Multivariate Regression Models for Count Data

Guilherme Parreira da Silva

Mestrando do PPGMNE
Programa de Pós Graduação em Métodos numéricos (PPGMNE)
Laboratório de Estatística e Geoinformação (LEG)
Universidade Federal do Paraná (UFPR)
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)

15 de Dezembro de 2020

PPGMNE PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS NUMÉRICOS EM ENGENHARIA

leg

UFPR UNIVERSIDADE FEDERAL DO PARANÁ

CAPES

Agenda

Motivation

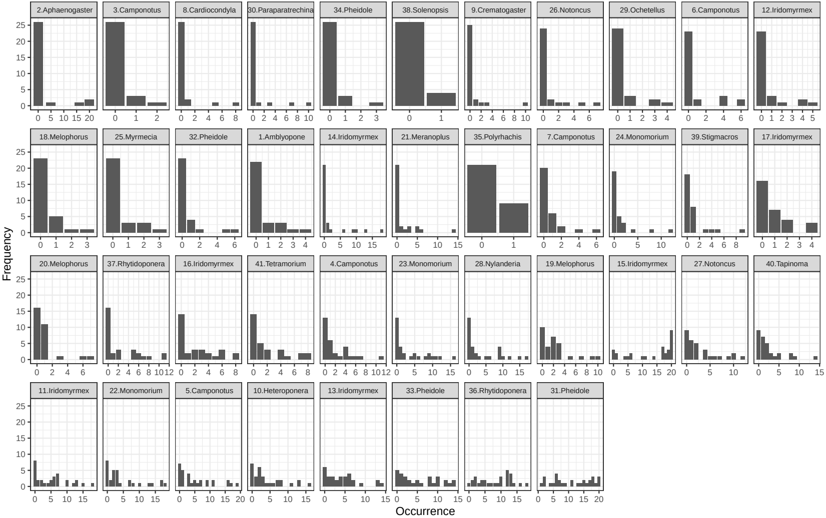
Dataset I - Australian Health Survey (AHS)

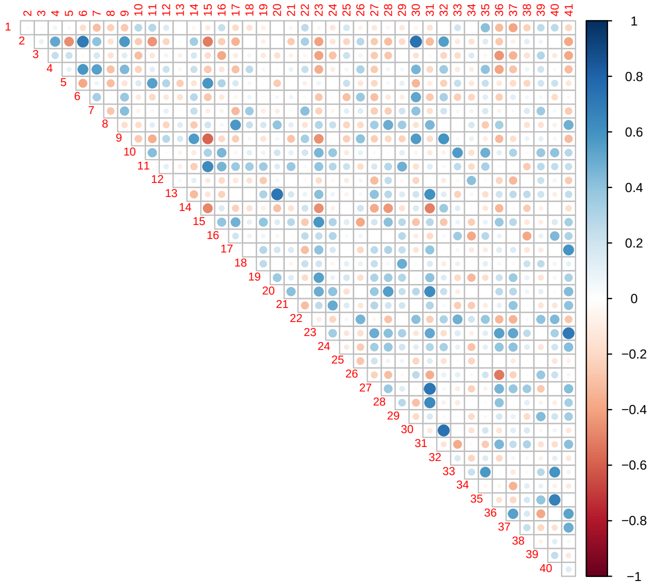
- Largest survey in Australia concerning health - 1987-88.
- Objectives:
  - To investigate whether a covariate set is associated with a set of response variables.
  - To investigate a possible relationship between response variables.
- Five response variables, which are **number of**:
  - Consultations with a doctor or specialist.
  - Consultations with health professionals.
  - Admissions to a hospital in the past 12 months.
  - Nights in a hospital during the most recent admission.
  - Medications used in the past two days.
- 10 covariates, among sociodemographic, income, health insurance and status.
- Sample size of 5190 respondents.

Dataset II - Ant Species Abudance

- Occurrence of 41 different species that fell into a pitfall trap.
- Objectives:
    - Investigate whether environmental covariates are related to the occurrence of ants.
    - Investigate whether different ant species occurs together.
- Sample size: 30 different sites in south-eastern Australia.
- 5 covariates that represent characteristics of each site.

