regression scenario and it will be used in this thesis. The pmf of this new parametrization with parameters $\mu \geq 0$, the mean, and $\nu \geq 0$, the dispersion, is given by:

$$\mathbb{P}(Y = y; \mu, \nu) = \frac{\lambda(\mu,\nu)^y}{(y!)^\nu} \frac{1}{Z(\lambda(\mu,\nu),\nu)}, y = 0,1,2,\ldots,$$

and the rate $\lambda(\mu,\nu)$ is a function of μ and ν and is obtained by the following equation:

$$\sum_{y=0}^{\infty} (y - \mu) \frac{\lambda^y}{(y!)^\nu} = 0. \tag{3.4}$$

The parameter ν plays an important role into this distribution. If $\nu = 0$ the COM-Poisson distribution tends to a Geometric distribution, if $\nu = 1$ a Poisson, and as $\nu \to \infty$ a Bernoulli. Therefore, when $\nu < 1$ we are in a situation of overdispersion, $\nu = 1$ equidispersion and finally, $\nu > 1$ underdispersion. Another important feature is that μ and ν are orthogonal. Orthogonality between parameters in a regression model makes inference easier once the value of one parameter does not influence the other, making the optimization process easier.

On the other hand, it is important to note that there are two infinite sums in the COM-Poisson pmf, one for $\lambda(\mu,\nu)$ and another for $Z(\lambda,\nu)$. In the first infinite sum, it is also necessary to find the root for λ in Equation 3.4. These features makes the pmf time consuming for a CMP random variable.

3.2 GENERALIZED LINEAR MODEL (GLM)

Probability mass functions are useful to describe behaviour of a variable, but they are not capable to study the relationship between a set of covariates and a response variable. To overcome this, a regression model is a suitable tool for this task, allowing (usually) the mean of the distribution to vary for each sample unit according to the covariates. In particular, a GLM can be used on this situation. A GLM is a class of models that can be characterized by four main components:

- The distribution of the response variable belongs to the exponential family.
- A link function that connects the linear predictor to the response variable. It has to be a one-to-one (admits inverse), monotonic and differentiable function.
- Variance is proportional to the mean.
- Observations are independent.

It extends the LM once can accommodate different probability distributions that belong to the exponential family, such as Poisson for count data, gamma and inverse

Gaussian for positive skewed continuous data and binomial for discrete proportion. A great advantage is that they can deal with heterogeneous variance of the residuals, once the variance is proportional to the mean. Once these distributions can assume a different range of values (only positive, negative/positive, $[0,1]$) it is necessary a link function to connect the response values to the linear predictor.

In summary, let $Y_1, Y_2, \ldots, Y_n$ be independent random variables and each of them $Y_i (i = 1, \ldots, n)$ follows a probability function f from the exponential family. The GLM is specified by:

$$\begin{aligned} Y_i &\sim f(\mu_i, \phi) \\ g(\mu_i) &= X\boldsymbol{\beta}, \end{aligned} \tag{3.5}$$

where μ_i is the mean for every sample unity, ϕ is the nuisance parameter for every distribution (which it is 1 for Poisson and Binomial), $g(.)$ is a suitable link function, $X\boldsymbol{\beta}$ is the linear predictor that comprises the design matrix of observed covariates X of size $n \times p$ and a vector of parameters $\boldsymbol{\beta}$ of size $p \times 1$.

As we will be dealing with a count response variable, we will consider f as one of the pmf described in section 3.1 and log as the link function. Even tough NB and COM-Poisson are not pure GLM (because they do not belong to the exponential family), they will be considered as GLM because they share the same model structure. The main difference between them is that Poisson can take advantage of the efficient IRLS algorithm to maximize the likelihood function and the others cannot.

The choice for the log link function is due to the fact that it links the average value of the response variable (μ_i) to the linear predictor ($X\boldsymbol{\beta}$). We can note that the linear predictor can range from $[-\infty; +\infty]$, while $\mu_i \, \varepsilon \, \mathbb{R}^+$, thus μ_i cannot receive negative values from the linear predictor. Once we apply $\log(\mu_i)$, the left hand side of Equation 3.5 will accept negative values too (as well positive values). It is important to observe that $\boldsymbol{\beta}$ will be in the scale of $\log(\mu_i)$, and no longer in μ_i. Therefore, once we estimate those parameters in log scale, we can exponentiate them to be able to interpret $\boldsymbol{\beta}$ as a mean ratio, multiplicatively in the scale of μ_i.

When we work with count data, a typical situation occurs when we are interested in the number of infected people in every neighbourhood in a given city. The number of infected people will vary according to the number of population (o_i) in the neighbourhood. As we know the total population a priori we can consider this information as an constant term in the linear predictor, which it is know as offset. In order to accommodate this situation in GLM specification, we only need to change Equation 3.5 to:

$$\begin{aligned} g(\mu_i) &= g(o_i) + x_i\boldsymbol{\beta} \\ \log(\mu_i) &= \log(o_i) + x_i\boldsymbol{\beta}. \end{aligned}$$

Note that o_i is in same scale as μ_i and there is no parameter associated to it. It guarantees that the relationship between μ_i and o_i is linear and no extra parameter will be estimated, so, it does not make the model more complicated to be estimated. The offset notation will not be used in this thesis to make notation shorter, however, it is easily extended to the other class of models presented.

3.3 GENERALIZED LINEAR MIXED MODELS (GLMM)

Even tough GLM is a flexible approach, the assumption of independent observations is still too restrictive. It is easy to see studies where dependent data occurs. For example, in longitudinal studies where a biomaker is measured more than one time in the same patient; in grouped data, when students from same class and same schools are more correlated to each other whether compared to students from other classes and schools.

In order to address dependent observations, overdispersion, among others, it was proposed the GLMM (BRESLOW; CLAYTON, 1993) class. It accommodates the extra variability by means of an unobservable variable, which it is considered in the model as a random effect (additional to linear predictor) and it is calculated for each individual. The original part of the model, or the usual linear predictor of the GLM model is known as fixed effect. Once the random effect is an unobserved random variable, it is necessary to include a hierarchical structure to Equation 3.5. The following notation is valid for repeated measurements and longitudinal studies:

$$\begin{aligned} Y_{il} \mid \mathbf{b}_{i\cdot} &\sim f(\mu_{il}, \phi) \\ g(\mu_{il}) &= X\boldsymbol{\beta} + Z\mathbf{b}_{i\cdot} \\ \mathbf{b}_{i\cdot} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{3.6}$$

where Y_{il} is the i-th unit sample measured in the l-th times (in a longitudinal study) or in the l-th group (in a repeated measurements study), $\mathbf{b}_{i\cdot}$ is a vector of random effects for each sample unit that follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ to accommodate the variance of each random effect and covariance among them, f is the conditional distribution of Y_{il} on $\mathbf{b}_{i\cdot}$, $g(\mu_{il})$ is the mean for every sample unit i and measurement l, and Z is the design matrix associated to the random effect.

4 MULTIVARIATE GENERALIZED LINEAR MIXED MODEL FOR COUNTING DATA

This chapter presents the new regression model used to analyse counting response variables, the MGLMM for counting data. The section 4.1 presents the structure of the model and the section 4.2 shows the proposed method for parameter estimation.

4.1 MULTIVARIATE GENERALIZED LINEAR MIXED MODEL (MGLMM) FOR COUNT DATA VIA ML

Let Y_{ir} being the multivariate outcome for subject i, where $i = 1,\ldots,n$ and response variable r, where $r = 1,\ldots,k$. Suppose a set of p known covariates is available for each response r, therefore, x_{irj} is the value of the *j-th* covariate for individual i and response r. The purpose of this thesis is to provide a joint model for a set of response variable. We start from the standard GLMM model specification with only a random intercept. The first thing we need to assume is the conditional distribution of the response variables:

$$Y_{ir} \mid b_{ir} \sim f(\mu_{ir};\phi_{ir}),$$

where all response variables shares the same distribution, however, with different mean and dispersion parameters for each response variable. The second thing we need to specify is the linear predictor:

$$g_r(\mu_{ir}) = x_{ir}^T\boldsymbol{\beta}_r + b_{ir},$$

where $g_r(.)$ is a suitable link function, $\boldsymbol{\beta}_r$ is a $p \times 1$ vector of parameter estimates and b_{ir} is the random intercept value for each individual and response variable. Lastly, the distribution of the random effects is specified by:

$$\begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{ir} \end{pmatrix} \sim \text{NM}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \underset{r\times r}{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1r}\sigma_1\sigma_r \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \ldots & \rho_{2r}\sigma_2\sigma_r \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{r1}\sigma_r\sigma_1 & \rho_{r2}\sigma_r\sigma_2 & \ldots & \sigma_r^2 \end{bmatrix}\right),$$

where each random effect has mean 0, variance σ^2, and $\rho_{rr'}(r \neq r')$ measures the correlation between each pair of random effects. Despite this is a general framework modelling, this thesis addresses only the same distribution (either Poisson, Binomial Negative or COM-Poisson) and link function (logarithm) for all responses variables.

An important question for this type of model is whether it is possible to estimate simultaneously the dispersion parameters of the pmf and the variance parameters of the random effects, once they measure variability related of the same random variable, and can cause identifiability problems.

4.2 INFERENCE AND ESTIMATION

In this section, we present the estimation method used to obtain the parameter regression estimates based on the likelihood function. The likelihood function can be understood as the probability of observing the sample data as a function of parameters values. The likelihood function is obtained by the joint probability distribution function of all random variables; when they are independent, the likelihood function is just the product of their probability distribution function. The main point here is that the observed data is treated as fixed, while parameters can assume a range of plausible values, building the likelihood function.

The set of parameter values that maximize the probability of observing the data is considered as the best estimates of the likelihood function. This estimation method is known as ML, and the best estimates for the parameters are known as the ML estimates (MLE). A full reference for this topic can be found in Pawitan (2001) and Casella e Berger (2002).

In order to estimate the model presented in section 4.1 under the ML method, it is necessary to obtain the joint distribution for both random variables (Y_{ir} and b_{ir}). Once we have a hierarchical structure, the joint distribution of Y_{ir} and b_{ir} can be factored into $f(\mathrm{Y}_{ir},\mathrm{b}_{ir}) = f(\mathrm{Y}_{ir} \mid \mathrm{b}_{ir})f(\mathrm{b}_{ir})$. As only Y is observed, we are interested on the marginal distribution of Y that can be obtained integrating out b_{ir}, that is, $f(\mathrm{Y}_{ir}) = \int f(\mathrm{Y}_{ir} \mid \mathrm{b}_{ir})f(\mathrm{b}_{ir})\mathrm{db}_{ir}$, which in practical terms, the joint distribution is averaged over the b_{ir} terms.

Now, we particularize the ML estimation method to the model described in section 4.1. The objective is to estimate the parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma^2}, \boldsymbol{\rho}\}$, where $\boldsymbol{\beta}$ is the regression parameter vector, $\boldsymbol{\phi}$ the extra parameters of each distribution (dispersion in most of cases), and $\boldsymbol{\sigma^2}$ and $\boldsymbol{\rho}$ are the parameters that compose the variance covariance matrix $\boldsymbol{\Sigma}$. The marginal likelihood function for the model described in section 4.1 for each sample unit is

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \int \prod_{r=1}^{k} f(\mathrm{y}_r \mid \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\phi}) f(\mathbf{b} \mid \boldsymbol{\Sigma}) \mathrm{d}\mathbf{b}, \tag{4.1}$$

the full likelihood for $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \prod_{i=1}^{N} f(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\phi}),$$

where N is the total number of sample units. Under independence between sample units, we have:

$$\mathbf{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \prod_{i=1}^{N} \int \prod_{r=1}^{k} f(\mathrm{y}_r \mid \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\phi}) f(\mathbf{b} \mid \boldsymbol{\Sigma}) \mathrm{d}\mathbf{b}. \tag{4.2}$$

The Equation 4.2 is composed by the product of two probability distributions. The first one is the probability distribution of the sample units, while the second is the probability distribution of the random effects. The distribution of the random effects is assumed to be a multivariate normal distribution. When both sample and random effects distribution is normal, the integral can be solved analytically, otherwise, numerical methods are required to solve the integral.

4.2.1 Numerical integration via Laplace approximation

Once we are dealing with non normal data, one possibility it is to use numerical integration methods to solve the integral in Equation 4.1 for each individual. Other possibilities include Penalized Quasi Likelihood (PQL) and Monte Carlo techniques that will not be covered here. It is important to note that the integral is of dimension k, and k can vary from two to r (the number of response variables). Methods based on numerical quadrature such as trapezium, 1/3 Simpson, Gauss-Hermite or adaptative Gauss-Hermite use too many points to evaluate the integrand and its complexity is proportional to the dimension of the integral (BONAT; JR, 2016). In order to solve Equation 4.1 we will use the LA, that it is a special case of the adaptive Gauss-Hermite quadrature (AGHQ), when it is used only one integration point (TIERNEY; KADANE, 1986). Altough, LA is faster than AGHQ, it is less accurate (SIGNORELLI; SPITALI; TSONAKA, 2020).

The main idea of LA is to replace the integrand in Equation 4.1 by a closed-form expression, which solves the integral. The resulting expression is maximized with respect to $\mathbf{b}$ keeping all other parameters constant and the expression is evaluated, consequently, the integral is solved. The approximation is obtained by

$$\int_{\mathbb{R}^k} \exp\{\mathbf{Q}(\mathbf{b})\} \mathrm{d}\mathbf{b} \approx (2\pi)^{n/2} \left| -\mathbf{Q}''(\hat{\mathbf{b}}) \right|^{-1/2} \exp\{\mathbf{Q}(\hat{\mathbf{b}})\}, \tag{4.3}$$

where $\mathbf{Q}(\mathbf{b})$ is a uni-modal and bounded function of the variable $\mathbf{b}$ that has same k-dimension as the integral. In this case, $\mathbf{Q}(\mathbf{b})$ is the integrand obtained by Equation 4.1, $\mathbf{b}$ the random effect, $\hat{\mathbf{b}}$ the maximized random effect values, while all other $\boldsymbol{\theta}$ parameters

remain constant. Therefore, $\mathbf{Q}(\hat{\mathbf{b}})$ is the maximum value of the function with respect to $\mathbf{b}$ and $\mathbf{Q}''(\hat{\mathbf{b}})$ is the curvature of the function in the maximum, or in other words, it corresponds to the value of the second derivative (Hessian) of the function in the maximum.

It is important to say that the integral of order k in Equation 4.1 has to be solved repeatedly in every step of the ML method for each one of the N sample units. Therefore, we have two optimizations procedures: one that it is called the outer maximization, that it is the maximization of the marginal likelihood (result from the integral), and another one that it is called the inner optimization, which is the maximization required inside the LA.

While the inner process maximization can be efficiently implemented by a Newton-Raphson (NR) method (that requires second-order derivatives) due to the fact that the integrand function is fully available, the outer process maximization is harder to obtain derivatives because the marginal likelihood is in fact a number, not a function; therefore, it will be maximized with algorithms that require only first-order derivatives. NR is described in subsection 4.2.2, derivatives calculation are described in subsection 4.2.3 and the outer maximization in subsection 4.2.4.

4.2.2 Inner Optimization - Newton's method

Optimizing a function consists to select parameter values that maximize a function, that is:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}\, \varepsilon\, \mathbb{R}} f(\boldsymbol{\theta}), \tag{4.4}$$

where f is a smooth function and $\boldsymbol{\theta}$ is the vector of unrestricted parameters. In this thesis, f is also nonlinear, so that we can classify this problem as an unconstrained nonlinear optimization (unconstrained because the parameters can assume any value).

In standard optimization programs, Equation 4.4 is usually written as $\min f(\boldsymbol{\theta})$ instead of $\max f(\boldsymbol{\theta})$. As we are under the context of the ML, we decided to present the previous equation as maximization. Moreover, $\max f(\boldsymbol{\theta}) \equiv -\min -f(\boldsymbol{\theta})$. Therefore, if a program has the objective to minimize a function and we want to use it, we need to add the minus sign in front of the output of the function to `maximize`, and then multiply it by -1 to get the results back.

Before going into details about maximization procedures, it is important to recall that $\max f(\boldsymbol{\theta}) \equiv f'(\boldsymbol{\theta}) = 0$. Therefore, maximize a function is equivalent to solve nonlinear equations of its derivative (find the root of a function). It makes room for the class of algorithms that solves nonlinear equations. Under this class, we will use the efficient Newton's method, also known as Newton-Raphson method, which has quadratic convergence (the smallest among this class). In simple words, quadratic

convergence means that the number of correct digits approximately doubles every iteration.

Newton's method method consists to find a $\boldsymbol{\theta}$ set that approximates to the root of f. Let start with $\boldsymbol{\theta_0}$ being the initial guess to the root, f the function to find the root and its derivative f' (NOCEDAL; WRIGHT, 2006). Equation 4.5 finds the next root candidate for f (this function is derived from the tangent line equation solving for $\boldsymbol{\theta}_1$ when y is zero: the root):

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - f(\boldsymbol{\theta}_0) f'(\boldsymbol{\theta}_0)^{-1}, \tag{4.5}$$

and $\boldsymbol{\theta}_1$ is typically a better root candidate than $\boldsymbol{\theta}_0$. This process is repeated until the convergence criteria are met using the following recurrence formula:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - f(\boldsymbol{\theta}_n) f'(\boldsymbol{\theta}_n)^{-1}. \tag{4.6}$$

Remember that we need to maximize the integrand in Equation 4.1 with respect to $\mathbf{b}$. Therefore, f is the first derivative of the integrand in Equation 4.1 with respect to $\mathbf{b}$, f'^{-1} is the inverse of the second derivative of the integrand, and finally, $\boldsymbol{\theta}_n$ are the random effects vector $\mathbf{b}$. While f essentially tells the direction of the new root, f' informs the size of the step (how far from the current estimate should $\boldsymbol{\theta}_{n+1}$ be). f' is the Hessian matrix, that informs the curvature of the function. Section subsection 4.2.3 discuss how to obtain the derivatives of a given f function. Once the integrand is not convex there is no guarantee that the solution found is a global maximum; only a local maximum is guaranteed if the convergence criteria are met in the end of the iterations.

For this method to work it is necessary a initial guess to be close to the optimum. Once we know that $\mathbb{E}[\mathbf{b}] = 0$ by the model definition, a suitable initial guess is to assume that $\mathbf{b}_0 = 0$. All other model parameters are held constant and are updated in every step of the outer maximization step. The initial guess for the constant parameters are chosen according to their type (small positive values for variance, 0 for correlation and mean parameters ($\boldsymbol{\beta}$), and values that correspond to equidispersion or small overdispersion for the dispersion parameter $\boldsymbol{\phi}$ when necessary.

4.2.3 Automatic differentiation

In this section, we will discuss some methods to obtain the derivatives of a function with focus on automatic differentiation (AD). The differentiation is used to obtain the derivatives of $\mathbf{Q}(\mathbf{u})$ in the LA and to maximize the marginal likelihood.

A derivative of a function can be defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{4.7}$$

and can be interpreted as the tangent line of the function $f(x)$. While the first derivative of a function points out the maximum value of the function with respect to its

parameters, the second derivative represents the curvature (how narrow the function is).

There a number of ways that it can be achieved: analytical (”by hand”), symbolical, numerical and AD. The first consists in memorize a number of rules obtained from definition and the hard work consists in applying these rules by hand to the function. The second way consists in the symbolical differentiation made by computers, for example: `WolphramAlpha` (LLC,). In these two methods, the output is the derived function.

In opposite, the later two methods return the value of the derivative and a computer program that calculates the derivative, respectively. The numerical differentiation consists in solving Equation 4.7 by picking a small value to h, such as $0,00001$ for example, and the result is the derivative value for x. Lastly, in simple words, the AD is a program that recursively applies the chain rule on a function, and applies the derivatives only in the minimal part of the function, such as, elementary arithmetic operation and elementary functions (such as exponential, logarithm, sine*and*cosine) (BAYDIN et al., 2017).

