Multivariate Regression Models in R: The mcglm package

Prof. Wagner Hugo Bonat

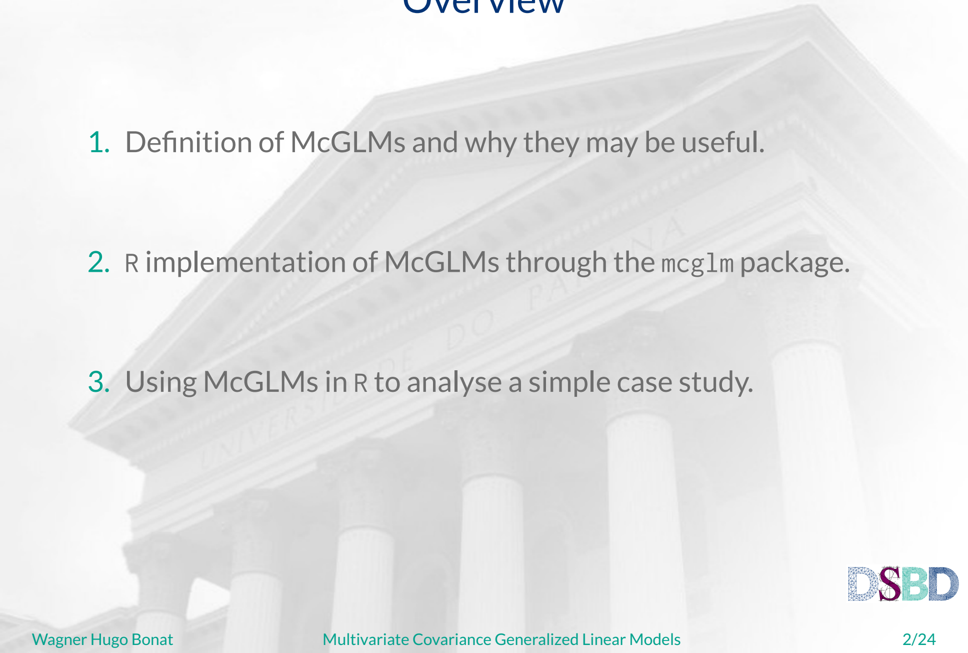
R Day

Laboratório de Estatística e Geoinformação - LEG

Universidade Federal do Paraná - UFPR

15 de maio de 2018

Overview



1. Definition of McGLMs and why they may be useful.
2. `R` implementation of McGLMs through the `mcglm` package.
3. Using McGLMs in `R` to analyse a simple case study.

What are McGLMs?