Ant Species - Barplot




2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Introduction

- Regression is a key concept under the statistical modelling.
- Univariate regression models are used to investigate the relationship between a set of covariates and one response variable.
- Standard regression models:
  - Linear model (LM) - (GALTON, 1886):
    - Deal only with continuous data.
    - Assumptions: Gaussian, independence and homogeneous variance.

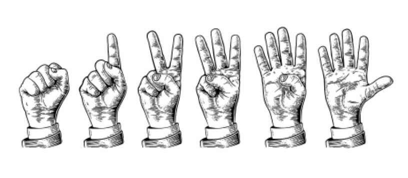
- Regression is a key concept under the statistical modelling.
- Univariate regression models are used to investigate the relationship between a set of covariates and one response variable.
- Standard regression models:
  - Linear model (LM) - (GALTON, 1886):
    - Deal only with continuous data.
    - Assumptions: Gaussian, independence and homogeneous variance.
  - Generalized linear model (GLM) - (NELDER; WEDDERBURN, 1972):
    - Link function connects the linear predictor to the response variable.
    - Variance is related to the mean.
    - Distribution belongs to the exponential family.

Count data - Definition

- Represents the number of times that an event occur in a fixed time interval, such as, time, space, distance, area, among others:
  - Example: Number of e-mails in the inbox in one day.

Count data - Definition

- Represents the number of times that an event occur in a fixed time interval, such as, time, space, distance, area, among others:
  - Example: Number of e-mails in the inbox in one day.



Count data - Definition

- Represents the number of times that an event occur in a fixed time interval, such as, time, space, distance, area, among others:
  - Example: Number of e-mails in the inbox in one day.

- The main way to describe this variable is based on the mean-variance relationship:
  - Overdispersion: Variance $>$ Mean.
  - Equidispersion: Variance $=$ Mean.
  - Subdispersion: Variance $<$ Mean.

- Equidispersion: Poisson.

Distributions to model count data

- Equidispersion: Poisson.
- Overdispersion: Negative Binomial (NB).

Distributions to model count data

- Equidispersion: Poisson.
- Overdispersion: Negative Binomial (NB).
- General models:
  - Extended Poisson Tweedie (BONAT et al., 2018).
  - Conway-Maxwell-Poisson (COM-Poisson) (SHMUELI et al., 2005).
  - Gamma Count (ZEVIANI et al., 2014).

Distributions to model count data

- Equidispersion: Poisson.
- Overdispersion: Negative Binomial (NB).
- General models:
  - Extended Poisson Tweedie (BONAT et al., 2018).
  - Conway-Maxwell-Poisson (COM-Poisson) (SHMUELI et al., 2005).
  - Gamma Count (ZEVIANI et al., 2014).
- Drawback:
  - The probability mass function (pmf) is not available in closed form.
  - Numerical methods are needed to compute the pmf.

Approaches for Multivariate data

- All models presented consider only one response variable.
- Different approaches to deal with multivariate responses:
  - Constructing multivariate distributions for couting data (FAMOYE, 2015; INOUYE et al., 2017).
  - Copula is a general framework to build multivariate distributions based on copulas functions (NIKOLOULOPOULOS; KARLIS, 2009).
  - BONAT (2016) proposed the Multivariate Covariance Generalized Linear Models (MCGLM).
  - Via Bayesian inferece:
    - BRMS package - Baeysian Regression Models using Stan (BÜRKNER, 2018).
    - MCMCglmm package - MCMC Generalised Linear Mixed Models (HADFIELD, 2010).

Approaches for Multivariate data

- All models presented consider only one response variable.
- Different approaches to deal with multivariate responses:
  - Constructing multivariate distributions for couting data (FAMOYE, 2015; INOUYE et al., 2017).
  - Copula is a general framework to build multivariate distributions based on copulas functions (NIKOLOULOPOULOS; KARLIS, 2009).
  - BONAT (2016) proposed the Multivariate Covariance Generalized Linear Models (MCGLM).
  - Via Bayesian inferece:
    - BRMS package - Baeysian Regression Models using Stan (BÜRKNER, 2018).
    - MCMCglmm package - MCMC Generalised Linear Mixed Models (HADFIELD, 2010).
  - GLMM using non observed random effects (BRESLOW; CLAYTON, 1993).