Let’s start with an example to illustrate AD. Suppose we want to differentiate the following function:

$$f(x) = e^{e^x+(e^x)^2+\sin(e^x+(e^x)^2}. \tag{4.8}$$

The natural way of doing that is to apply analytical differentiation, and the result is:

$$f'(x) = e^{e^x+(e^x)^2}e^x + 2(e^x)^2 + \cos(e^x + (e^x)^2)(e^x + 2(e^x)^2). \tag{4.9}$$

As we can see, the derived function in Equation 4.9 is larger and more complex to evaluate than the function in Equation 4.8. Moreover, there are some identical calculations (for example, $e^{\{e^x+(e^x)^2\}}$) that happen in both Equation 4.9 and Equation 4.8 that could be saved if a smart programming technique is used. As there are some repeated calculations, we can define some intermediate variables to avoid redundant calculus:

$$\begin{aligned} a &= \exp^x \\ b &= a^2 \\ c &= a + b \\ d &= \exp^c \\ e &= \sin c \\ f &= d + e. \end{aligned}$$

To represent how these variables are related to each other, it is convenient to draw a computational graph, that represents a function in its elementary operations, as described in Figure 5.

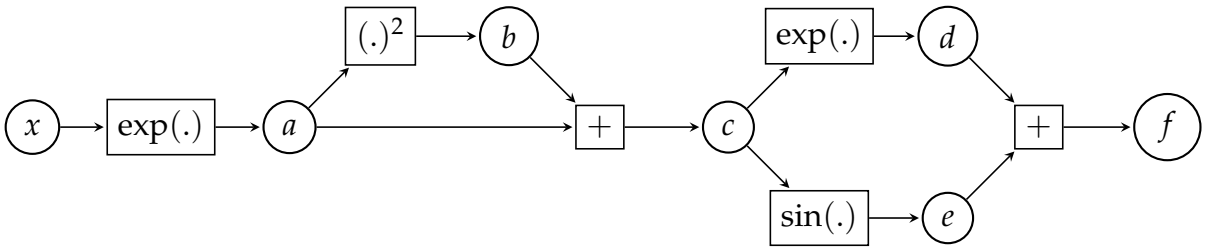


FIGURE 5 – COMPUTATIONAL GRAPH OF THE FUNCTION $f(x) = e^{\{e^x+(e^x)^2+\sin(e^x+(e^x)^2\}}$

We can then start writing out the derivatives of f with respect to x. From Figure 5 it is easily seen that $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x}$, and $\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a}$, and we repeat the chain rule until $\frac{\partial f}{\partial d}$ and $\frac{\partial f}{\partial e}$. Given that we defined all the derivatives, we can start solving it from the final part of the graph to its beginning and reusing the intermediate values of the derivatives. This sequence is described in Equation 4.10.

$$\begin{aligned}
\frac{\partial f}{\partial d} &= \frac{\partial (d+e)}{\partial d} = 1. \\
\frac{\partial f}{\partial e} &= \frac{\partial (d+e)}{\partial e} = 1. \\
\frac{\partial f}{\partial c} &= \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c} = 1e^c + 1cos(c). \\
\frac{\partial f}{\partial b} &= \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} = \frac{\partial f}{\partial c}1. \\
\frac{\partial f}{\partial a} &= \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a} = (e^c + \cos(c))2\mathrm{a} + (e^c + \cos(c))1. \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a}\frac{\partial a}{\partial x} = (e^c + \cos(c))2\mathrm{a} + (e^c + \cos(c))e^x.
\end{aligned} \tag{4.10}$$

The terms in blue in Equation 4.10 are calculated only once and are reused. This procedure is also known as backpropagation in the context of neural networks (DEISENROTH; FAISAL; ONG, 2020). The reverse mode of AD works for Equation 4.1, once the function is the type of $f : \mathrm{R}^n \to \mathrm{R}$ (many inputs to one output). When we have many outputs to one input, the forward mode is advisable, and the main difference is that the calculation order is the opposite; but this is not the case in here.

Some advantages/characteristics of using AD are:

- The function $f(x)$ does not need to be in closed form.
- The computation of $f'(x)$ can be done in the same time complexity as computing $f(x)$.

- Exact to machine precision.
- In TMB $\text{cost}(\nabla f) \leq 4\text{cost}(f)$.

One disadvantage of AD is that it is hard to compute the derivative of a function with if statements. However, we do not have this problem in the statistical modelling, once the likelihood function does not contain if statements. For parameters that have restriction in the parameter space, reparametrization can be used to handle it and are discussed on subsection 4.3.1. For example, instead of modelling σ^2, where $\sigma^2 \varepsilon \mathrm{R}_*^+$, we can use $\exp\{\sigma^2\}$ with input $\log(\sigma^2)$ to make sure that $\sigma^2 > 0$, instead of using $\text{if}(\sigma^2 < 0, \sigma^2 = 0.1)$.

4.2.4 Outer optimization - quasi Newton's method

In subsection 4.2.2 we described the Newton's method that requires second order derivatives to maximize a function. However, the Hessian is not always simple to compute or available for a function f. In order to overcome this, quasi-Newton methods were proposed as alternatives to Newton's method. Instead of using the Hessian itself, they use an approximation to the Hessian. So, they still use the first order derivative information (such as the gradient descendent optimization method), with an easier matrix to compute, that can substitute the Hessian to obtain some idea about the curvature of the function.

One characteristics of quasi-Newton's method is that they belong to the class of line search methods: first the direction of the new $\boldsymbol{\theta}_{n+1}$ is chosen based on the gradient, and then the step size is obtained via an approximation to the Hessian (NOCEDAL; WRIGHT, 2006). Another optimization strategy belongs to the trust region methods, where firstly a region for the next $\boldsymbol{\theta}_{n+1}$ is chosen, and then the direction.

A general quasi-Newton method is described in the following way:

1. $\Delta\boldsymbol{\theta}_n = -\alpha_n \mathbf{B}_n^{-1} f'(\boldsymbol{\theta}_n)$, with α chosen to satisfy the Wolfe conditions.
2. $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \Delta\boldsymbol{\theta}_n$.
3. The difference of gradients in the new point $f'(\boldsymbol{\theta}_{n+1})$ is calculated $\mathbf{y}_n = f'(\boldsymbol{\theta}_{n+1}) - f'(\boldsymbol{\theta}_n)$ and is used to update the approximate Hessian $\mathbf{B}_{n+1}$, or directly its inverse $\mathbf{H}_{n+1} = \mathbf{B}_{n+1}^{-1}$ using the Sherman-Morrison formula.

The question that arises here is how to obtain an approximate matrix to Hessian (or its inverse), say $\mathbf{B}_n$ (or $\mathbf{B}_n^{-1}$), that it is easy to obtain and does not slow down too much the computation. The idea is to use the secant equation:

$$f^{'}(\boldsymbol{\theta}_{n+1}) - f^{'}(\boldsymbol{\theta}_n) = \mathbf{B}_{n+1}(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \tag{4.11}$$

and solve it for $\mathbf{B}_{n+1}$. Obtaining $\mathbf{B}_{n+1}$ from the secant equation is analogous to approximate the Hessian using finite difference. It is also analogous to a Taylor series approximation of order two over the function f around $\boldsymbol{\theta}_n$. Therefore, either Newton's method or quasi-Newton methods are based on a quadratic approximation of f around the maximum $\hat{\boldsymbol{\theta}}$; as more quadratic the nonlinear function surface is, easier is the maximization process. Solve Equation 4.11 for $\mathbf{B}_{n+1}$ is easy with one dimension (it is only a division) and not trivial for a n dimensional parameter set. It is important to recall that $\mathbf{B}_{n+1}$ has to be positive definite to represent the Hessian, and inverting the matrix is necessary to compute SEs in the statistical paradigm.

Therefore, different quasi-Newton methods were proposed in order to obtain $\mathbf{B_{n+1}}$. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm obtains $\mathbf{B_{n+1}}$ via:

$$\mathbf{B}_{n+1} = \mathbf{B}_n + \frac{\mathbf{y}_n \mathbf{y}_n^\top}{\mathbf{y}_n^\top \Delta\boldsymbol{\theta}_n} - \frac{\mathbf{B}_n \Delta\boldsymbol{\theta}_n \Delta\boldsymbol{\theta}_n^\top \mathbf{B}_n^\top}{\Delta\boldsymbol{\theta}_n^\top \mathbf{B}_n \Delta\boldsymbol{\theta}_n}.$$

A problem with this, it is that we would still need to obtain the inverse of approximated Hessian. However, it can be calculated directly in BFGS algorithm via:

$$\begin{aligned} \mathbf{B}_{n+1}^{-1} = \mathbf{B}_n^{-1} & \\ + \frac{(\Delta\boldsymbol{\theta}_n^\top \mathbf{y}_n + \mathbf{y}_n^\top \mathbf{B}_n^{-1} \mathbf{y}_n)(\Delta\boldsymbol{\theta}_n \Delta\boldsymbol{\theta}_n^\top)}{(\Delta\boldsymbol{\theta}_n^\top \mathbf{y}_n)^2} & \\ - \frac{\mathbf{B}_n^{-1} \mathbf{y}_n \Delta\boldsymbol{\theta}_n^\top + \Delta\boldsymbol{\theta}_n \mathbf{y}_n^\top \mathbf{B}_n^{-1}}{\Delta\boldsymbol{\theta}_n^\top \mathbf{y}_n}. & \end{aligned} \tag{4.12}$$

It is important to say that quasi-Newton methods do not require to explicitly invert the approximation to the Hessian. They can be programmed in a way that the computation of $\mathbf{B}_{n+1}^{-1}$ is straight-forward, as it was shown for BFGS in Equation 4.12. Moreover, if the initial guess for $\mathbf{B}_0$ is an identity matrix, the first step of the algorithm is the same as the gradient descendent, and the cost of inverting an identity matrix is negligible in the first step. A requirement for most quasi-Newton methods (inclusively BFGS) is to store the $\mathbf{B}_n$ or $\mathbf{B}_n^{-1}$ matrix because it is updated (and improved) in every step of the algorithm. Newton's method instead require the solution of a nonlinear system.

Moreover, as BFGS does not require matrix inversion, its computational complexity is $\mathcal{O}(n^2)$, compared to $\mathcal{O}(n^3)$ in Newton's method, that requires solving a nonlinear system of equations in order to obtain the inverse of the Hessian. Thus, the cost per iteration in BFGS is smaller than Newton's method. However, the convergence is linear (as the majority of quasi-Newton methods), which is slower than the quadratic convergence of Newton's method. Therefore, while the cost per iteration is higher in Newton's method, the number of iterations is smaller.

The cheaper optimization technique to compute will depend on the problem and specially in the cost of inverting the Hessian (greater the cost, better may be to use a quasi-Newton method).

Another important quasi-Newton method is PORT. This algorithm was meant to be portable over different types of computer (PORT) and is based on a FORTRAN library by David Gay from Bell Labs. The approximated Hessian is calculated on the following way:

$$\begin{aligned}\mathbf{B}_{n+1} = \mathbf{B}_n & \\ &+ \frac{(\mathbf{y}_n - \mathbf{B}_n\Delta\boldsymbol{\theta}_n)(\mathbf{B}_n\Delta\boldsymbol{\theta}_n)^\top + \mathbf{B}_n\Delta\boldsymbol{\theta}_n(\mathbf{y}_n - \mathbf{B}_n\Delta\boldsymbol{\theta}_n)^\top}{\Delta\boldsymbol{\theta}_n^\top\mathbf{B}_n\Delta\boldsymbol{\theta}_n} \\ &- \frac{\Delta\boldsymbol{\theta}_n^\top(\mathbf{y}_n - \mathbf{B}_n\Delta\boldsymbol{\theta}_n)(\mathbf{B}_n\Delta\boldsymbol{\theta}_n)(\mathbf{B}_n\Delta\boldsymbol{\theta}_n)^\top}{(\Delta\boldsymbol{\theta}_n^\top\mathbf{B}_n\Delta\boldsymbol{\theta}_n)^2}\end{aligned}$$

4.3 SOFTWARE IMPLEMENTATION

In this section, we present the software implementation to fit the model present in section 4.1. The software used is `R` version 4.0.2 (R Core Team, 2020) along the Template Model Builder - `TBM` (KRISTENSEN et al., 2016) package. `TMB` is a `R` package that offers a collection of tools to build complex random effects statical models through `C++` templates. `TMB` is based on state-of-the art software: `CppAD` (Bell BM, 2005), `Eigen C++` (GUENNEBAUD; JACOB et al., 2010), `BLAS` (BLACKFORD et al., 2002), among others libraries written in `C++`, which are responsible for obtaining the derivatives through AD, linear algebra computations and parallelization, respectively.

Once the user supplies an objective function in a C++ template file (which, usually is the negative log-likelihood function), it obtains the marginal likelihood via LA integrating out the **b** random effects. The internal optimization of the LA is made via Newton's method, with first and second derivatives provided via AD from `TMB`. After that, the marginal can be optimized to obtain the MLE parameters with any general-purpose algorithm (using first derivative information provided from `TMB`), such as `PORT` or `BFGS` implemented in `nlminb` and `optim` functions in R, respectively, presented in subsection 4.2.4. Moreover, the standard deviation of the parameters (or a function of it) can be obtained via Delta method; profiling is available too.

4.3.1 Reparametrization

When dealing with optimization procedures the parameters estimated can vary on $\mathbb{R}$, as it was shown in Equation 4.4. However, parameters can have bounds, i.e. limitation on its parameter space, as it was shown on section 4.1. For example, along the optimization procedures, the algorithm can return a value greater than 1 for the

correlation parameter ρ, and the function will return an error. In order to accommodate this type of situations, it is necessary to use reparametrization when needed for every parameter.

For $\boldsymbol{\beta}$ it is not necessary to use any kind of reparametrization, once it is allowed to vary on $\mathbb{R}$. Dispersion parameters for both NB and COM-Poisson distributions are allowed to vary only on $\mathbb{R}^+$. For this case, a typical reparametrization is to apply the `exponential` function to the returned parameter value from the optimization procedure in the pmf to ensure that the dispersion parameter will vary on $\mathbb{R}^+$. As a side effect, the initial guess for this parameter has to be in `log` scale, not in the natural value, because inside the density this value will be exponentiate, so cancelling out the `log/exp` functions.

For σ we could also use the `log` and `exp` functions, and for ρ the Fisher Z-transformation with its inverse. However, `TMB` uses an efficient reparametrization for these two parameters under the normal multivariate mode, and it will be used.

Instead of estimating $\boldsymbol{\rho}$ directly, a different unrestricted parameter $\boldsymbol{\varrho}$ is estimated in a lower triangular matrix with unit diagonal L, and an unstructured symmetric positive definite correlation matrix $\boldsymbol{\Omega}$ can be obtained via:

$$\boldsymbol{\Omega} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{L}'\mathbf{D}^{-\frac{1}{2}},$$

where:

$$\mathbf{D} = \text{diag}(\mathbf{L}\mathbf{L}'),$$

For example, to estimate a model with 4 response variables, we want to estimate 6 ((4*4-4)/2) correlation parameters. Thus, the lower triangular matrix L has order 4 and is filled row-wise, such that:

$$\mathbf{L} = \begin{pmatrix} 1 & & & \\ \varrho_0 & 1 & & \\ \varrho_1 & \varrho_2 & 1 & \\ \varrho_3 & \varrho_4 & \varrho_5 & 1 \end{pmatrix}.$$

According to the structure defined in subsection 4.3.1, the restrictions imposed into $\boldsymbol{\Omega}$, and the way the normal multivariate was coded into `TMB`, the variances parameters of the random effects do not require reparametrization, however, they need to be supplied as standard deviations. In order to obtain the full variance-covariance matrix $\boldsymbol{\Sigma}$ we can use:

$$\boldsymbol{\Sigma} = \mathbf{W}\boldsymbol{\Omega}\mathbf{W},$$

where W is a diagonal matrix with entries being equal to the random effects's standard deviation:

$$\mathbf{W} = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \sigma_3 & \\ & & & \sigma_4 \end{pmatrix}.$$

4.3.2 Software Implementation Example

This section presents an illustration of TMB package usage. In particular, it is shown how to simulate data, how to write the joint likelihood and its estimation process.

The simulation process can be done from TMB via template file or R directly. We chose to simulate the data from R once we are used to it. The function defined in Code 4.1 simulates data from a bivariate Poisson regression model. Code comments are provided after symbol comments (# for R and // for C++).

Code 4.1 – R FUNCTION THAT SIMULATES A BIVARIATE POISSON RANDOM VARIABLE

```
rpoisson_bi <- function(beta1, beta2, true_rho, n, s2_1, s2_2, seed){
  # beta_i is the value of the regression parameter of the i_th response
      variable (maximum of two values allowed)
  # true_rho is the correlation between the two random variables
  # n is the bivariate sample size
  # s2_i is the variance of the i_th random effect
  # seed is the seed number to guarantee reproducibility
  set.seed(seed)
  if (length(beta1)==1){
    X <- matrix(rep(1,n), ncol = 1)
    colnames(X) <- "(Intercept)"
  } else {
    x1 <- rnorm(n)
    X <- model.matrix(~x1)
  }
  # Random effects with full covariance matrix
  Sigma <- matrix(NA, 2, 2)
  Sigma[1,1] <- s2_1
  Sigma[2,2] <- s2_2
  Sigma[1,2] <- true_rho*sqrt(s2_1)*sqrt(s2_2)
  Sigma[2,1] <- true_rho*sqrt(s2_1)*sqrt(s2_2)
  U <- rmvnorm(n, mean = c(0,0), sigma = Sigma)
  mu1 <- exp(X%*%beta1 + U[,1])
  mu2 <- exp(X%*%beta2 + U[,2])
  ## Response variables
  Y1 <- rpois(n, lambda = mu1)
  Y2 <- rpois(n, lambda = mu2)
  return(list(Y1 = Y1, Y2 = Y2,
```

```
28                X1 = X,
29                X2 = X))
30 }
```

In parallel, it is necessary to create the TMB template file with the joint negative log-likelihood function for a bivariate Poisson regression model, as it is shown in Code 4.2. Note that the probability distribution functions are written in the same R style.

Code 4.2 – JOINT NEGATIVE LOG-LIKELIHOOD FUNCTION FOR A BIVARIATE POISSON REGRESSION MODEL IN C++ TEMPLATE FILE IN TMB

```
1  // Poisson Bivariate (2 responses) same fixed Effects.
2  #include <TMB.hpp>
3  template<class Type>
4  Type objective_function<Type>::operator() ()
5  {
6    using namespace density;
7    // Macros created to read the data from R easier than pure C++ code
8    DATA_VECTOR(Y1);
9    DATA_VECTOR(Y2);
10   DATA_MATRIX(X1);
11   DATA_MATRIX(X2);
12   PARAMETER_VECTOR(beta1);
13   PARAMETER_VECTOR(beta2);
14   PARAMETER_VECTOR(sigma);
15   PARAMETER_VECTOR(rho);
16   PARAMETER_MATRIX(U);
17
18   // It allows to run the code in parallel
19   parallel_accumulator<Type> nll(this);
20
21   int n = Y1.size(); // Number of rows
22   // Density for 1st response variable
23   vector<Type> mu1(n);
24   mu1 = exp(X1*beta1 + U.col(0).array());
25   nll -= sum(dpois(Y1, mu1, true));
26   // Density for 2nd response variable
27   vector<Type> mu2(n);
28   mu2 = exp(X2*beta2 + U.col(1).array());
29   nll -= sum(dpois(Y2, mu2, true));
30
31   // Random Effect with correlation matrix parametrization from TMB
32   for(int i = 0; i < n; i++)
33     nll += VECSCALE(UNSTRUCTURED_CORR(rho), sigma)(U.row(i));
34
35   // It converts "theta" to rho, the correlation parameter
36   matrix<Type> Cor(2,2);
37   Cor = UNSTRUCTURED_CORR(rho).cov();
```

```
38   ADREPORT(Cor);
39
40   return nll;
41 }
```

Once the joint negative log-likelihood was written, the pre and post-processing is made via R as Code 4.3 shows. As we chose to simulate data, we use the function rpoisson_bi in Code 4.1 to generate the data from the bivariate Poisson regression model with predefined parameter values. Instead, we could have just imported some bivariate count data.

After that, we provide the inputs (initial values for parameters based on the true parameter value and dataset) for MakeADFun function in order to integrate out the random effects (argument random in MakeADFun), and consequently, obtains the marginal likelihood and its first derivative. Its output is passed to an optimizer function, in this case, nlminb, and all remaining parameters are estimated. Finally, SEs are obtained via delta method.