Objectives

General:

- Propose the Multivariate generalized linear mixed model (MGLMM) - A multivariate modelling framework to deal with count data under the GLMM approach.

Specific:

- Computational implementation.
- Simulation studies.
- Analyse three datasets.

MGLMM

- Let Y_{ir} be the multivariate outcome for subject i, $i = 1, \ldots, n$ and response variable r, $r = 1, \ldots, k$.
- Let p be known covariates set is available for each response r.
- Let x_{irj} be the value of the *j-th* covariate for individual i and response r.

Joint model based on a GLMM with a random intercept:

$$\mathrm{Y}_{ir} \mid \mathrm{b}_{ir} \sim f(\mu_{ir}; \phi_r),$$

where f is a pmf, e.g. Poisson, NB, COM-Poisson.

- Linear predictor:

$$\mathrm{g}_r(\mu_{ir}) = x_{irj}^T \boldsymbol{\beta}_r + \mathrm{b}_{ir},$$

where:

- $\mathrm{g}_r(\mu_{ir})$ is a suitable link function (log).
- $\boldsymbol{\beta}_r$ is a $\mathrm{px1}$ vector of covariate.
- b_{ir} is the random intercept value for each sample unit and response variable.

The random effects distribution:

$$\begin{pmatrix} \mathrm{b}_{i1} \\ \mathrm{b}_{i2} \\ \vdots \\ \mathrm{b}_{ir} \end{pmatrix} \sim \mathrm{NM}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} ; \sum_{r\times r} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1r}\sigma_1\sigma_r \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2r}\sigma_2\sigma_r \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{r1}\sigma_r\sigma_1 & \rho_{r2}\sigma_r\sigma_2 & \dots & \sigma_r^2 \end{bmatrix} \right).$$

- Can ϕ_r and σ_r^2 be estimated simultaneously?

Estimation and inference

Maximum likelihood (ML) estimation

- Joint distribution:

$$f(\mathrm{Y}, \mathrm{b}) = f(\mathrm{Y}|\mathrm{b})f(\mathrm{b}).$$

- Marginal distribution:

$$f(\mathrm{Y}) = \int f(\mathrm{Y}|\mathrm{b})f(\mathrm{b})\mathrm{db}.$$

Maximum likelihood (ML) estimation

- Joint distribution:

$$f(\mathrm{Y}, \mathrm{b}) = f(\mathrm{Y}|\mathrm{b})f(\mathrm{b}).$$

- Marginal distribution:

$$f(\mathrm{Y}) = \int f(\mathrm{Y}|\mathrm{b})f(\mathrm{b})\mathrm{db}.$$

- The goal is to estimate the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^{\mathbf{2}}, \boldsymbol{\rho})^\top$.
- Marginal likelihood:

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

  where $\mathbf{y}$ is a k-response vector and $\mathbf{b}$ a k-random effect vector.
- Full likelihood:

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

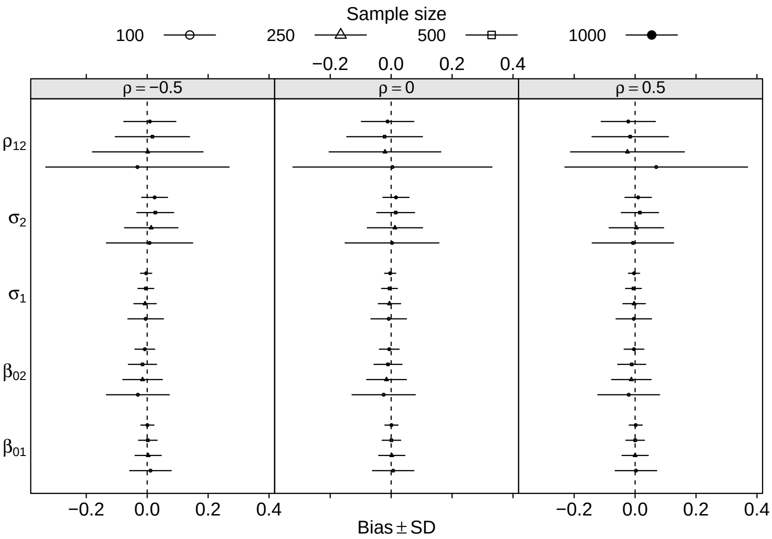
  where N is the total number of sample units.