Code 4.3 – R CODE TO ESTIMATE THE PARAMETERS OF A BIVARIATE POISSON REGRESSION MODEL

```
1  rm(list = ls())
2  library(TMB)
3  library(mvtnorm)
4  setwd("~/GoogleDrive/Mestrado/dissertacao/TMB/Simulation_Study/Poisson")
5  model <- "02_poisson_bivariate" #1st try
6  # Compile the .cpp template file
7  compile(paste0(model, ".cpp"), flags = "-O0 -ggdb")
8  dyn.load(dynlib(model))
9  # Load a function to simulate a bivariate Poisson RV
10 source("~/GoogleDrive/Mestrado/dissertacao/TMB/Simulation_Study/functions_
       to_simulate.R")
11 ## Data simulation ----------------------------------------
12 n_resp <- 2       # Number of response variables
13 n <- 1000          # Number of observations
14 beta1 <- log(2)   # Beta parameter of first random variable
15 beta2 <- log(.5)  # Beta parameter of second random variable
16 seed <- 778       # Seed for reproducibility
17 ## Random effects covariance matrix
18 s2_1 <- 0.3       # Var for first random effect  .54 is the sd
19 s2_2 <- 0.15      # Var for second random effect .38 is the sd
20 true_rho <- .5    # Correlation between two random variables
21 ## Input data ---------------------------------------------
22 data <- rpoisson_bi(beta1, beta2, true_rho, n, s2_1, s2_2,seed)
23 ## Start Parameters ----------------------------------------
24 # Start parameters equal to the true values. Note that it is passed the
       standard deviation, not the variance of the random effect
```

```
params <- list(beta1 = beta1,
               beta2 = beta2,
               sigma = c(sqrt(s2_1),sqrt(s2_2)),
               rho = true_rho,
               U = matrix(0, ncol = n_resp, nrow = n))
## Compiling ----------------------------------------------
# This function creates the negative loglik and gradient function
obj <- MakeADFun(data = data,
                 parameters = params,
                 DLL = model,
                 hessian = T,
                 silent = T,
                 random = c("U"))
# The object created is optimized through nlminb
fit1 <- nlminb(start = obj$par, objective = obj$fn,gradient = obj$gr,
               control = list(eval.max = 1e8, iter.max = 1e8,
                              abs.tol = 1e-04, rel.tol = 1e-04))
fit1
# It calculates the standard error of the parameters
rep <- sdreport(obj)
summary(rep, c("fixed"), p.value = T)
summary(rep, c("report"), p.value = T)
```

Code 4.4 presents a multivariate version of Code 4.1.

Code 4.4 – C++ TEMPLATE FILE IN TMB FOR A MULTIVARIATE POISSON REGRESSION MODEL

```
// Poisson Multivariate (n responses) same fixed Effects.
#include <TMB.hpp>
template<class Type>
Type objective_function<Type>::operator() ()
{
  using namespace density;
  DATA_MATRIX(X);          // n x p
  PARAMETER_MATRIX(beta);  // p x r
  DATA_MATRIX(Y);          // n x r
  PARAMETER_MATRIX(U);     // n x r
  PARAMETER_VECTOR(rho);   // r(r-1)/2
  PARAMETER_VECTOR(sigma); // r

  // Type nll = 0;
  parallel_accumulator<Type> nll(this);

  int n = Y.rows(); // Number of rows
  int c = Y.cols();  // Number of cols
  matrix<Type> Xbeta(n, c);
  Xbeta = X*beta;
```

```
21
22    vector<Type> Yj(n);
23    vector<Type> mu(n);
24    for (int j = 0; j<c; j++){ //Density for response
25      Yj = Y.col(j);
26      mu = exp(Xbeta.col(j).array() + U.col(j).array());
27      nll -= sum(dpois(Yj, mu, true));
28    }
29
30    for(int i = 0; i < n; i++) // Density for R.E.
31      nll += VECSCALE(UNSTRUCTURED_CORR(rho), sigma)(U.row(i));
32
33    matrix<Type> Cor(c,c);
34    Cor = UNSTRUCTURED_CORR(rho).cov();
35    ADREPORT(Cor);
36
37    return nll;
38  }
```

5 RESULTS

This chapter presents the simulation study results for bivariate regression count models for the distributions presented in section 3.1 to study the properties of the MLE estimators. Moreover, it presents the results of the multivariate regression count model for the three datasets discussed in chapter 2 for each distribution.

5.1 SIMULATION STUDY

In this section, we present a simulation study for every pmf described in section 3.1 to study the properties of the MLE estimators (bias and consistency) in the proposed model presented in section 4.1. In particular, it was considered a bivariate regression model for count data. We designed 12 simulation scenarios with four different sample sizes, 100, 250, 500, 1000, and three different correlation between random effects, $\rho = -0.5, 0, 0.5$. For the regression structure, it was considered only a intercept for each response, with $\beta_{01} = \log(7)$ and $\beta_{02} = \log(1.5)$. The variance of random effects was $\sigma_1^2 = .3$ and $\sigma_2^2 = .15$. The dispersion parameter for NB and COM-Poisson was equal to $\phi = 1$ and $\nu = .7$ respectively, which induces a small overdispersion.

It was generated 150, 200 and 300 datasets for Poisson, COM-Poisson and NB distribution for each design. The primary idea was to generated 100 datasets for each distribution. However, as the SE of the estimates were not calculate in every repetition (especially for NB), it was necessary to increase the number of datasets generated proportionally to the number of SE failure for each distribution in order to obtain at least 100 valid estimations. This result will be explored in the results section for every distribution.

The following three subsection presents the results for Poisson, NB and COM-Poisson.

5.1.1 Poisson

Figure 6 presents the average bias and confidence interval based on the mean SE by sample size and simulation scenario for bivariate Poisson regression model. Figure 6 shows that for all simulation scenarios both the average bias and SE tend to 0 as the sample size increases. This shows the consistency and unbiasedness of the MLE estimator.

FIGURE 6 – AVERAGE BIAS AND CONFIDENCE INTERVAL BASED ON THE MEAN SE BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE POISSON REGRESSION MODEL

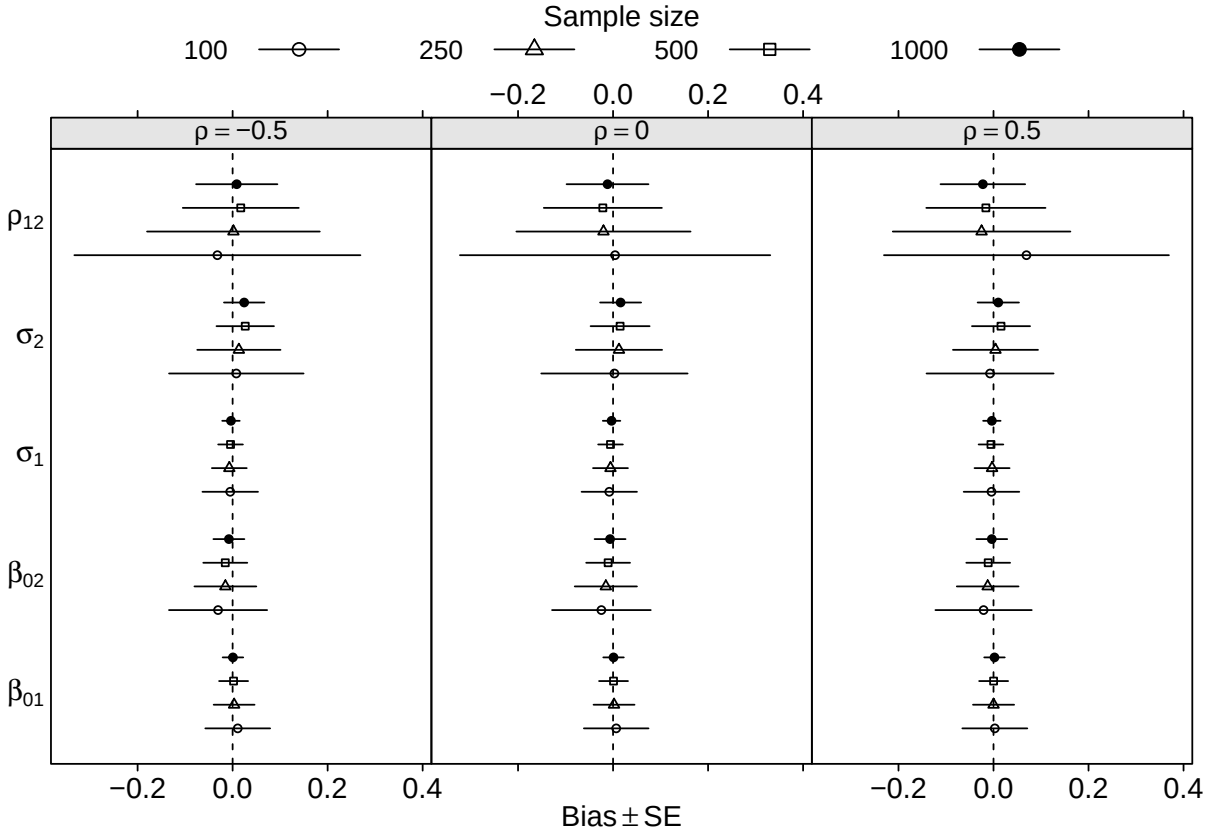
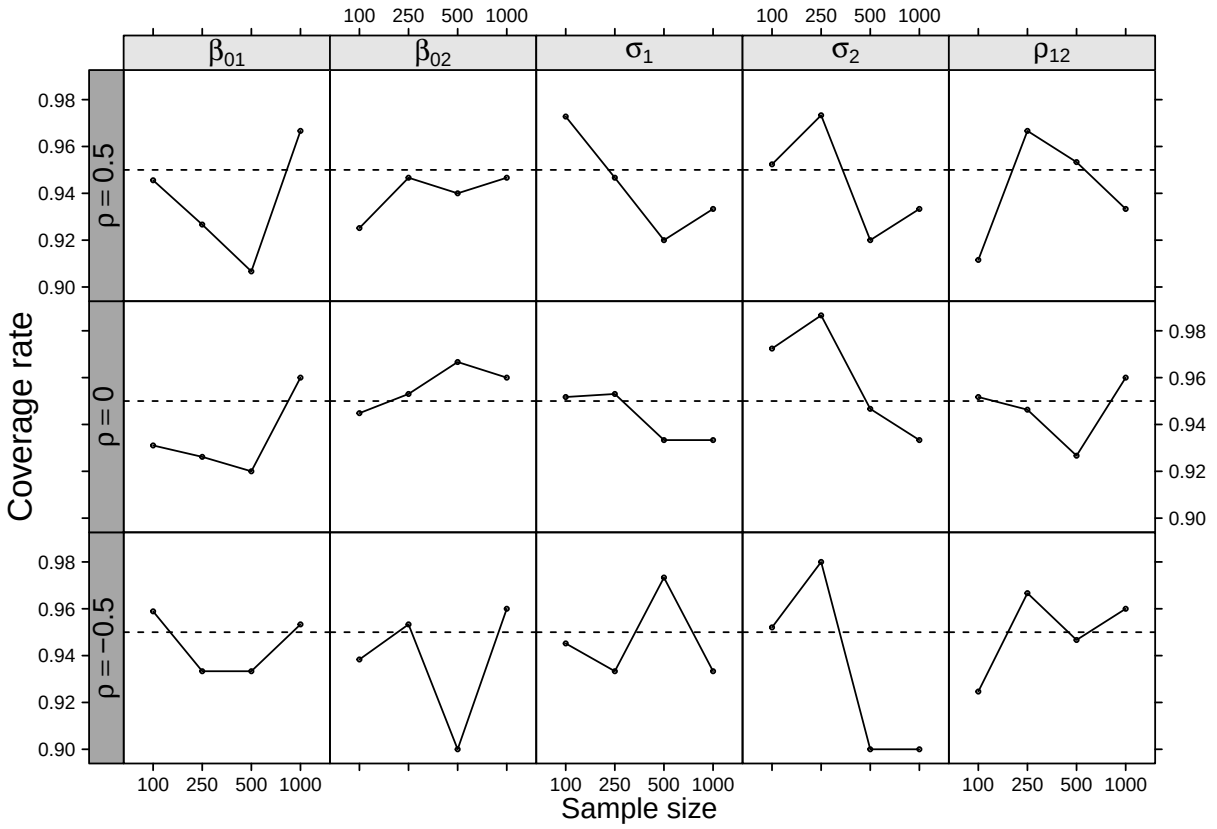


Figure 7 presents the coverage rate for each parameter by sample size and simulation scenario for bivariate Poisson regression model. Overall, all empirical coverage rates are close to the nominal level of 95%, varying between 90% and 98% approximately. In particular, the coverage rate of regression parameters β_{0r} are slightly greater than the nominal level. For the variance of random effects σ_r , the coverage rate is slightly slower than the nominal level. Finally, for the correlation between random effects ρ, the coverage rate is slightly lower when $\rho = .5$, and slightly greater when $\rho = \{0, -0.5\}$ compared to the 95% nominal level.

Even for Poisson distribution, which is the simplest case to estimate because there is no dispersion parameter, there were some iterations that did not calculate SE for some parameters or produced extreme values due to large SEs that were not considered into the results. It was used the `PORT` algorithm to estimate the model because it was more stable and faster than `BFGS` in most of situations (it was not clear when `BFGS` was better than `PORT`). It also happened in Kristensen et al. (2016), where 9 study cases from different model settings ranging from linear regression to multivariate stochastic volatility models were considered, and `PORT` had a better performance than `BFGS`.

FIGURE 7 – COVERAGE RATE FOR EACH PARAMETER BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE POISSON REGRESSION MODEL



The SE failure problem occurred when SE of a parameter was not calculated and this sample was discarded. Extreme values occurred when the algorithm was able to compute the SEs, but the estimates were too extreme due to a large SE. For all distributions, the iteration was discarded when the ρ's SE was greater than 8. Table 6 presents this results.

TABLE 6 – SUMMARY OF FAILURES IN ESTIMATING THE BIVARIATE POISSON REGRESSION MODEL

Problem type	ρ	Sample Size	Fail	Simulations	% Fail
SE fail	-0.5	100	3	150	2.00
SE fail	0.0	100	5	150	3.33
SE fail	0.0	250	1	150	0.67
SE fail	0.5	100	3	150	2.00
Extreme values	-0.5	100	1	150	0.67

There were 12 SE fail type and only one due to extreme values. Note that 12 failures happened when the sample size was 100, and only one with a sample size of 250. Moreover, the percentage of failure did not surpass 3.5% for any scenarios considered.

5.1.2 NB

Figure 8 presents the average bias and confidence interval based on the mean SE by sample size and simulation scenario for bivariate NB regression model. Figure 8 shows that for all simulation scenarios both the average bias and SE tend to 0 as the sample size increases. This shows the consistency and unbiasedness of the MLE estimator. Concerning the variance of the random effect, in general σ_1 and σ_2 were slightly underestimated.

FIGURE 8 – AVERAGE BIAS AND CONFIDENCE INTERVAL BASED ON THE MEAN SE BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE NB REGRESSION MODEL

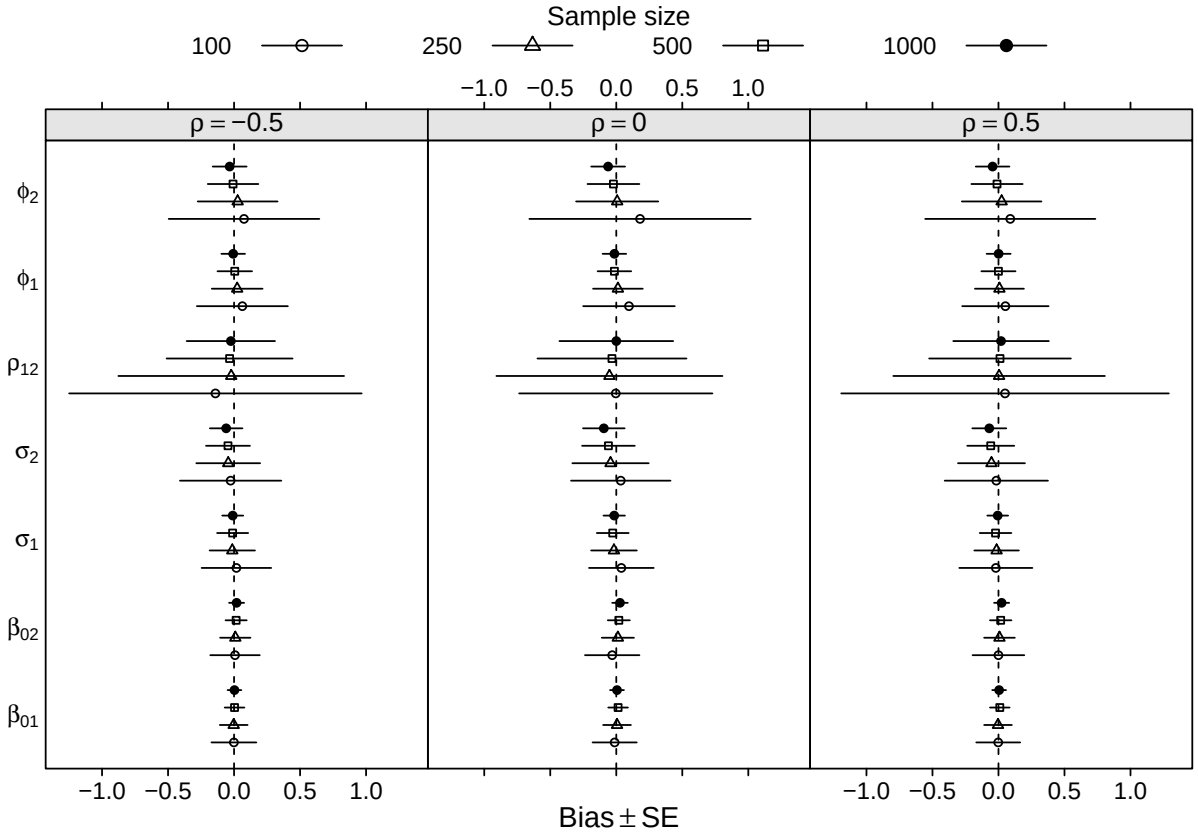
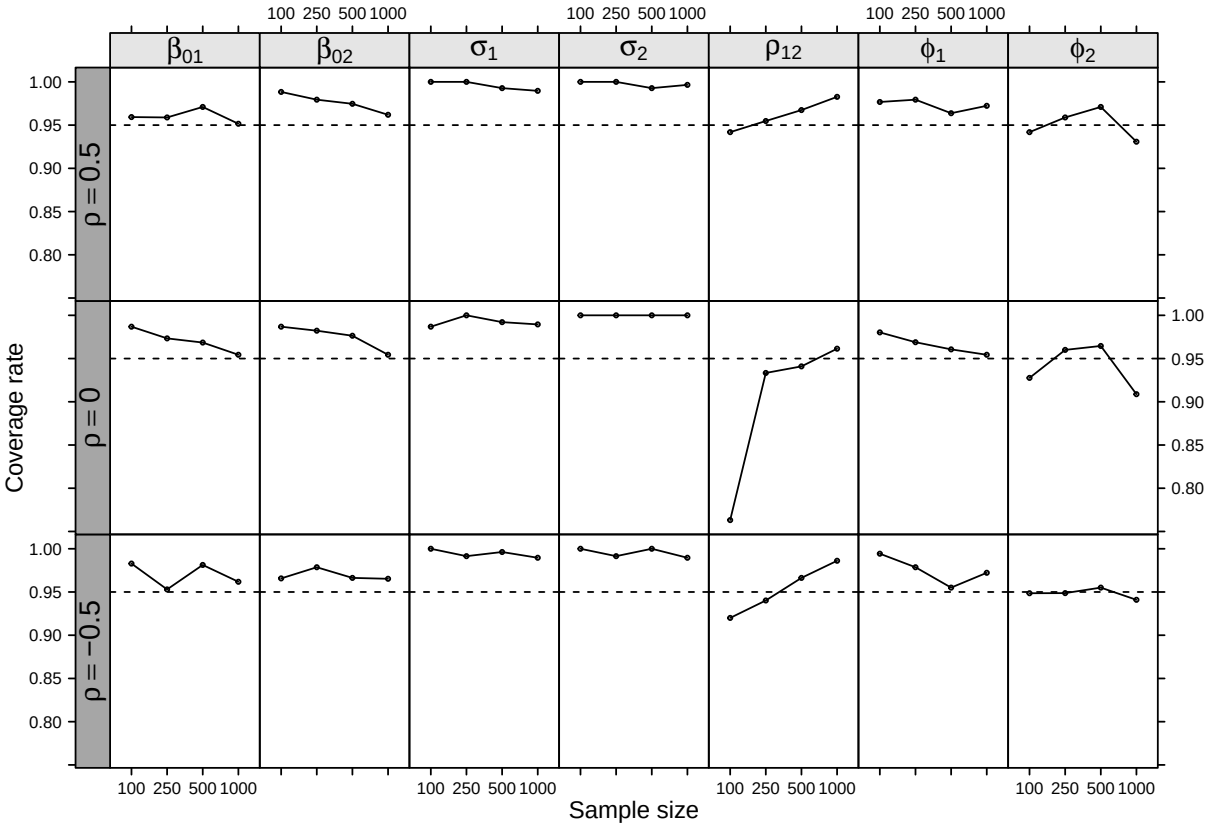


Figure 9 presents the coverage rate for each parameter by sample size and simulation scenario for bivariate NB regression model. Overall, all empirical coverage rates are close to the nominal level of 95%, varying between 90% and 99% approximately. In particular, the coverage rate of regression parameters β_{0r}, variance of random effects σ_r and correlation between random effects ρ are slightly greater than the nominal level. In opposite, there was a coverage rate close to 80% when sample size was equal to 100 and $\rho = 0$.

FIGURE 9 – COVERAGE RATE FOR EACH PARAMETER BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE NB REGRESSION MODEL



Estimation problems were more severe for the bivariate NB regression model compared to the Poisson. In addition to the rules used for Poisson to classify extreme values, it was necessary to remove those iterations when the dispersion parameter ϕ was greater than 5 (it was simulated at 1). Table 7 presents the problems summary.

TABLE 7 – SUMMARY OF FAILURES IN ESTIMATING THE BIVARIATE NB REGRESSION MODEL

Problem type	ρ	Sample Size	Fail	Simulations	% Fail
SE fail	-0.5	100	102	300	34.00
SE fail	-0.5	250	62	300	20.67
SE fail	-0.5	500	33	300	11.00
SE fail	-0.5	1000	12	300	4.00
SE fail	0.0	100	134	300	44.67
SE fail	0.0	250	72	300	24.00
SE fail	0.0	500	44	300	14.67
SE fail	0.0	1000	15	300	5.00
SE fail	0.5	100	105	300	35.00
SE fail	0.5	250	54	300	18.00
SE fail	0.5	500	24	300	8.00
SE fail	0.5	1000	11	300	3.67
Extreme values	-0.5	100	23	300	7.67
Extreme values	-0.5	250	4	300	1.33
Extreme values	0.0	100	14	300	4.67
Extreme values	0.0	250	3	300	1.00
Extreme values	0.0	500	2	300	0.67
Extreme values	0.5	100	23	300	7.67
Extreme values	0.5	250	3	300	1.00

Table 7 shows that from the 740 problems in estimation, 668 were due to SE failure problem and 72 due to extreme values. Considering the scenarios simulated, when the sample size was 100 and $\rho = 0$, it resulted in the largest number of problems at all (148 from 300), followed by $\rho = .5$ with 128 and $\rho = -.5$ with 125. We can also note that as the sample size increases, decreases the number of problems.

5.1.3 COM-Poisson

Figure 10 presents the average bias and confidence interval based on the mean SE by sample size and simulation scenario for bivariate COM-Poisson regression model. Figure 10 shows that for all simulation scenarios the SE tend to 0 as the sample size increases. This shows the consistency of the MLE estimator. However, almost every estimator is biased. While ν_2 is overestimated, ν_1 is underestimated (and the bias does not decrease with an increase of sample size). The correlation parameter ρ has a typical behaviour: when the data was simulated with $\rho = 0$ there was almost no bias, for $\rho = .5$ a negative bias and for $\rho = -.5$ a positive bias; the model forces the correlation parameter to become zero. The standard deviation of random effect σ_2 is overestimated, while σ_1 is slightly underestimated. Regarding the regression parameters, β_{02} is being slightly under estimated and β_{01} shows a negligible bias.

An interesting possible relationship among parameters is that when ν was overestimated (making the model more underdispersed: $\nu > 1$), σ was also overestimated (increasing the variance of the model). In contrast, when ν was underestimated (making the model even more overdispersed: $\nu < 1$), σ was also underestimated (decreasing the variance of the model). It seems that when ν makes the dispersion of the model greater, σ makes the variance lower, and vice-versa.

FIGURE 10 – AVERAGE BIAS AND CONFIDENCE INTERVAL BASED ON THE MEAN SE BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE COM-POISSON REGRESSION MODEL

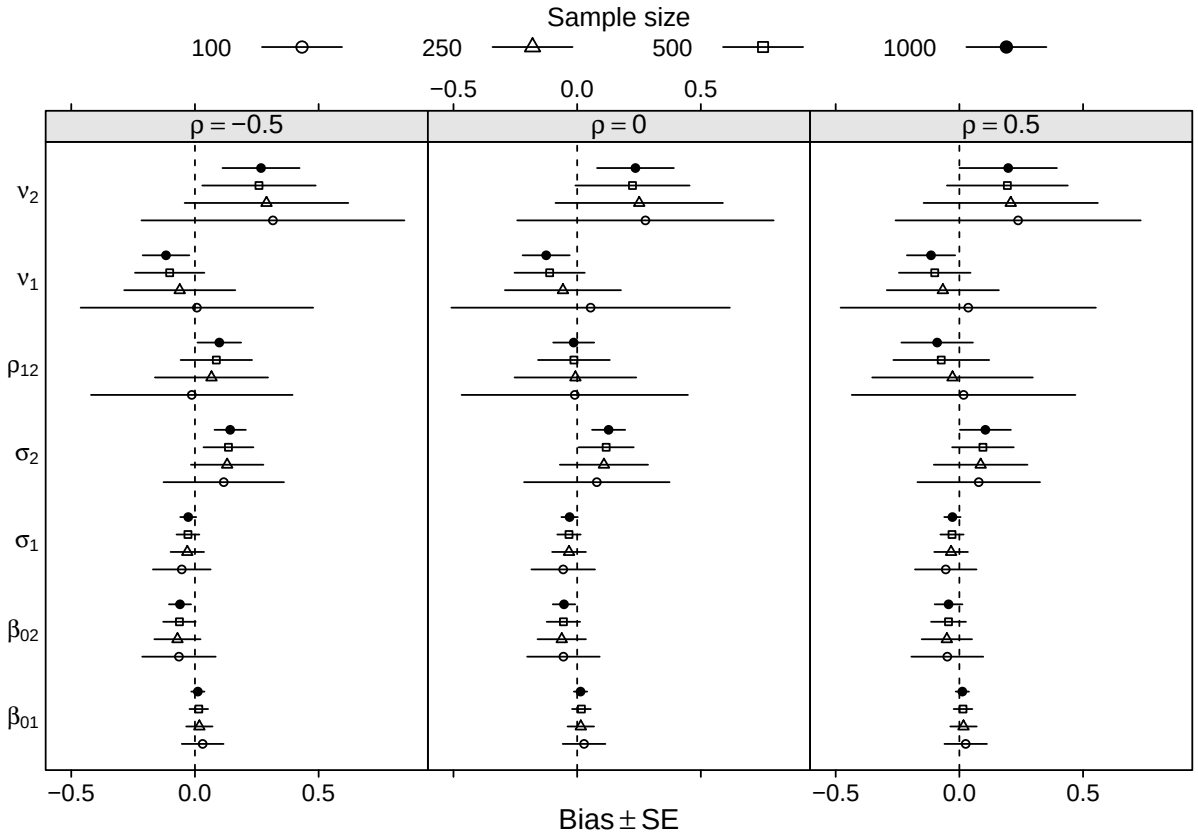
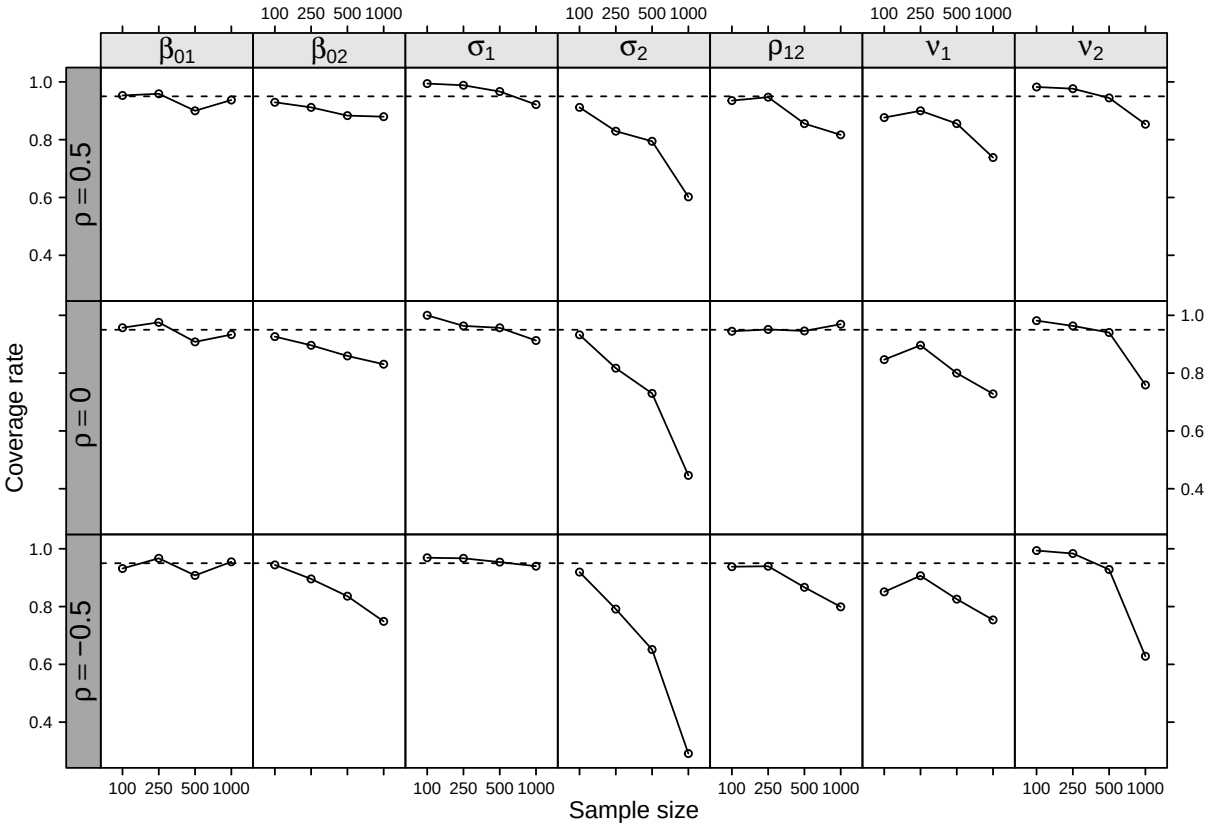


Figure 11 presents the coverage rate for each parameter by sample size and simulation scenario for bivariate COM-Poisson regression model. Overall, all empirical coverage rates are not close to the nominal level of 95%, which agrees to the results presented in Figure 10, where the bias did not decrease even for a higher sample size. In particular, the coverage rate of σ_2 had the worst results among all parameters (the coverage rate decreases as sample sizes increased), followed by ν_2. Not surprisingly, these parameters had the two largest bias in Figure 10. In contrast, results for β_{01}, σ_1 and ρ (when the data was generated with $\rho = 0$) had coverage rates close to 95% level.

FIGURE 11 – COVERAGE RATE FOR EACH PARAMETER BY SAMPLE SIZE AND SIMULATION SCENARIO FOR BIVARIATE NB REGRESSION MODEL



Estimation problems were more severe for the COM-Poisson regression model compared to the Poisson and less severe if compared to NB. In addition to the rules used for Poisson to classify extreme values, it was necessary to remove those iterations when the dispersion parameter ν was greater than 4 and the SE of ρ was greater than 2. Table 8 presents the problems summary.

TABLE 8 – Summary of failures in estimating the Bivariate COM-Poisson Regression model

Problem type	ρ	Sample Size	Fail	Simulations	% Fail
Fail SE	-0.5	100	30	200	15.0
Fail SE	-0.5	250	18	200	9.0
Fail SE	-0.5	500	5	200	2.5
Fail SE	-0.5	1000	1	200	0.5
Fail SE	0.0	100	28	200	14.0
Fail SE	0.0	250	35	200	17.5
Fail SE	0.0	500	14	200	7.0
Fail SE	0.0	1000	5	200	2.5
Fail SE	0.5	100	18	200	9.0
Fail SE	0.5	250	26	200	13.0
Fail SE	0.5	500	20	200	10.0
Fail SE	0.5	1000	9	200	4.5
Extreme values	-0.5	100	9	200	4.5
Extreme values	0.0	100	9	200	4.5
Extreme values	0.0	250	1	200	0.5
Extreme values	0.0	500	1	200	0.5
Extreme values	0.5	100	12	200	6.0
Extreme values	0.5	250	4	200	2.0

Table 8 shows that from the 245 problems in estimation, 209 were due to SE failure problem and 36 due to extreme values. Independently of the problem type, when sample size was equal to 100 and $\rho = -.5$ we had the largest number of problems in estimation (39), followed when $\rho = 0$ with 37 problems and sample size = 250 and $\rho = 0$ with 36 problems. We can also note that as the sample size increases, decreases the number of problems.

5.2 DATA ANALYSES

This section presents the data analyses of the three datasets presented in chapter 2. Initial values had to be chosen carefully for the models described in chapter 4. Firstly and for every dataset, it was fitted a MCGLM model (BONAT, 2016) in order to obtain initial parameter estimates for the regression and variance parameters (based on the variance of the residuals). On the other hand, the correlation parameter was set to 0. We did not use MCGLM to obtain initial estimates for the correlation parameter due to the difference of methodologies. These parameter estimates values were considered as initial values to the Poisson model.

Then, for every distribution the estimation was made in two steps. In the first, the data was fitted with a sample of size 350 for NHANES data, 300 for AHS data

and no sample was used for ANT data because the data had only 30 observations. An estimation process with sample size smaller than the ones used did not converge for each data respectively, so, we had to use different sample sizes for each data set. In the second step, the initial values from the sample model were used to fit the model with the whole dataset, and the first estimation was carried out with the `PORT` routine, followed by `BFGS` and `PORT`. In every step, initial values were used from the last optimization, and the final result reported was based on the maximum value of the logLik.

The initial values for the NB model were the final estimates from the Poisson model (with initial value to the dispersion parameter $\phi = 1$, i.e., small overdispersion). In the same way, the initial values for the COM-Poisson model were the final estimates from the NB model (with initial value to the dispersion parameter $\nu = 1$, i.e., equidispersion).

This flow is described in Figure 12. The reported results are only from the full data, as the samples were used to obtain initial estimates.

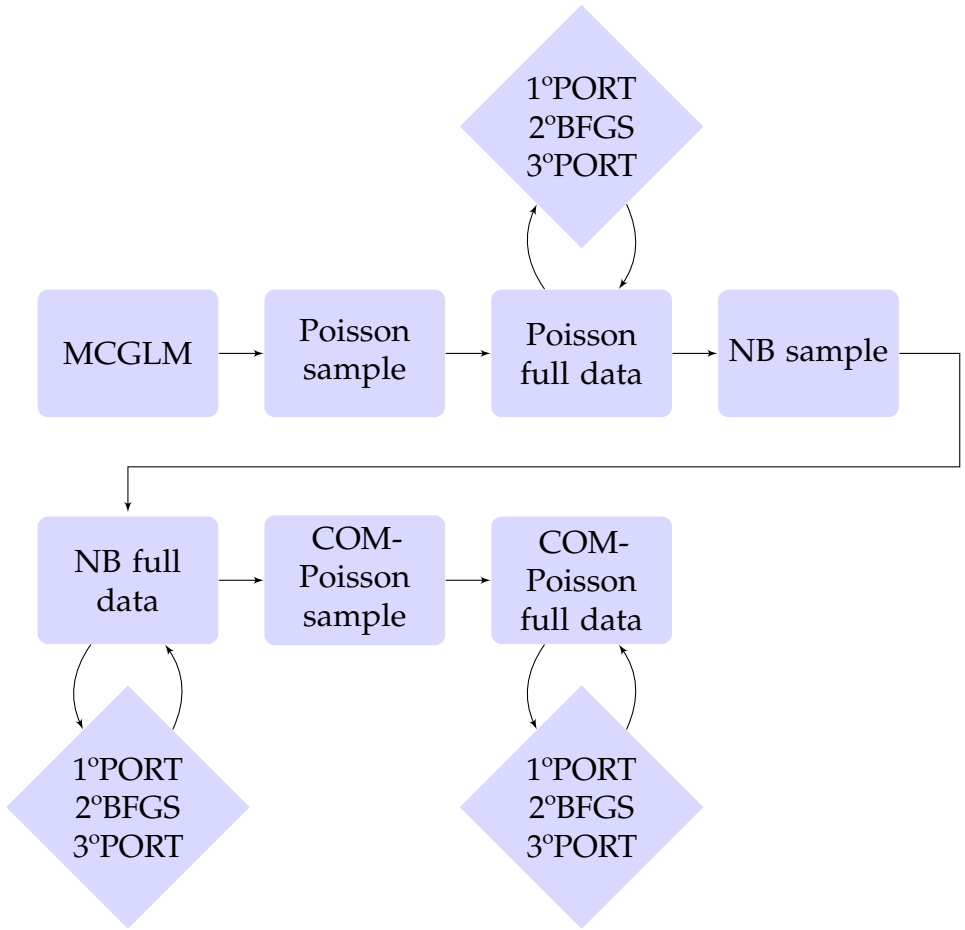


FIGURE 12 – MODEL ESTIMATION PROCESS FOR EACH DATASET

In addition to the models presented in chapter 4, a simpler version of them was also fitted. It was used two simplifications in the behalf of the variance of the random

effect and one for the dispersion parameter. As the Poisson distribution does not have a dispersion parameter, only the full specification of the model was used. In the first case was considered a model with common variance specification, where the variance of the random effect was equal to all response variables. In the second case, we used a fixed variance specification, where the variance was fixed to 1 and it was not estimated along the model. The final and third scenario was when the dispersion parameter was fixed to $\phi = 1$ for the NB (indicating small overdispersion for this distribution) and $\nu = 1.5$ for the COM-Poisson model (indicating small underdispersion for this distribution).

We then show the results of the models presented in Figure 12 and its variations with log-likelihood value (`logLik` - bigger the best), Akaike Information Criterion (AIC - lower the best), Bayesian Information Criterion (BIC - lower the best), number of parameters estimated (np) and whether the SE was calculated or not for the estimated parameters. While the logLik value is only concerned about the shape of the function, AIC penalizes it by the number of parameters used ($\text{AIC} = -2 * \text{logLik} + 2 * \text{np}$) and BIC penalizes even more for samples greater than 8 ($\text{BIC} = -2 * \text{logLik} + \log(\text{n}) * \text{np}$), where n is the sample size.

5.2.1 Results of NHANES data

Table 9 presents the model fit measures for NHANES data from different distributions and parametrization.

TABLE 9 – MODEL FIT MEASURES FOR NHANES DATA FROM DIFFERENT DISTRIBUTIONS AND PARAMETRIZATION

Model	np	AIC	BIC	logLik	SE
Poisson	21	7145.1	7253.4	-3551.6	✓
NB	24	7150.2	7273.9	-3551.1	✓
Fixed dispersion NB	21	8092.4	8200.6	-4025.2	✗
Common variance NB	22	7203.4	7316.8	-3579.7	✗
Fixed variance NB	21	7924.2	8032.5	-3941.1	✓
COM-Poisson	24	4615.9	4739.6	-2284.0	✓
Fixed dispersion COM-Poisson	21	6991.7	7099.9	-3474.8	✗
Common variance COM-Poisson	22	5608.2	5721.6	-2782.1	✓
Fixed variance COM-Poisson	21	6649.8	6758.0	-3303.9	✓

In overall, the model which best fitted to the data with respect to the three fit measures was the COM-Poisson with full specification. We can also note a large difference in absolute numbers against the competitors. The second best specification was with a common variance for a COM-Poisson model, suggesting that the dispersion parameter of COM-Poisson has a major importance in fitting this model. Moreover, we

can note a very close `logLik` between the NB and the Poisson: it happened due to a large value for the dispersion parameter of the NB, approaching for the Poisson model. Among the model specification variations, the fixed dispersion was the worst scenario for both NB and COM-Poisson models compared to fixed or common variance.