1. Marginal likelihood:
   - Laplace Approximation.
2. Optimization:
   - BFGS and PORT.
3. Computational tools:
   - Software and programming language R.
   - TMB package written in C++ with CppAD and Eigen C++ libraries (KRISTENSEN et al., 2016).
   - Automatic Differentiation.

Results

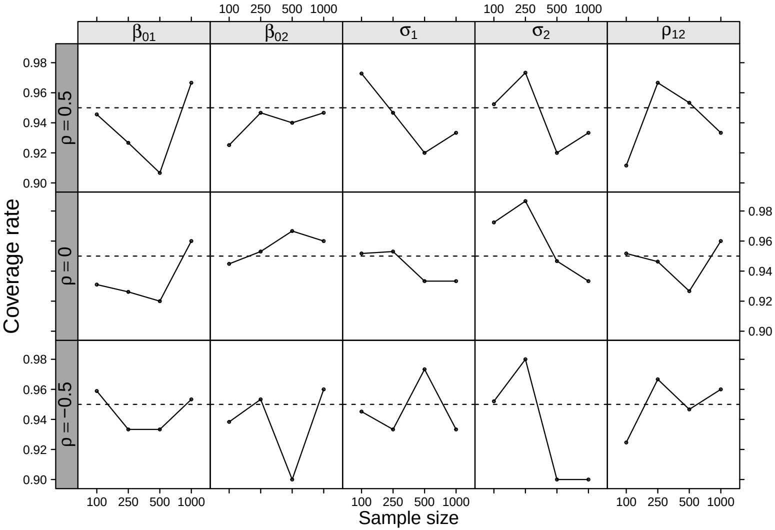
Simulation Study Design - Poisson

- Objective:
    - To investigate the property of the estimators:
        - Bias.
        - Consistency.
        - Coverage rate.



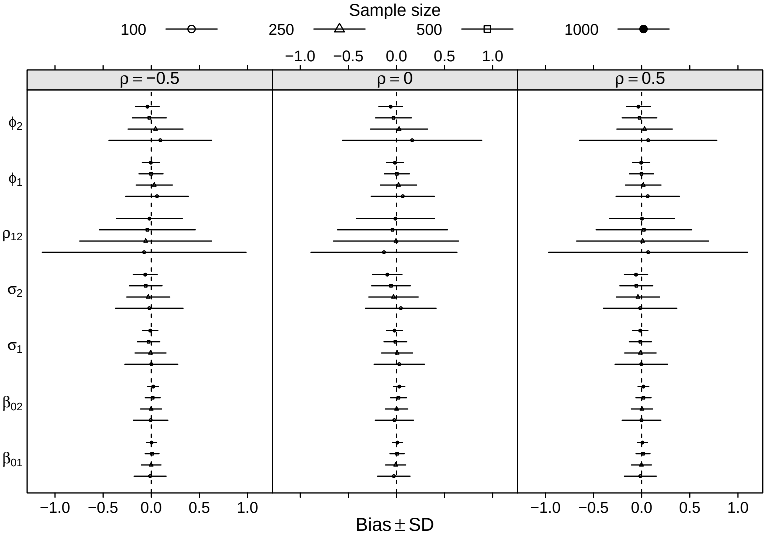
Parameter Configuration: 2 response variables; $\beta_{01} = \log(7)$; $\beta_{02} = \log(1.5)$; $\rho = \{-.5, 0, .5\}$;

$\sigma_1^2 = .3$ ($\sigma_1 = .55$); $\sigma_2^2 = .15$ ($\sigma_2 = .39$); Sample size $= \{100, 250, 500, 1000\}$.