This results agree with underdispersion characteristic of the data presented in Table 1. It is important to note that only beforehand COM-Poisson is able to deal with underdispersion data due to the ν parameter. For Poisson and NB models, the random effects structure added one extra variance parameter for each response variable. For Poisson, the extra variance parameter gives only the ability to model overdispersion; while for NB, it allows to model even greater variability than the traditional NB model. Therefore, the analysis of this dataset confirmed the expected results by the specified model and distributions.

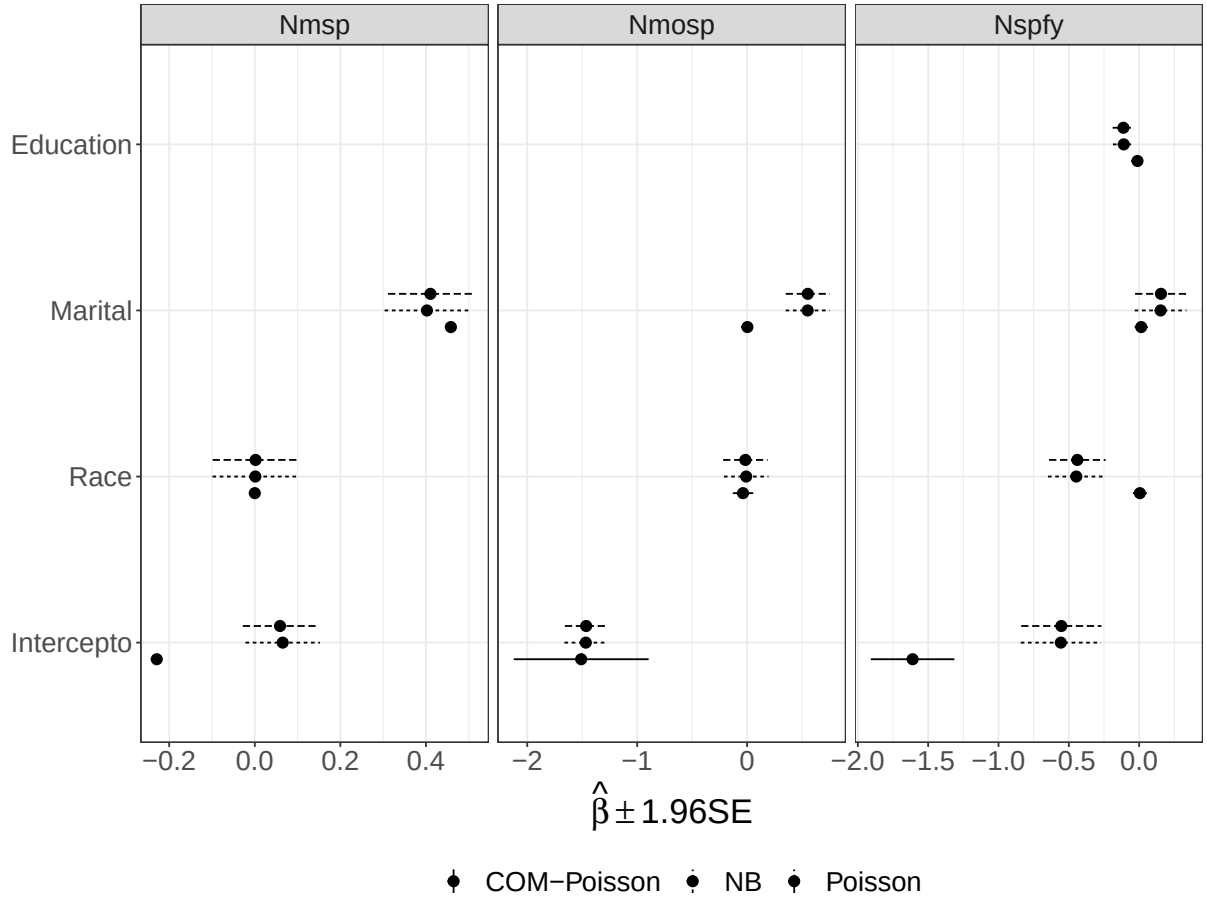
The check mark ✓ in last column specifies that the SE of the parameters were calculated (at least for one parameters of the model). The ✗ indicates that it was not possible to calculate the SE for any parameter of the model. Even in the COM-Poisson model, some estimates did not have the SEs calculates. Then, we modified the linear predictor excluding those estimates from each response variable that did not have the SEs calculated, and refitted the model. The same linear predictor was used to fit the best model from each distributions,i n this case, Poisson, and NB. The results are presented in Table 10.

TABLE 10 – MODEL FIT MEASURES FOR NHANES DATA FROM THE BEST PARAMETRIZATION FOR EACH DISTRIBUTION

Model	np	AIC	BIC	logLik	SE
Poisson	16	7147.8	7230.3	-3557.9	✓
NB	19	7153.1	7251.0	-3557.5	✓
COM-Poisson	19	4448.6	4546.6	-2205.3	✓

We can see that reducing the linear predictor improved the COM-Poisson model fit, while it worsened the Poisson and NB in terms of `logLik` and `AIC`, but improved the `BIC` because of a greater penalty due to the number of parameters. In order to better understand the behaviour of the parameter estimates, Figure 13 compares the regression estimates, in Table 11 compares the dispersion and Equation 5.1 presents a matrix where out of the diagonal are presented correlation estimates and the SE of each random effect, along with the standard deviation and SE in the diagonal.

FIGURE 13 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL



From Figure 13 we can see that for most of estimates, the confidence intervals were smaller for COM-Poisson compared to NB and Poisson models. As these distributions can not handle underdispersed data, they tend to overestimate the SE of the parameters. The only confidence interval that was greater for COM-Poisson compared to NB and Poisson was the intercept for Nmosp. Moreover, we see that the intercept point estimate was not close between COM-Poisson and the other two models in every scenario.

Regarding the relationship between the covariates and the response variables, the models considered agree in the direction of the relationship to each response variable. Race (1 = White, 0 = Others) covariate has almost no effect for Nmsp and Nmosp; for Nspfy has a negative effect for Poisson and NB models and null effect for COM-Poisson model. Marital status (1 = Married, 0 = Others) has a positive effect for all three response variables for NB and Poisson models; while for COM-Poisson, has a positive effect only for Nmsp, and a null effect in Nmosp and Nspfy. Finally, education level (range from 1 = Less Than 9^{th} Grade to 5 = College Graduate or above) has a negative effect on Nspfy in NB and Poisson models and no effect for COM-Poisson model. In those

models that can not handle underdispersion, more information was captured by the regression parameters than from the COM-Poisson model.

TABLE 11 – DISPERSION PARAMETER ESTIMATES AND SEs FOR EACH MODEL AND OUTCOME OF NHANES DATA

	NB (ϕ)		COM-Poisson ($\hat{\nu}$)	
Outcome	Estimate	SE	Estimate	SE
Nmsp	4958.17	21947.1	46.513	0.059
Nmosp	753.63	3878.4	20.964	5.829
Nspfy	1996.65	12993.9	19.675	1.286

From Table 11 we can see that ϕ is large enough to approximate the NB to a Poisson model followed by a even larger SE; it justifies the similar model measures presented in Table 9 and Table 10. For COM-Poisson, large $\hat{\nu}$ values indicate underdispersion and a small SE indicates that $\hat{\nu}$ is not zero.

$$\Sigma'_{\text{Poisson}} = \begin{bmatrix} .18(.03)^* & .94(.1)^* & .97(.04)^* \\ & .23(.08)^* & .92(.13)^* \\ & & .63(.07)^* \end{bmatrix} \quad \Sigma'_{\text{NB}} = \begin{bmatrix} .18(.03)^* & .97(.08)^* & .99(.02)^* \\ & .23(.08)^* & .97(.09)^* \\ & & .63(.07)^* \end{bmatrix}$$

$$\Sigma'_{\text{COM-Poisson}} = \begin{bmatrix} .48(<.01)^* & <-.01(<.01)^* & <.01(<.01) \\ & 1.21(.15)^* & <-.01(.01) \\ & & 1.25(.06)^* \end{bmatrix} \quad (5.1)$$

Equation 5.1 presents the correlation between random effects in the upper diagonal and the standard deviation in the diagonal. Stars represent statistical significance at 5% level. It was necessary to make a distinction between $\mathbf{\Sigma}$ and Σ' because the last one uses standard deviation in the diagonal, and the first uses variance.

Equation 5.1 shows that the correlation estimates for Poisson and NB model are close to one and are significant at 5% level. In contrast, the correlation estimates for COM-Poisson model are close to zero and only one is significant (between Nmsp and Nmosp). It seems that the underdispersion information not modelled by Poisson and NB models are somewhat present in the correlation estimates. Moreover, the standard deviation of the random effect was greater for COM-Poisson compared to the Poisson and NB models. It may occur because at the same time that the COM-Poisson has greater variability due to the random effect standard deviation, it balances out with smaller dispersion of $\hat{\nu}$ parameter.

5.2.2 Results of ANT data

Table 12 presents the model fit measures for ANT data from different distributions and parametrizations.

TABLE 12 – MODEL FIT MEASURES FOR ANT DATA FROM DIFFERENT DISTRIBUTIONS AND PARAMETRIZATIONS

Model	np	AIC	BIC	logLik	SE
Poisson	1107	5128.8	10835.8	-1457.4	✓
NB	1148	5281.5	11199.9	-1492.7	✗
Fixed dispersion NB	1107	5639.6	11346.7	-1712.8	✗
Common variance NB	1108	5292.1	11004.3	-1538.1	✗
Fixed variance NB	1107	5305.0	11012.0	-1545.5	✗
COM-Poisson	1148	5167.0	11085.4	-1435.5	✗
Fixed dispersion COM-Poisson	1107	5065.3	10772.3	-1425.6	✓
Common variance COM-Poisson	1108	5114.5	10826.7	-1449.3	✓
Fixed variance COM-Poisson	1107	5162.0	10869.0	-1474.0	✓

For this dataset we do not see large differences among the tested models as we saw in subsection 5.2.1. The best model considering the three measures for this dataset was the COM-Poisson with fixed dispersion, followed closely by COM-Poisson if we look only at `logLik` value. Among NB models, the one with bigger `logLik` and smaller `AIC` was the full model; moreover, the SE was not calculated to any model. The Poisson model was even better than all NB models, fixed variance COM-Poisson model, and AIC and BIC compared to COM-POISSON.

After a first filter of regression estimates who were not able to calculate the SEs based on the COM-Poisson model with fixed dispersion, we obtained the results presented in Table 13. In this case, we also decided to include the COM-Poisson model due to a close `logLik` value and because it was far the best model for NHANES data. A second filter attempted did not succeed as it decreased too much the `logLik` values and increased the `AIC` and `BIC`.

TABLE 13 – MODEL FIT MEASURES FOR ANT DATA FROM THE BEST PARAMETRIZATION FOR EACH DISTRIBUTION

Model	np	AIC	BIC	logLik	SE
Poisson	1057	5085.08	6566.15	-1485.54	✓
NB	1098	5166.98	6705.50	-1485.49	✓
COM-Poisson	1098	4965.87	6504.38	-1384.93	✓
Fixed dispersion COM-Poisson	1057	5016.70	6497.77	-1451.35	✓

From Table 13 we can see that the best final models were from COM-Poisson

distribution; Poisson and NB models had similar `logLik` results. The COM-Poisson model had a higher `logLik` and `BIC`, and smaller `AIC` values compared to the fixed dispersion model. A logLik ratio test (LRT) between these models resulted in $p < .00001$ ($\chi_{41} = 132.84$), which gives evidence in favour of the full specified model. Therefore, we will present the estimates of the full models for each distribution.

Figure 14 presents the regression parameter estimates and 95% confidence intervals by outcome and final model for the first 12 response variables. The same graphic for the remaining response variables are presented in Appendix A, once similar patterns were found for all response variables. Firstly, we can see that not all response variables share the same linear predictor (covariate feral mammal dung is not presented to the second, third and ninth response variable). Secondly, there was still some regression estimates that it was not possible to calculate the SE and are presented by a hollow circle; this is necessary to make the distinction where the SE was calculated and it was very small (filled circle, such as bare ground for COM-Poisson and response variable 1-8 and 10-12).

In overall, the confidence interval was smaller for COM-Poisson model compared to its counterparts and the point estimates were close among all models. The feral mammal dung covariate had the largest confidence intervals compared to the other covariates. There are still regression coefficients which are very close to zero, suggesting that a better variable selection method may be applied to improve the model fit.

FIGURE 14 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL

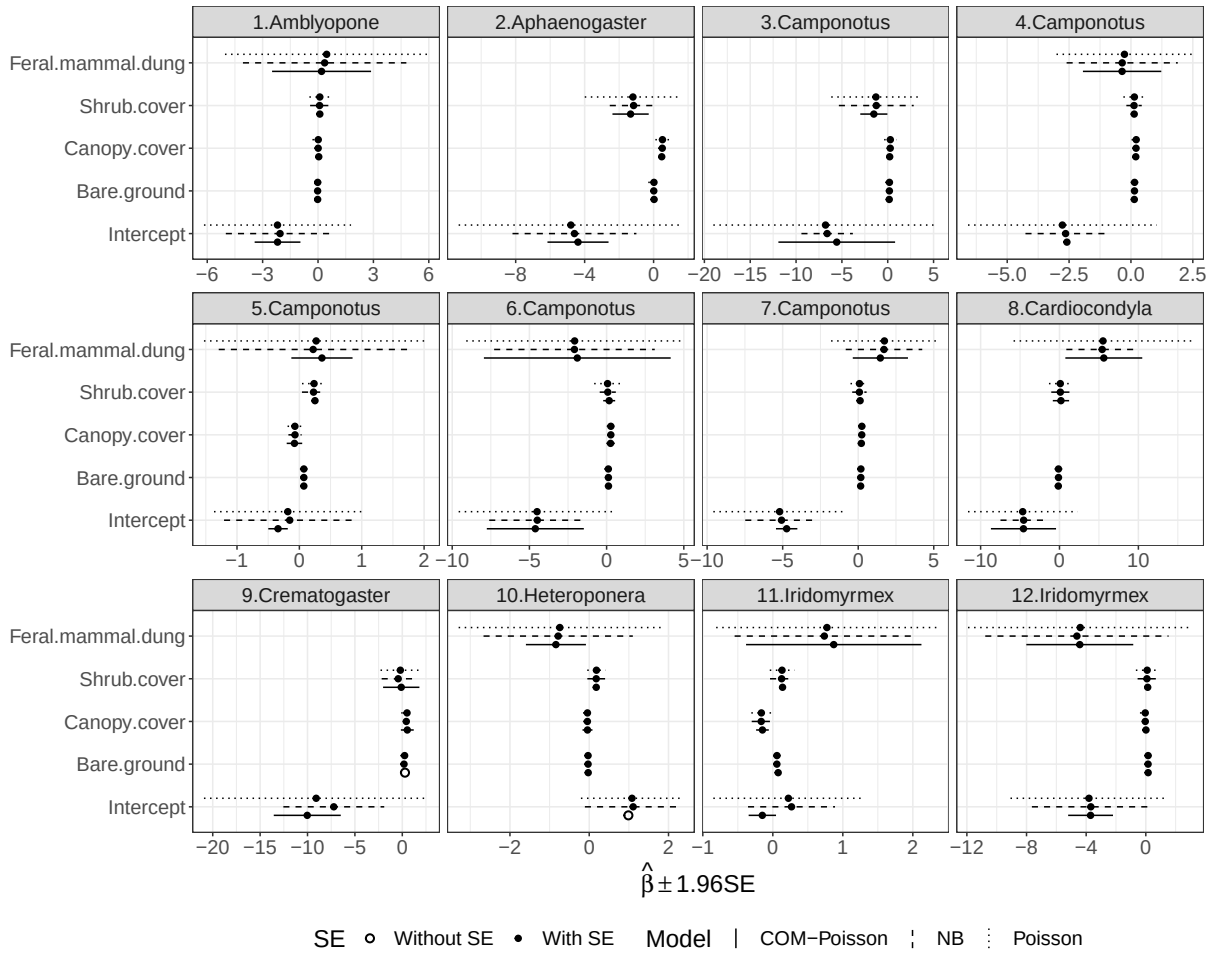


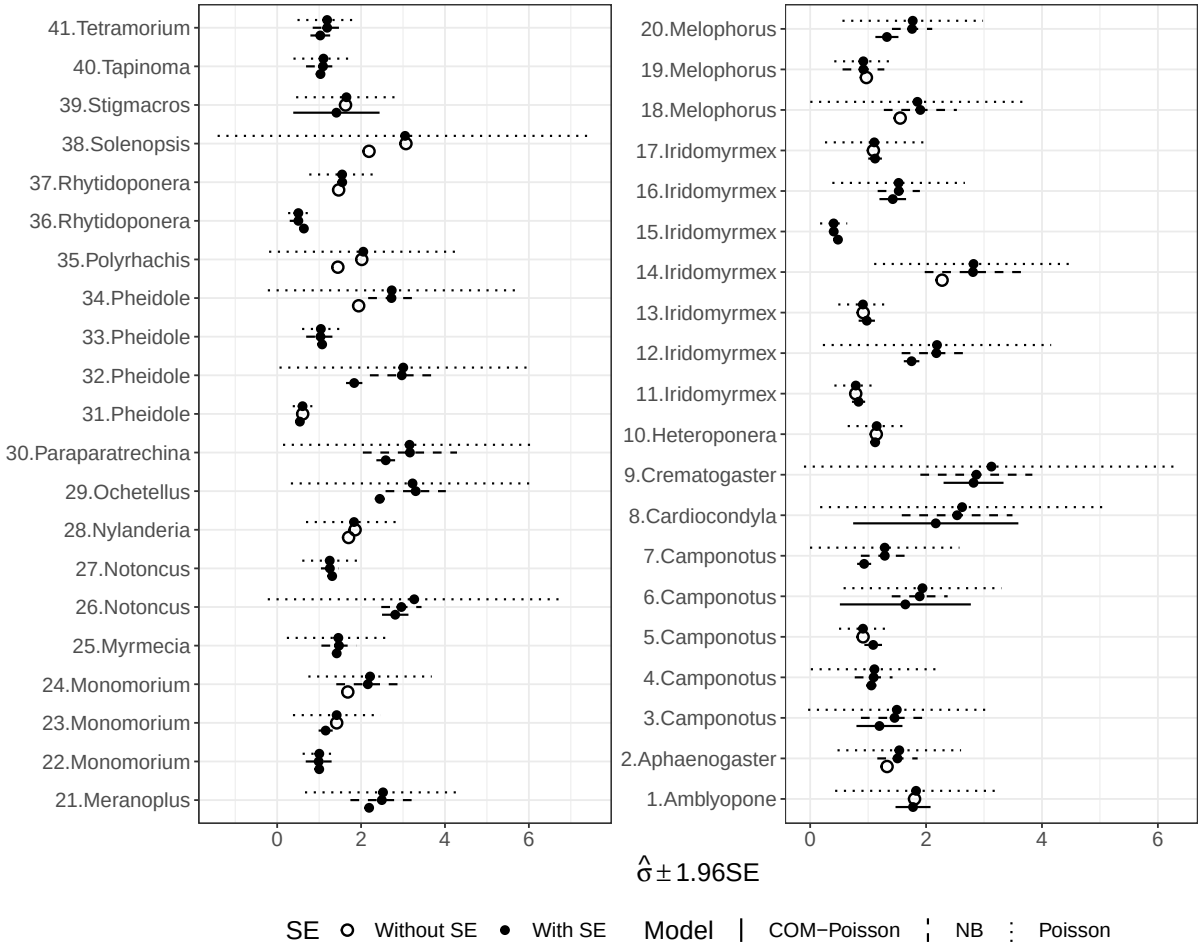
Table 14 presents the dispersion estimates for each model and outcome. Even tough this data can be considered as overdispersed according to the GDI = 11.543 presented in Table 3, we see that ϕ approaches the infinity (suggest an equidispersed model by NB distribution), and ν is greater than 1 for all response variables and indicates underdispersion. Even for 35.Polyrhachis and 38.Solenopsis response variables which had a DI smaller than 1, ν was equal to 50.264 (.004) and 5.722 (NaN) respectively. The SEs that were not calculate can be justified by the presence of response variables that do not have enough variability, such as 35 and 38, that can compromise the whole curvature of the `logLik` function, or even response variables such as 8 and 30 that have a large distance between counts, and a small frequency of them as was shown in Figure 2.

TABLE 14 – DISPERSION OF PARAMETER ESTIMATES AND SEs FOR EACH MODEL AND OUTCOME OF ANT DATA

	NB(ϕ)		COM-Poisson(ν)	
Outcome	Estimate	SE	Estimate	SE
1.Amblyopone	1.6e+06	2.7e+09	4.629	0.093
2.Aphaenogaster	3.0e+08	NaN	1.353	3.021
3.Camponotus	3.9e+04	4.0e+05	8.233	7.941
4.Camponotus	3.1e+06	NaN	16.780	NaN
5.Camponotus	2.3e+08	NaN	9.440	2.514
6.Camponotus	1.0e+07	8.9e+08	4.805	19.806
7.Camponotus	1.2e+05	1.5e+07	23.899	12.705
8.Cardiocondyla	1.1e+06	8.6e+08	2.552	NaN
9.Crematogaster	1.3e+06	5.6e+08	2.894	4.797
10.Heteroponera	1.2e+08	NaN	12.014	1.301
11.Iridomyrmex	2.3e+09	NaN	4.541	0.576
12.Iridomyrmex	1.1e+06	2.5e+09	6.157	2.468
13.Iridomyrmex	2.0e+09	NaN	5.206	2.876
14.Iridomyrmex	3.1e+07	2.0e+09	3.950	0.338
15.Iridomyrmex	4.6e+10	NaN	1.835	0.172
16.Iridomyrmex	5.0e+07	NaN	7.759	7.950
17.Iridomyrmex	4.8e+05	1.0e+08	21.509	6.488
18.Melophorus	4.1e+04	4.6e+06	16.608	5.139
19.Melophorus	7.2e+08	NaN	2.452	NaN
20.Melophorus	3.0e+05	5.3e+07	33.173	1.365
21.Meranoplus	4.6e+07	NaN	2.086	0.996
22.Monomorium	2.6e+09	NaN	3.415	NaN
23.Monomorium	4.1e+07	1.6e+09	18.908	4.680
24.Monomorium	4.3e+05	1.1e+08	16.110	3.537
25.Myrmecia	1.8e+06	7.3e+08	3.150	0.279
26.Notoncus	2.3e+06	7.5e+08	2.598	0.863
27.Notoncus	4.9e+06	NaN	22.130	1.065
28.Nylanderia	1.5e+08	NaN	5.204	NaN
29.Ochetellus	5.9e+05	3.0e+08	4.368	5.511
30.Paraparatrechina	6.5e+06	NaN	2.034	NaN
31.Pheidole	3.7e+09	NaN	32.746	NaN
32.Pheidole	1.4e+06	5.1e+08	6.434	NaN
33.Pheidole	8.3e+07	3.7e+08	18.084	NaN
34.Pheidole	1.1e+05	3.0e+07	4.779	NaN
35.Polyrhachis	5.4e+03	2.2e+05	50.264	0.004
36.Rhytidoponera	1.8e+09	NaN	12.928	0.547
37.Rhytidoponera	9.6e+07	5.0e+09	6.523	NaN
38.Solenopsis	3.5e+04	5.6e+06	5.722	NaN
39.Stigmacros	1.7e+05	2.9e+06	26.556	12.873
40.Tapinoma	6.5e+07	NaN	6.848	0.121
41.Tetramorium	6.4e+06	5.1e+09	18.882	5.593

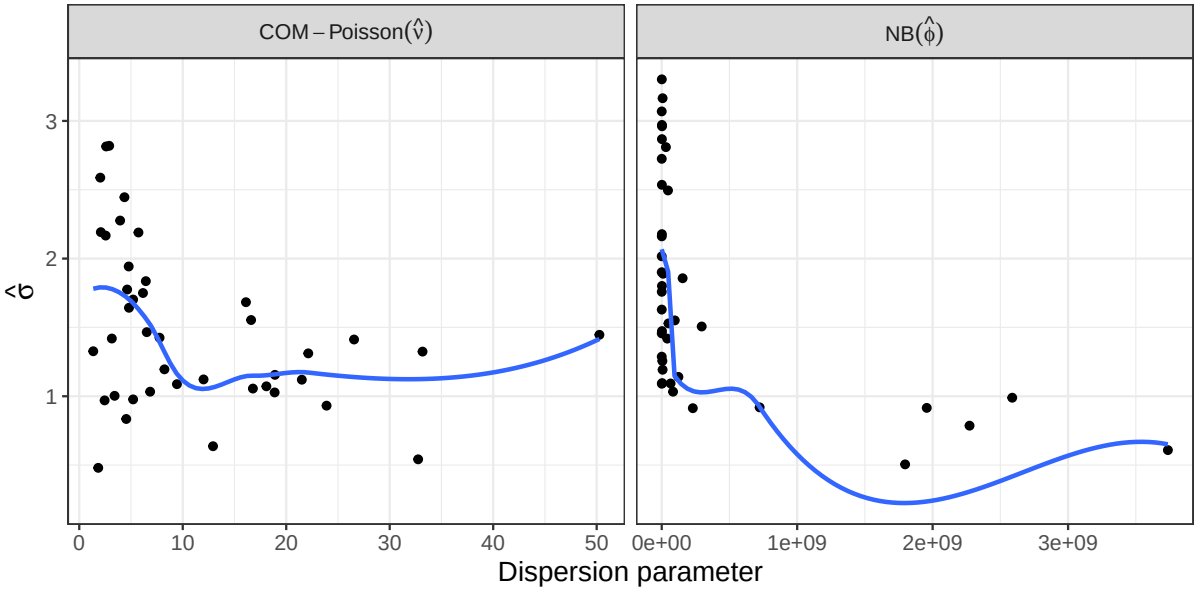
Figure 15 presents the standard deviation estimates of random effect and 95% confidence intervals by outcome and final model. We see that point estimates are not very close to zero, and except from Poisson, no variance parameter is equal to zero (according to Wald confidence intervals). It means that this parameter is adding extra variability to the model, making a counterweight to the results in Table 14, where those parameters were indicating equi and underdispersion for NB and COM-Poisson model. While the dispersion parameter decreases the variability, the variance parameter increases.

FIGURE 15 – STANDARD DEVIATION ESTIMATES OF RANDOM EFFECT AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL



This relationship is presented in Figure 16 and accounted by a -0.36 Pearson correlation coefficient for COM-Poisson model, and -.317 for NB. The $\hat{\phi}$ limit was shortened for NB removing $\hat{\phi} = 4.6e + 10$ in order to better see the pattern. The line in blue represents a local polynomial regression fitting just to aids the eye.

FIGURE 16 – SCATTER PLOT BETWEEN DISPERSION PARAMETER AND STANDARD DEVIATION OF RANDOM EFFECT FOR NB AND COM-POISSON MODELS



The correlation coefficient is presented in Figure 17, Figure 18 and Figure 19 for Poisson, NB and COM-Poisson models respectively. Among the 820 correlation coefficients calculated, COM-Poisson had 376 significant correlation coefficients, NB 330 and Poisson 119. It shows that correlation coefficient from COM-Poisson had smaller SE than their counterparts. The stars in the graphic represent the significant coefficients at 5% level. The correlation patterns presented in theses figures are somewhat different from the ones found in Figure 3. It shows the importance to calculate the correlation coefficient in a model that accounts the effect of the linear predictor. For example, the sample correlation was nearly zero between responses 1 and 2, 2 and 3; while for the COM-Poisson correlation between the random effect of these two response variables was negative and significant at 5% level, for NB only one was significant, and for Poisson neither was significant possibly because of a high SE, once the $\hat{\rho}$ was equal to -.7 between the random effects of Y1 and Y2, and -.54 between the random effects of Y2 and Y3.

FIGURE 17 – CORRELOGRAM OF ANT SPECIES OCCURENCE FROM POISSON MODEL. STARS REPRESENT CORRELATION SIGNIFICATIVE AT 5% LEVEL

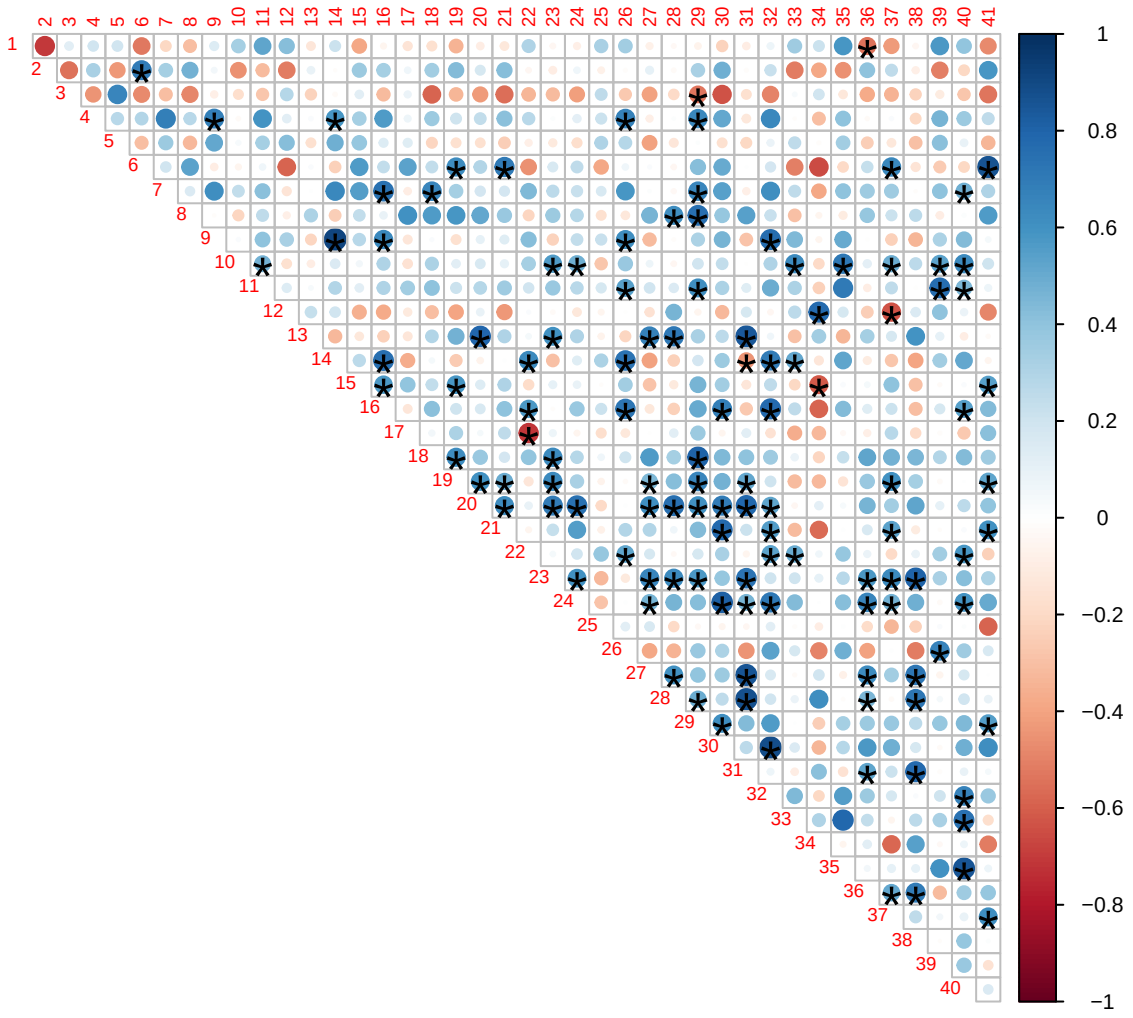


FIGURE 18 – CORRELOGRAM OF ANT SPECIES OCCURENCE FROM NB MODEL. STARS REPRESENT CORRELATION SIGNIFICATIVE AT 5% LEVEL

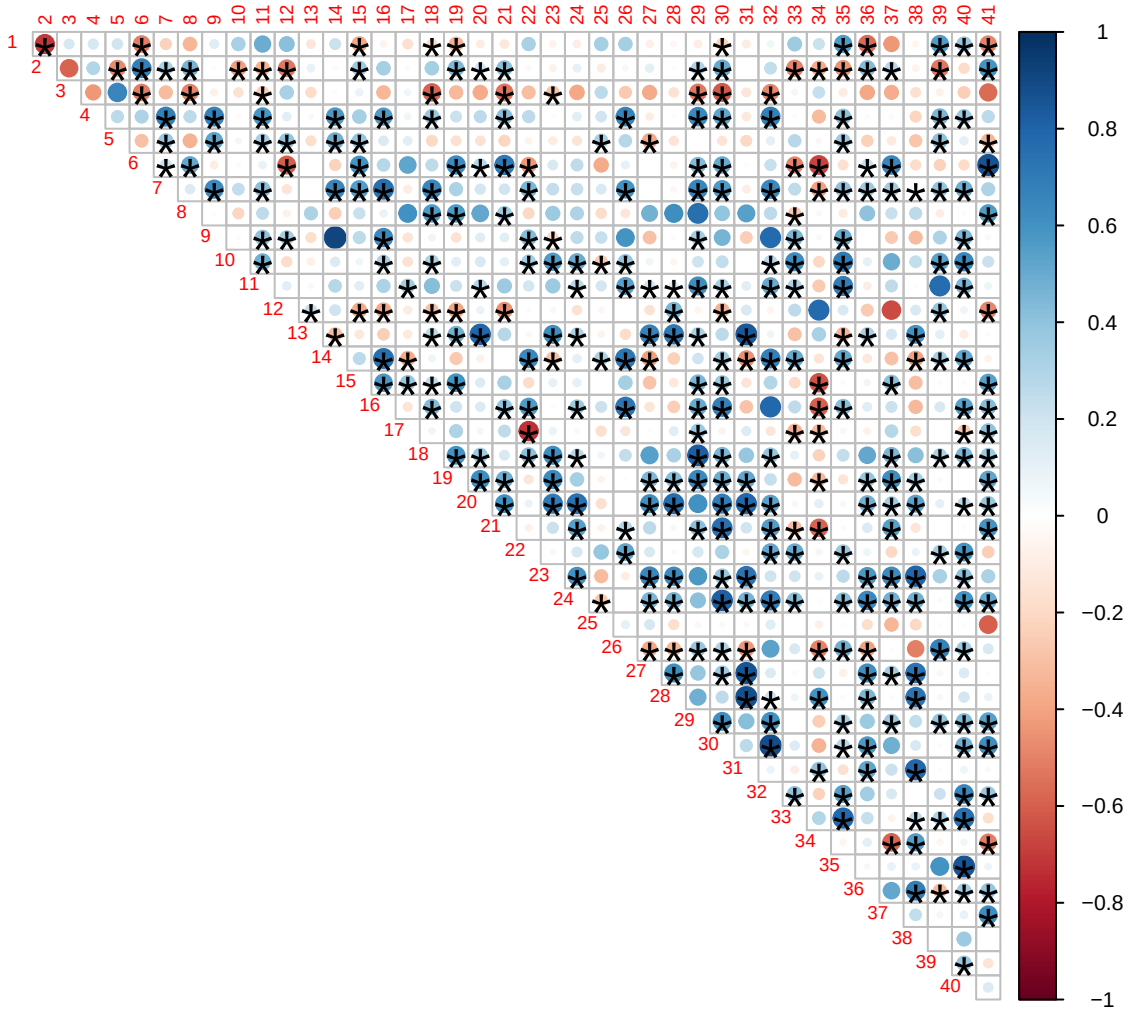
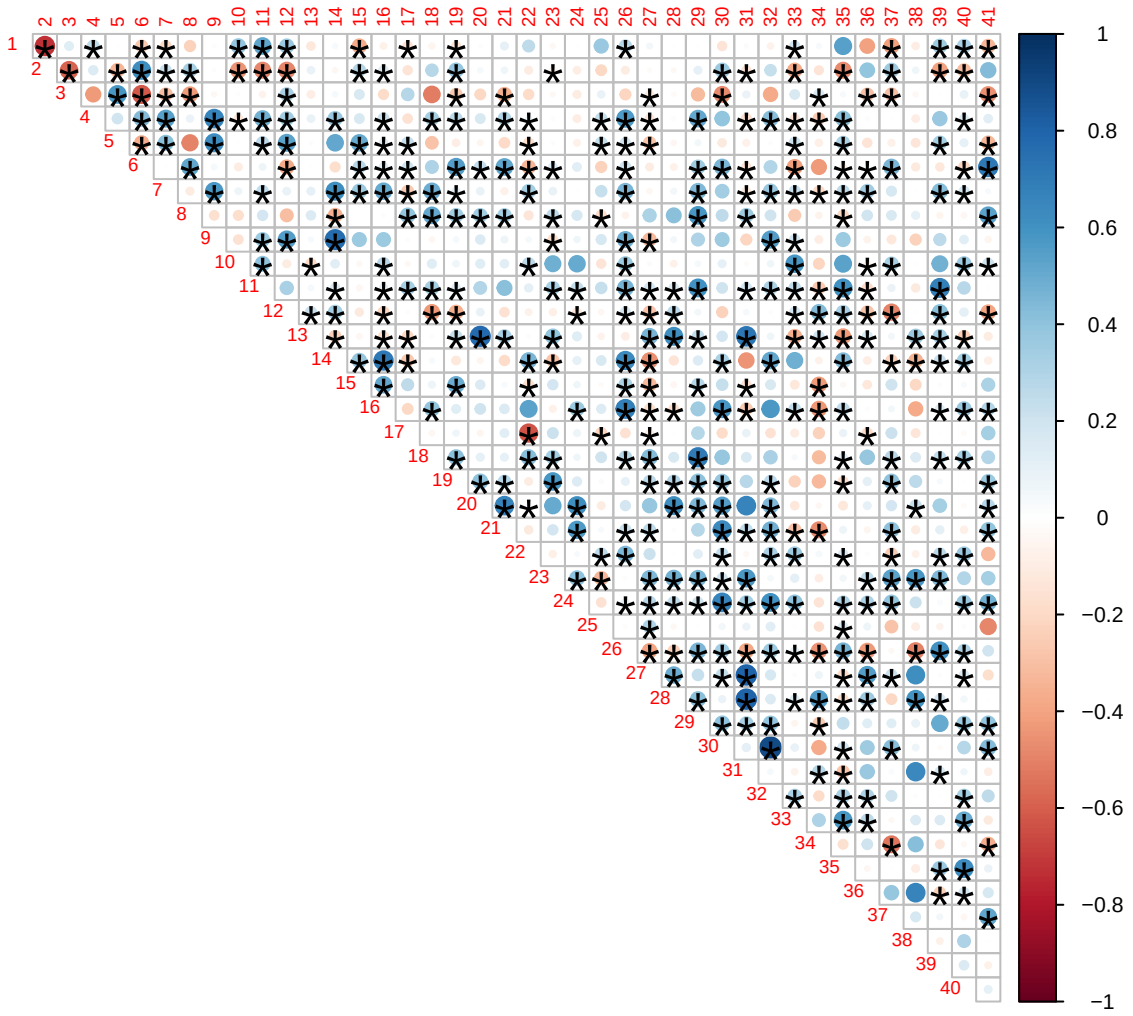


FIGURE 19 – CORRELOGRAM OF ANT SPECIES OCCURENCE FROM COM-POISSON MODEL. STARS REPRESENT CORRELATION SIGNIFICATIVE AT 5% LEVEL



5.2.3 Results of AHS data

Table 15 presents the model fit measures for AHS data from different distributions and parametrization.