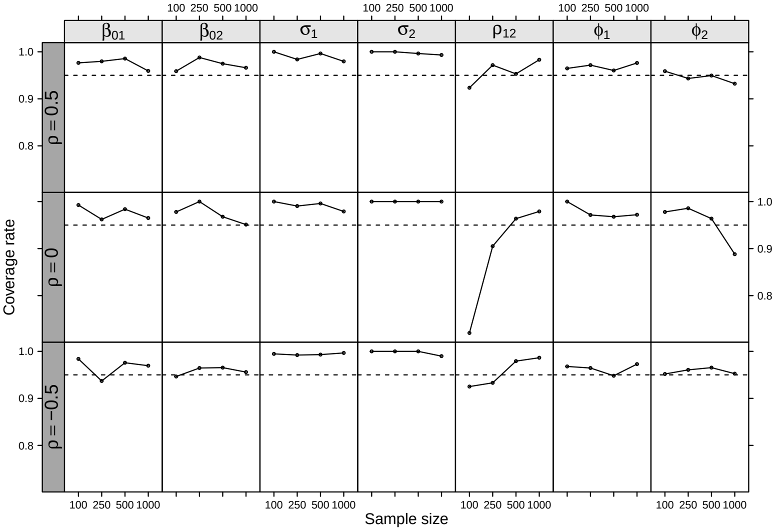
Parameter Configuration: 2 response variables; $\beta_{01} = \log(7)$; $\beta_{02} = \log(1.5)$; $\rho = \{-.5, 0, .5\}$;

$\sigma_1^2 = .3$ ($\sigma_1 = .55$); $\sigma_2^2 = .15$ ($\sigma_2 = .39$); Sample size $= \{100, 250, 500, 1000\}$.



Parameter Configuration: 2 response variables; $\beta_{01} = \log(7)$; $\beta_{02} = \log(1.5)$; $\rho = \{-.5, 0, .5\}$;

$\sigma_1^2 = .3$ ($\sigma_1 = .55$); $\sigma_2^2 = .15$ ($\sigma_2 = .39$);$\boldsymbol{\phi_r = 1}$; Sample size $= \{100, 250, 500, 1000\}$.



Parameter Configuration: 2 response variables; $\beta_{01} = \log(7)$; $\beta_{02} = \log(1.5)$; $\rho = \{-.5, 0, .5\}$; $\sigma_1^2 = .3$ ($\sigma_1 = .55$); $\sigma_2^2 = .15$ ($\sigma_2 = .39$);$\boldsymbol{\phi_r = 1}$; Sample size $= \{100, 250, 500, 1000\}$.

Reference I

BONAT, W. H. **mcglm: Multivariate Covariance Generalized Linear Models**. Traducao. [s.l: s.n.].

BONAT, W. H. et al. Extended Poisson–Tweedie: Properties and regression models for count data. **Statistical Modelling**, v. 18, n. 1, p. 24–49, 2018.

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

BÜRKNER, P.-C. Advanced Bayesian Multilevel Modeling with the R Package brms. **The R Journal**, v. 10, n. 1, p. 395–411, 2018.

FAMOYE, F. A Multivariate Generalized Poisson Regression Model. **Comm. Statist. Theory Methods**, v. 44, n. 3, p. 497–511, fev. 2015.

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

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.

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

KRISTENSEN, K. et al. TMB: Automatic Differentiation and Laplace Approximation. **Journal of Statistical Software**, v. 70, n. 5, p. 1–21, 2016.

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

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

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)**, v. 54, n. 1, p. 127–142, 2005.

ZEVIANI, W. M. et al. The Gamma-count distribution in the analysis of experimental underdispersed data. **Journal of Applied Statistics**, v. 41, n. 12, p. 2616–2626, 2014.

- Thank you for your atention.
  - Advisor: Wagner Hugo Bonat.
  - Co-advisor: Paulo Justiniano Ribeiro Júnior.
  - Contact:
    - guilhermeparreira.silva@gmail.com.
    - https://github.com/guilhermeparreira.
  - Support:

PPGMNE PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS NUMÉRICOS EM ENGENHARIA

leg

UFPR UNIVERSIDADE FEDERAL DO PARANÁ

CAPES