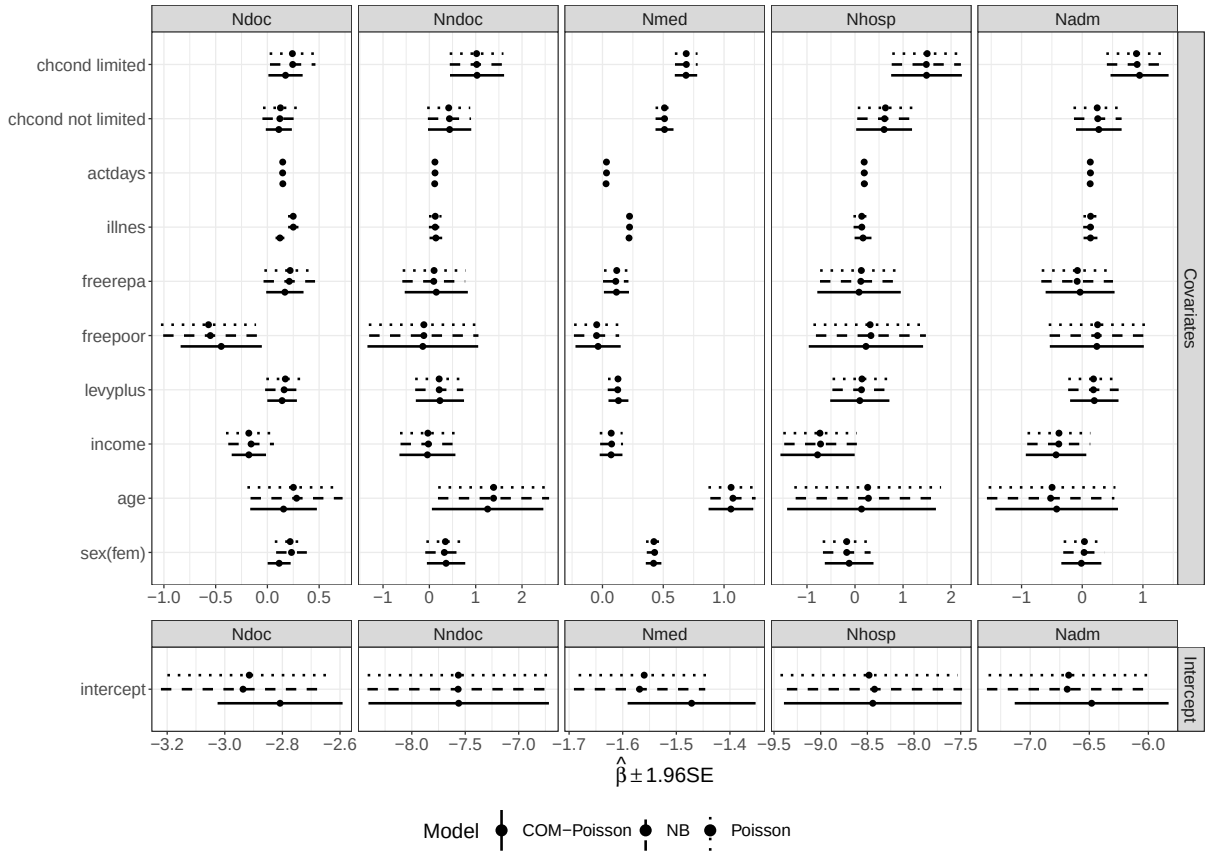
TABLE 15 – Model fit measures for AHS data from different distributions and parametrization

Model	np	AIC	BIC	logLik	SE
Poisson	70	33814	34272	-16837	✓
NB	75	33826	34318	-16838	✓
Fixed dispersion NB	70	35277	35736	-17569	✓
Common variance Var NB	71	37349	37815	-18604	✓
Fixed variance NB	70	37431	37890	-18645	✓
COM-Poisson	75	33106	33598	-16478	✓
Fixed dispersion COM-Poisson	70	33664	34123	-16762	✓

It was not necessary to make any variable selection based on those that did not compute the SE, because the SE was calculated for all coefficients, possibly due to an increase in sample size. The common and fixed variance did not converge for COM-Poisson. Therefore, COM-Poisson model was the best.

Figure 20 presents the regression parameter estimates and 95% confidence intervals by outcome and final model. First of all, it was necessary to separate one grid for covariates and another for intercepts due to smaller range of values for intercepts compared to covariates. After that, we can see that confidence interval amplitude is almost the same for all models, but slightly smaller for COM-Poisson models over the competitors in almost all estimates for Ndoc response variable; for the others response variables we did not see much a difference. The point estimates were very close between Poisson and NB models, with a difference for COM-Poisson model in some estimates, for example: intercept for Ndoc, Nmed and Nadm; freepoor, age, sex, illnes and chcond limited for Ndoc. This may be related to the bias found in Figure 10.

FIGURE 20 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL



We can recall that $\hat{\beta}$ is statistically significant if the interval does not contain the zero. Moreover, it changes (increase or decrease) the `log` of the response variable (and not the response variable directly) due to the `log` link function. To interpret the

parameter in the scale of the response, we need to exponentiate it. For example: it is expected that female individuals increase the mean of Ndoc in exp(.11) = 1.12 times compared to male individuals; in another words, it is expected that the mean in Ndoc is 12% higher in female compared to male.

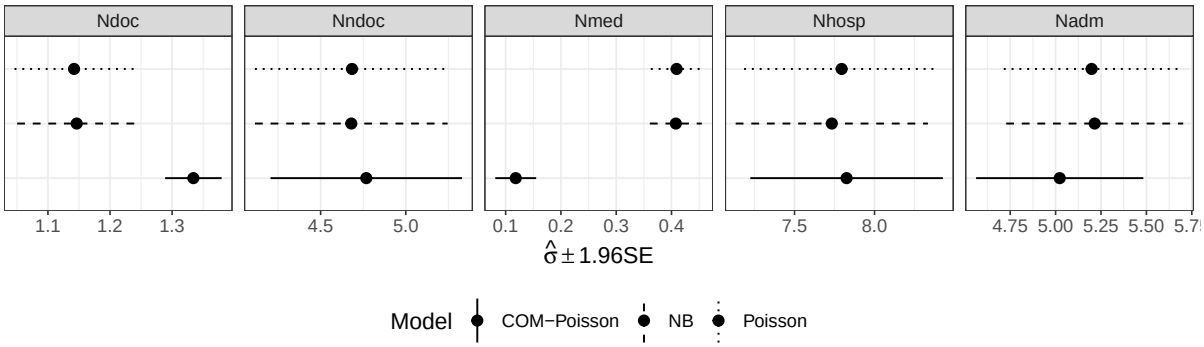
Table 16 presents the dispersion estimates for each model and outcome. Even tough this data can be considered as overdispersed according to the GDI = 17.94 presented in Table 5, we see that ϕ approaches the infinity (suggest an equidispersed model by NB distribution), and ν is greater than 1 for all response variables and indicates underdispersion. The expected result would ϕ and ν approaching to zero, which results in overdispersion of both distributions. However, as explored in subsection 5.2.2, this behaviour may occur due to the variance of the random effect. For Com-Poisson model, the only estimate which indicate overdispersion was for Nmed, which had a DI equals to 1.99 (small overdispersion).

TABLE 16 – DISPERSION OF PARAMETER ESTIMATES AND SEs FOR EACH MODEL AND OUTCOME OF AHS DATA

	NB(ϕ)		COM-Poisson(ν)	
Outcome	Estimate	SE	Estimate	SE
Ndoc	1.7e+04	1.8e+05	9.160	0.432
Nndoc	5.9e+03	6.0e+04	2.837	0.336
Nmed	3.6e+03	2.4e+04	0.674	0.032
Nhosp	8.5e+03	1.2e+05	5.110	0.615
Nadm	8.5e+03	8.1e+04	6.665	0.431

Figure 21 presents the standard deviation estimates of random effect and 95% confidence intervals by outcome and final model. In general terms, NB and Poisson models are very similar with respect to interval range and point estimates; COM-Poisson had a slightly smaller confidence interval for Ndoc and Nmed, and a larger and smaller punctual estimates respectively, which may be related to a possible ν-σ relationship. For Nmed, when the standard deviation is close, but not zero, we had $\nu = .674$, indicating that ν_3 captured the overdispersion of Nmed response variable. Lastly, Nhosp had the largest standard deviation among all random effects, which agrees to the largest sample variance found in Table 5.

FIGURE 21 – STANDARD DEVIATION ESTIMATES OF RANDOM EFFECT AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL



The correlation coefficients and its SEs in parenthesis are presented in Equation 5.2 for Poisson, NB and COM-Poisson models respectively. Among the 10 correlation coefficients calculated, COM-Poisson had 6 significant correlation coefficients, NB had 2 and Poisson 2. The stars in the matrix represent significant coefficients at 5% level. Even tough we do not have a main interest on interpreting each correlation coefficient, it is important to know which of them is significant, because it may be related to a smaller SE and absolute value that it is not close to zero.

We can see an almost perfect significant correlation between Nadm and Nhosp, and a strong correlation between Nmed and Ndoc (even stronger for COM-Poisson compared to Poisson and NB) for all three models. This pattern is also seen in Table 5. Besides them, for COM-Poisson model, Nmed is positive correlated with Nndoc, Nhosp and Nadm; also, a significant negative small correlation is seen for Ndoc and Nndoc. For Poisson and NB models, all nonsignificant coefficient are no greater than $\mid .07 \mid$.

$$
\begin{array}{r}
\\
\text{Poisson}_\rho = \\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\text{Ndoc} & \text{Nndoc} & \text{Nmed} & \text{Nhosp} & \text{Nadm}
\end{array} \\
\left[\begin{array}{ccccc}
1 & -0.02(0.04) & 0.59(0.07)^* & 0.04(0.03) & 0.05(0.03) \\
 & 1 & 0.07(0.05) & -0.04(0.03) & -0.06(0.03) \\
 & & 1 & 0.04(0.04) & 0.05(0.04) \\
 & & & 1 & 1(<.01)^* \\
 & & & & 1
\end{array}\right]
\end{array}
\begin{array}{l}
\\
\text{Ndoc} \\
\text{Nndoc} \\
\text{Nmed} \\
\text{Nhosp} \\
\text{Nadm}
\end{array}
$$

$$
\begin{array}{r}
\\
\text{NB}_\rho = \\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\text{Ndoc} & \text{Nndoc} & \text{Nmed} & \text{Nhosp} & \text{Nadm}
\end{array} \\
\left[\begin{array}{ccccc}
1 & -0.02(0.04) & 0.58(0.07)^* & 0.04(0.03) & 0.05(0.03) \\
 & 1 & 0.07(0.05) & -0.05(0.03) & -0.06(0.03) \\
 & & 1 & 0.05(0.04) & 0.06(0.04) \\
 & & & 1 & 1(<.01)^* \\
 & & & & 1
\end{array}\right]
\end{array}
\begin{array}{l}
\\
\text{Ndoc} \\
\text{Nndoc} \\
\text{Nmed} \\
\text{Nhosp} \\
\text{Nadm}
\end{array}
\tag{5.2}
$$

$$
\begin{array}{r}
\\
\text{COM-Poisson}_\rho = \\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\text{Ndoc} & \text{Nndoc} & \text{Nmed} & \text{Nhosp} & \text{Nadm}
\end{array} \\
\left[\begin{array}{ccccc}
1 & -0.07(0.03)^* & 0.78(0.09)^* & -0.02(0.02) & -0.02(0.02) \\
 & 1 & 0.34(0.14)^* & -0.04(0.03) & -0.05(0.03) \\
 & & 1 & 0.25(0.09)^* & 0.25(0.09)^* \\
 & & & 1 & 0.99(<.01)^* \\
 & & & & 1
\end{array}\right]
\end{array}
\begin{array}{l}
\\
\text{Ndoc} \\
\text{Nndoc} \\
\text{Nmed} \\
\text{Nhosp} \\
\text{Nadm}
\end{array}
$$

6 FINAL CONSIDERATIONS

The main focus of this master thesis was to propose MGLMM. This new class for statistical modelling can model more than one response variable in the same model accommodating a random effect that follows a multivariate normal distribution that can account the correlation between the random effects and the variance of them. This extra feature is not available in a GLMM model, where it is necessary to model one response variable at a time. In particular, in this thesis we addressed only count data problems, considering Poisson, NB and COM-Poisson distributions. This model was implemented using the `TMB` package in R. It is estimated via ML method using numerical integration via LA with inner and outer optimization based on Newton's method and general purpose algorithm respectively, such as BFGS and PORT routines, which derivatives are provided through AD.

In order to evaluate the properties of the ML estimators we conducted a simulation study for each distribution, considering three different values for the correlation parameter and four different sample sizes. They were all evaluated by means of average bias and confidence interval based on the mean SE and coverage rate with nominal level of 95%. For Poisson distribution we achieved unbiased and consistent estimators with intervals for bias ranging at most between (-.2;.2), except for sample size equals to 100 and ρ parameter. The coverage rate was close to 95% in all scenarios studied, varying between 90% and 98%. For NB it was seen a greater variability than compared to the Poisson distribution, especially because of the dispersion parameter ϕ in NB, necessary to model overdispersion. For NB distribution we also achieved unbiased and consistent estimators with bias intervals ranging in most of cases between (-.5;.5). In particular, a greater confidence interval width was seen for the correlation parameter and ϕ_2 parameter; and small bias for σ_2 and ν_2 parameters. The coverage rate in most cases was equal or greater than 95% confidence level; with a coverage rate below 80% for the correlation parameter $\rho = 0$ when sample size was equal to 100.

For the COM-Poisson model, while the consistency was achieved, the unbiasedness property was not verified for most parameters. The regression parameter was underestimated for β_{02} and showed no bias for β_{01}. The correlation parameter ρ was always estimated towards zero: when $\rho = -.5$ it was overestimated, $\rho = +.5$ it was underestimated, and for $\rho = 0$ it showed no bias. The standard deviation of random effect σ_2 was overestimated while σ_1 was underestimated. This behaviour may be correlated to the dispersion parameter ν: when more variance was captured from σ_2 less dispersion was captured from ν_2; on the other hand, when less variance was captured from σ_1, more dispersion was captured from ν_1. The dispersion parameter ν

deserves a bit more of attention to interpret it.

The dispersion parameter ν_2 was overestimated where ν_1 was underestimated in all scenarios. This bias may due to the reparametrization used by Huang (2017), where in 1 out of 3 examples the dispersion parameter was overestimated by 1.6% (1.754 compared to 1.727 original value); unfortunately, no simulation study was provided. On the other hand, there is a recently proposed mean parametrization of the COM-Poisson distribution by Ribeiro et al. (2020) that shows a simulation study where there is a small bias for the dispersion parameter in the overdispersion case for a sample size of 1000. A performance comparison between these two reparametrization shows that Ribeiro et al. (2020) is more efficient than Huang (2017), because the new parametrization only requires simple algebra, without adding any extra computational complexity to the original parametrization. Huang (2017) requires an extra solution of a linear system that contains an infinite sum on it.

As most of parameters were biased, the coverage rate was not close to 95% in all scenarios. For σ_2, β_{02} and ν_r the coverage rate was between 70% and 90%; for ρ the coverage rate was close to 80% in 2 out of 3 scenarios ($\rho = \{-0.5, 0.5\}$). The other parameters, β_{01}, σ_1 and ρ when $\rho = 0$ had coverage rate close to 95%.

Along the simulation, some estimates did not have the SEs calculated or produced extreme values, and therefore, they were discarded. In the Poisson model it occurred in .72% cases, for NB was in 20.56% and COM-Poisson in 10.21%. These percentages were calculated by number of problems divided by total number of simulations, that will change according to the number of replications for each distribution. For example, for COM-Poisson the total number of simulation is 200 times 12 (the number of scenarios).

After that, three datasets were analysed by each model and variations of them: fixing dispersion, fixed variance and common variance for all random effects. The first data set analysed was from the NHANES survey, which comprises of three response variables, being one equidispersed and two underdispersed. The COM-Poisson was the best model compared to their counterparts according to `logLik`, `AIC` and `BIC`, especially because of its ability to model underdispersion. The SE was smaller compared to the Poisson and NB models. The dispersion parameter ν indicated underdispersion, which it is expected due to the nature of the data and the correlation parameter ρ was almost zero.

The second dataset was the ANT which contains 41 response variables that counts for the number of ANT species in 30 sites in Australia. The multivariate response can be considered as overdispersed by the GDI. The COM-Poisson model was also the best model regarding `AIC` and `logLik`; the model with best `BIC` was the COM-Poisson with fixed dispersion. The SE was still smaller for COM-Poisson model compared to

the Poisson and NB models in almost all comparisons. The ν parameter was greater than 1 for all response variables indicating underdispersion. This may be due to a greater variance of the random effects, making a balance between these two parameters corresponding to a negative empirical correlation between them.

The third dataset is from the AHS survey with five overdispersed random variables and 5190 participants. The COM-Poisson model produced the best fit according to the fit measures used. The SE of β and ρ was similar among the COM-Poisson, NB and Poisson models. For σ, 2 out 5 five SE was smaller for COM-Poisson. The ν parameter was underdispersed in 4 out of 5 response variables and overdispersed in 1 response variable, followed by a small σ. The COM-Poisson model produced more significant correlation values than their counterparts.

Therefore, we suggest to use the MGLMM model framework for count data. In particular, best results were obtained with the COM-Poisson model in three real datasets tested. The main advantage of it is the possibility to model all response variable at the same time and measure the correlation between the random effects of them.

6.1 FUTURE WORK

There are some possibilities to improve the proposed model:

- To better study the relationship between σ and the dispersion parameter in order to propose a coefficient that indicates whether a response variable is under-equi or overdispersed.
- To propose a correlation coefficient between the response variables based on the correlation coefficient between the random effects, which may rely on the linear predictor and the dispersion of the conditional distribution.
- To use a simple to compute distribution that models underdispersion.
- Increase the simulation study considering underdispersed scenarios and more regression parameters.
- To implement a variable selection technique.

6.2 ESTIMATION PROBLEMS

The main problem for estimating these models were in maximizing the `logLik` of the slow to compute COM-Poisson pmf. Due to it, a lot of care was necessary to take into account in the estimation process, such as RAM and number of threads used. Along this process it was always necessary to make a balance between the RAM memory

consuming and number of threads used. At the same that increasing the number of threads used in the estimation decreases the computational time, it also increases the RAM usage. Therefore, it was necessary to find the correct number of threads used depending on the dataset and machine configuration.

Regarding the computational time to estimate those models, we bring some results for the AHS dataset. For Poisson and NB models, it was possible to estimate them on a personal machine: a Ubuntu 20.04 with intel(R) core (TM) processor i7-8750H CPU @2.20GHz with 24Gb of RAM. For Poisson it took 1 hour and 45 minutes to run with 12 threads. When we request 12 threads for a PC to estimate the model, it means that it can use up to 12 threads to complete this task, but it may have used less threads depending on the complexity of the data. For NB model it took 1hour and 20 minutes requesting only one thread. We are not considering neither the time used to fit the sample of the data in order to obtain initial estimates for the parameters and the time used to obtain the SEs. Even though NB had smaller computational time than Poisson, it usually does not happen. This may have occurred because other models could be running at the same time on the machine.

We tried to fit the COM-Poisson model in this same machine requesting four threads, but it kept around 2 weeks running and did not converge. In contrast, leaving it for a Debian GNU/Linux 8 (jessie) server 92GB RAM with a AMD Opteron 6136 processor using two threads it took 5 days to compile. When we tried to use more threads the kernel killed the session due to fault of memory. For all these models, the first run of PORT had the best performance with respect to time and in maximizing the log-likelihood function over BFGS.

BIBLIOGRAPHY

ALDRICH, J. et al. Ra fisher and the making of maximum likelihood 1912-1922. *Statistical science*, Institute of Mathematical Statistics, v. 12, n. 3, p. 162–176, 1997. Cited on page 17.

BATES, D. et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1–48, 2015. Cited on page 17.

BAYDIN, A. G. et al. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, JMLR. org, v. 18, n. 1, p. 5595–5637, 2017. Cited on page 40.

Bell BM. *CppAD: a package for C++ algorithmic differentiation*. [S.l.], 2005. Disponível em: ⟨http://www.coin-or.org/CppAD⟩. Cited on page 44.

BLACKFORD, L. S. et al. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, v. 28, n. 2, p. 135–151, 2002. Cited on page 44.

BONAT, W. H. *mcglm: Multivariate Covariance Generalized Linear Models*. [S.l.], 2016. R package version 0.3.0. Disponível em: ⟨https://CRAN.R-project.org/package=mcglm⟩. Cited 2 times on pages 16 and 59.

BONAT, W. H. Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, v. 84, n. 4, p. 1–30, 2018. Cited 2 times on pages 26 and 27.

BONAT, W. H. et al. Extended poisson–tweedie: Properties and regression models for count data. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, v. 18, n. 1, p. 24–49, 2018. Cited on page 15.

BONAT, W. H.; JR, P. J. R. Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, Wiley Online Library, v. 27, n. 2, p. 83–89, 2016. Cited on page 37.

BONAT, W. H.; JøRGENSEN, B. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, v. 65, n. 5, p. 649–675, 2016. ISSN 1467-9876. Disponível em: ⟨http://dx.doi.org/10.1111/rssc.12145⟩. Cited on page 15.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993. Cited on page 34.

BROOKS, M. E. et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, v. 9, n. 2, p. 378–400, 2017. Disponível em: ⟨https://journal.r-project.org/archive/2017/RJ-2017-066/index.html⟩. Cited on page 17.

BüRKNER, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, v. 10, n. 1, p. 395–411, 2018. Cited on page 16.

CAMPBELL, J. T. The poisson correlation function. *Proceedings of the Edinburgh Mathematical Society*, Cambridge University Press, v. 4, n. 1, p. 18–26, 1934. Cited on page 16.

CASELLA, G.; BERGER, R. L. *Statistical inference*. [S.l.]: Duxbury Pacific Grove, CA, 2002. v. 2. Cited on page 36.

COLIN, A. C.; TRIVEDI, P. K. *Regression Analysis of Count Data, Econometric Society Monograph No. 30*. [S.l.]: Cambridge: Cambridge University Press, 1998. Cited on page 26.

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, v. 12, n. 2, p. 132–136, 1962. Cited on page 31.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. *Mathematics for machine learning*. [S.l.]: Cambridge University Press, 2020. Cited on page 41.

DEMPSTER, A. P. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 205–232, 1968. Cited on page 16.

EL-SHAARAWI, A. H.; ZHU, R.; JOE, H. Modelling species abundance using the poisson–tweedie family. *Environmetrics*, Wiley Online Library, v. 22, n. 2, p. 152–164, 2011. Cited on page 15.

FAMOYE, F. A Multivariate Generalized Poisson Regression Model. *Comm. Statist. Theory Methods*, Taylor & Francis, v. 44, n. 3, p. 497–511, Feb 2015. ISSN 0361-0926. Cited 2 times on pages 16 and 19.

FISHER, R. A. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics*, Wiley Online Library, v. 6, n. 1, p. 13–25, 1934. Cited 2 times on pages 20 and 23.

GALTON, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, JSTOR, v. 15, p. 246–263, 1886. Cited on page 14.

GAY, D. M. Usage summary for selected optimization routines. *Computing science technical report*, v. 153, p. 1–21, 1990. Cited on page 17.

GIBB, H. et al. Does morphology predict trophic position and habitat use of ant species and assemblages? *Oecologia*, Springer, v. 177, n. 2, p. 519–531, 2015. Cited on page 21.

GRUNWALD, G. K. et al. A statistical model for under-or overdispersed clustered and longitudinal count data. *Biometrical Journal*, Wiley Online Library, v. 53, n. 4, p. 578–594, 2011. Cited on page 15.

GUENNEBAUD, G.; JACOB, B. et al. *Eigen v3*. 2010. Http://eigen.tuxfamily.org. Cited on page 44.

HADFIELD, J. D. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, v. 33, n. 2, p. 1–22, 2010. Disponível em: ⟨http://www.jstatsoft.org/v33/i02/⟩. Cited on page 16.

HARDIN, J.; HILBE, J. *Generalized Linear Models and Extensions*. Stata Press, 2018. ISBN 9781597182256. Disponível em: ⟨https://books.google.com.br/books?id=AhKQtgEACAAJ⟩. Cited on page 31.

HUANG, A. Mean-parametrized conway–maxwell–poisson regression models for dispersed counts. *Statistical Modelling*, Sage Publications Sage India: New Delhi, India, v. 17, n. 6, p. 359–380, 2017. Cited 2 times on pages 31 and 79.

INOUYE, D. I. et al. A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 9, n. 3, p. e1398, 2017. Cited on page 16.

JØRGENSEN, B.; KOKONENDJI, C. C. Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, Springer, v. 100, n. 1, p. 43–78, 2016. Cited on page 15.

KOKONENDJI, C. C.; PUIG, P. Fisher dispersion index for multivariate count distributions: A review and a new proposal. *Journal of Multivariate Analysis*, Elsevier, v. 165, p. 180–193, 2018. Cited 2 times on pages 20 and 23.

KRISTENSEN, K. et al. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, v. 70, n. 5, p. 1–21, 2016. Cited 3 times on pages 17, 44, and 52.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 58, n. 4, p. 619–678, 1996. ISSN 00359246. Disponível em: ⟨http://www.jstor.org/stable/2346105⟩. Cited on page 16.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986. Cited on page 16.

LLC, W. *WolframAlpha*. Last visited on 12/8/2020. Disponível em: ⟨https://www.wolframalpha.com/⟩. Cited on page 40.

LOEYS, T. et al. The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 65, n. 1, p. 163–180, 2012. Cited on page 15.

MUñOZ-PICHARDO, J. M. et al. A multivariate poisson regression model for count data. *Journal of Applied Statistics*, Taylor & Francis, v. 0, n. 0, p. 1–17, 2021. Disponível em: ⟨https://doi.org/10.1080/02664763.2021.1877637⟩. Cited on page 16.

NATIONAL Health and Nutrition Examination Survey datasets 2007. 2007. ⟨https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2007⟩. Accessed: 2020-08-14. Cited on page 19.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Cited on page 14.

NHANES - National Health and Nutrition Examination Survey. 2007. ⟨https://www.cdc.gov/nchs/nhanes/about_nhanes.htm⟩. Accessed: 2020-08-14. Cited on page 19.

NIKOLOULOPOULOS, A. K.; KARLIS, D. Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 39, n. 1, p. 172–187, 2009. Cited on page 16.

NOCEDAL, J.; WRIGHT, S. *Numerical optimization*. [S.l.]: Springer Science & Business Media, 2006. Cited 3 times on pages 17, 39, and 42.

PAWITAN, Y. *In all likelihood: statistical modelling and inference using likelihood*. [S.l.]: Oxford University Press, 2001. Cited on page 36.

PINHEIRO, J. et al. *nlme: Linear and Nonlinear Mixed Effects Models*. [S.l.], 2017. R package version 3.1-131. Disponível em: ⟨https://CRAN.R-project.org/package=nlme⟩. Cited on page 17.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: ⟨https://www.R-project.org/⟩. Cited 3 times on pages 17, 21, and 44.

RIBEIRO, J. E. E. et al. Reparametrization of com–poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, v. 20, n. 5, p. 443–466, 2020. Disponível em: ⟨https://doi.org/10.1177/1471082X19838651⟩. Cited on page 79.

RIDOUT, M.; DEMÉTRIO, C. G.; HINDE, J. Models for count data with many zeros. In: INTERNATIONAL BIOMETRIC SOCIETY INVITED PAPERS CAPE TOWN, SOUTH AFRICA. *Proceedings of the XIXth international biometric conference*. [S.l.], 1998. v. 19, p. 179–192. Cited on page 15.

SHMUELI, G. et al. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 1, p. 127–142, 2005. Cited 2 times on pages 15 and 31.

SIGNORELLI, M.; SPITALI, P.; TSONAKA, R. Poisson–tweedie mixed-effects model: A flexible approach for the analysis of longitudinal rna-seq data. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, p. 1471082X20936017, 2020. Cited on page 37.

TAMIOSO, P. R. et al. Inducing positive emotions: Behavioural and cardiac responses to human and brushing in ewes selected for high vs low social reactivity. *Applied Animal Behaviour Science*, v. 208, p. 56 – 65, 2018. ISSN 0168-1591. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S0168159118304490⟩. Cited on page 14.

THYGESEN, U. H. et al. Validation of ecological state space models using the laplace approximation. *Environmental and Ecological Statistics*, Springer, v. 24, n. 2, p. 317–339, 2017. Cited on page 17.

TIERNEY, L.; KADANE, J. B. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, Taylor & Francis, v. 81, n. 393, p. 82–86, 1986. Disponível em: ⟨https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240⟩. Cited 2 times on pages 17 and 37.

WANG, Y. et al. *mvabund: Statistical Methods for Analysing Multivariate Abundance Data*. [S.l.], 2020. R package version 4.1.3. Disponível em: ⟨https://CRAN.R-project.org/package=mvabund⟩. Cited on page 21.

WEDDERBURN, R. W. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, Oxford University Press, v. 61, n. 3, p. 439–447, 1974. Cited on page 16.

WINKELMANN, R. *Econometric analysis of count data*. [S.l.]: Springer Science & Business Media, 2008. Cited 2 times on pages 16 and 31.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. Regression models for count data in r. *Journal of statistical software*, Foundation for Open Access Statistics, v. 27, n. 8, p. 1–25, 2008. Cited on page 15.

ZEVIANI, W. M. et al. The gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, Taylor & Francis, v. 41, n. 12, p. 2616–2626, 2014. Cited on page 15.

Appendix

APPENDIX A – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME AND FINAL MODEL FOR ANT DATASET

FIGURE 22 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME 13-24 AND FINAL MODEL FOR EACH DISTRIBUTION

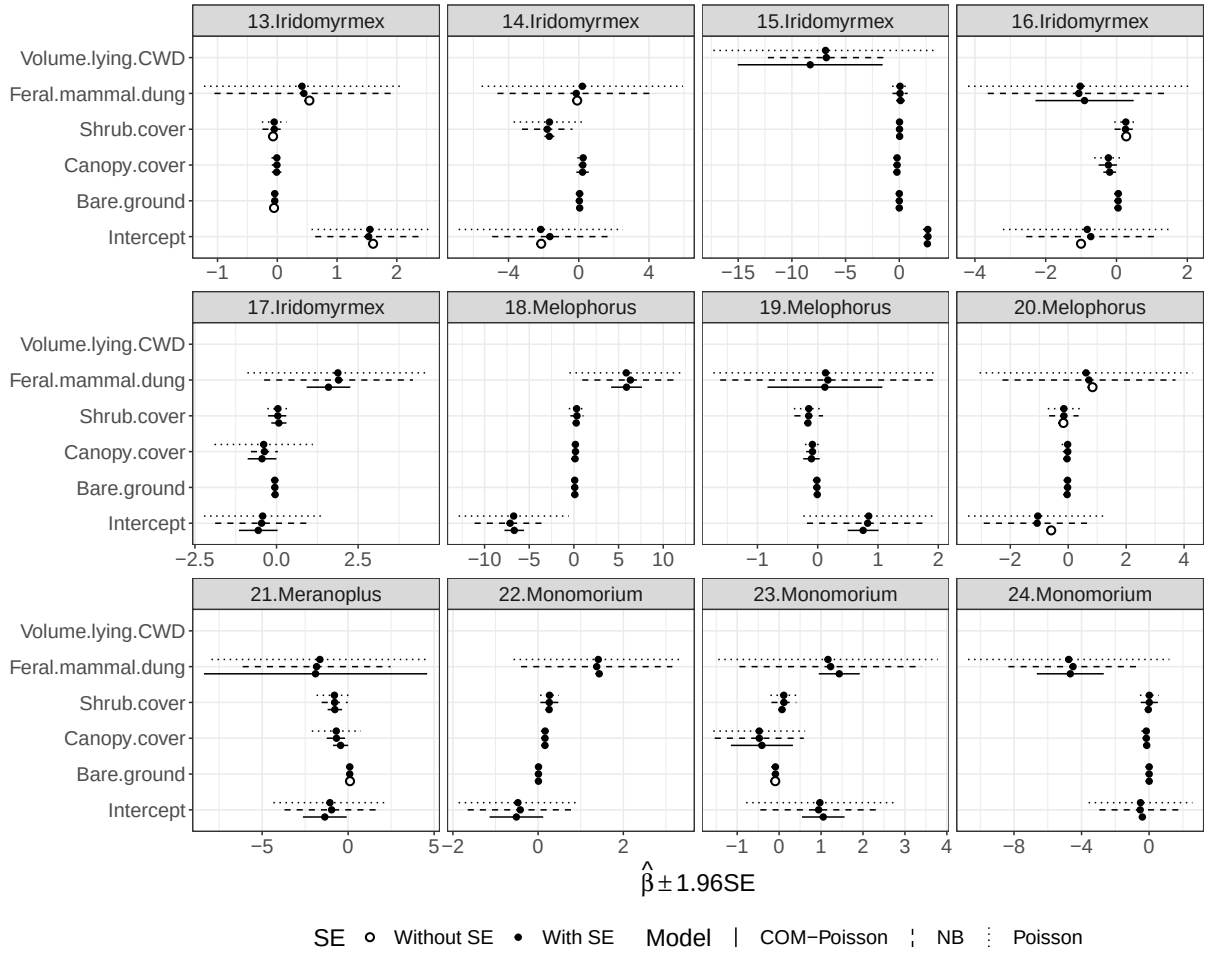


FIGURE 23 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME 25-36 AND FINAL MODEL FOR EACH DISTRIBUTION

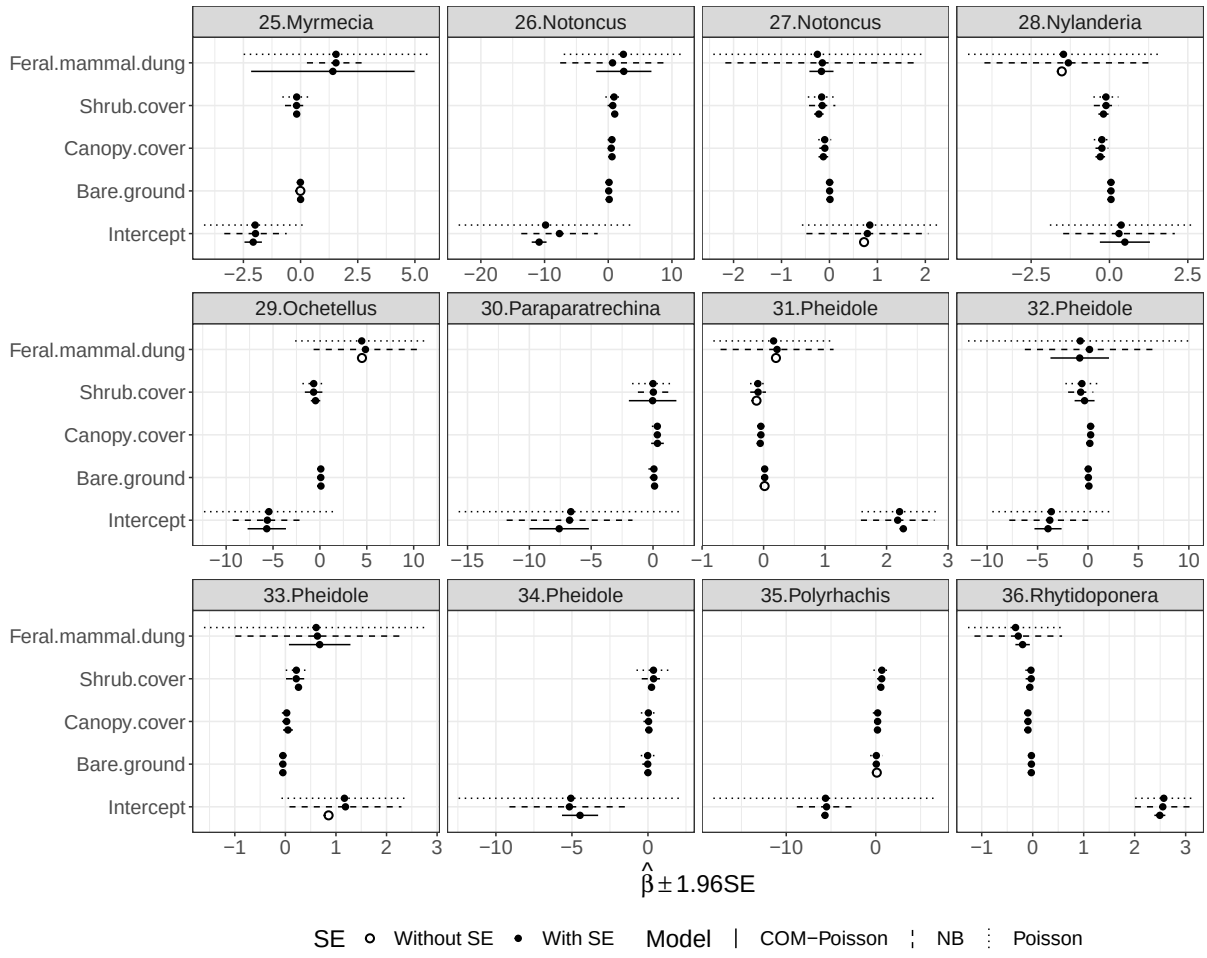
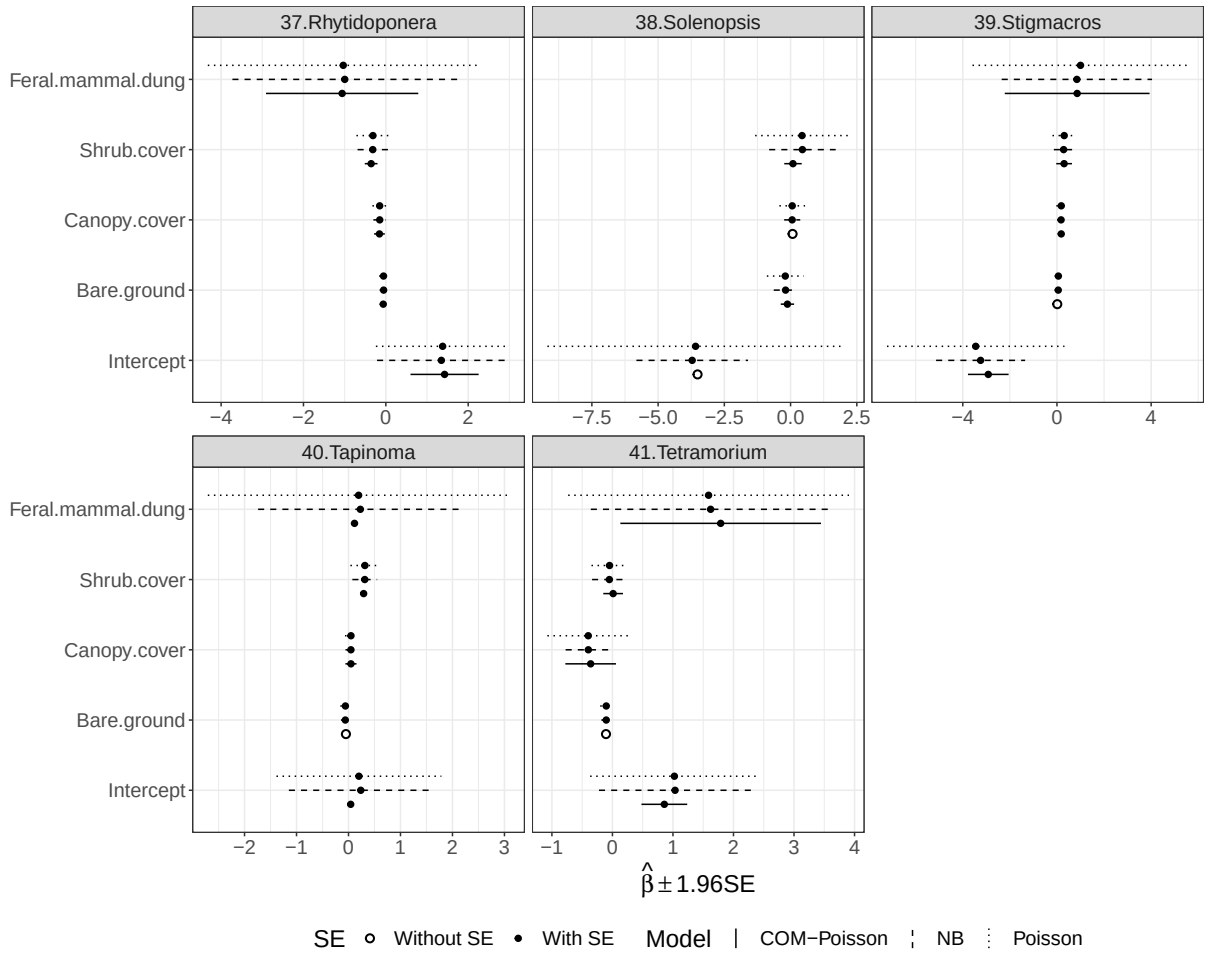


FIGURE 24 – REGRESSION PARAMETER ESTIMATES AND 95% CONFIDENCE INTERVALS BY OUTCOME 37-42 AND FINAL MODEL FOR EACH DISTRIBUTION



Annex

ANNEX A – TO BE USED

ANNEX B – TO BE USED