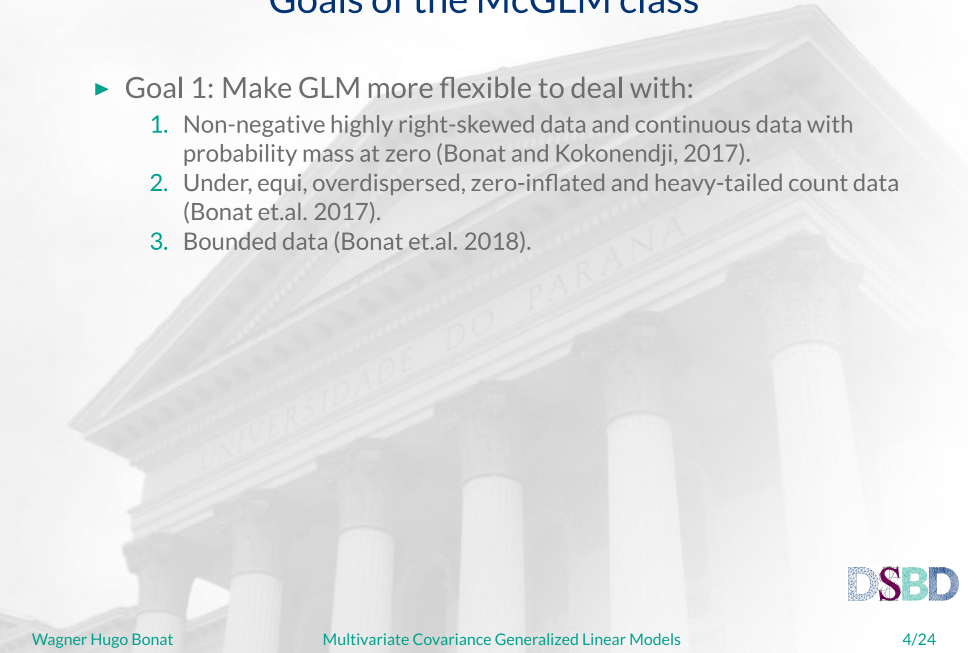
- McGLMs stand for

  Multivariate Covariance Generalized Linear Models.

- General statistical modelling framework proposed by Bonat and Jørgensen (2016).
- McGLMs extend the orthodox Generalized Linear Models (Nelder and Wedderburn, 1972) in many different ways.

Goals of the McGLM class

- Goal 1: Make GLM more flexible to deal with:
  1. Non-negative highly right-skewed data and continuous data with probability mass at zero (Bonat and Kokonendji, 2017).
  2. Under, equi, overdispersed, zero-inflated and heavy-tailed count data (Bonat et.al. 2017).
  3. Bounded data (Bonat et.al. 2018).



Goals of the McGLM class

- Goal 1: Make GLM more flexible to deal with:
    1. Non-negative highly right-skewed data and continuous data with probability mass at zero (Bonat and Kokonendji, 2017).
    2. Under, equi, overdispersed, zero-inflated and heavy-tailed count data (Bonat et.al. 2017).
    3. Bounded data (Bonat et.al. 2018).
- Goal 2: Extend GLM to cGLM to deal with:
    1. Mixed models (Bonat et. al., 2017).
    2. Repeated measures and longitudinal data (Bonat et. al., 2017).
    3. Time series.
    4. Spatial and space-time data (Bonat and Jørgensen, 2016).
    5. Genetic and Twin data (Bonat and Hjelmborg, 2018).

Goals of the McGLM class

- Goal 1: Make GLM more flexible to deal with:
  1. Non-negative highly right-skewed data and continuous data with probability mass at zero (Bonat and Kokonendji, 2017).
  2. Under, equi, overdispersed, zero-inflated and heavy-tailed count data (Bonat et.al. 2017).
  3. Bounded data (Bonat et.al. 2018).
- Goal 2: Extend GLM to cGLM to deal with:
  1. Mixed models (Bonat et. al., 2017).
  2. Repeated measures and longitudinal data (Bonat et. al., 2017).
  3. Time series.
  4. Spatial and space-time data (Bonat and Jørgensen, 2016).
  5. Genetic and Twin data (Bonat and Hjelmborg, 2018).
- Goal 3: Extend cGLM to McGLM to deal with:
  1. Multiple response variables.
  2. Response variables of mixed types.

Generalized Linear Models

- Let $\boldsymbol{Y}$ be an $N \times 1$ response vector.
- Let $\boldsymbol{X}$ be an $N \times k$ design matrix.
- Let $\boldsymbol{\beta}$ be a $k \times 1$ parameter vector.
- Orthodox GLM

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \\ \mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}(\tau_0 \boldsymbol{I})\mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}. \end{aligned} \tag{1}$$

- g is the link function.
- $\mathrm{V}(\boldsymbol{\mu}; p) = \mathrm{diag}(\vartheta(\boldsymbol{\mu}; p))$ where $\vartheta(\boldsymbol{\mu}; p)$ is the variance function.
- p and τ_0 are the power and dispersion parameters.
- $\boldsymbol{I}$ is an identity matrix.

Variance functions

- Tweedie family (variance function), characterized by

$$\vartheta(\boldsymbol{\mu}; p) = \mu^p,$$

  Gaussian ($p = 0$), gamma ($p = 2$) and inverse Gaussian ($p = 3$).
- Deals with symmetric, skewed and zero-inflated continuous outcomes.
- Binomial family, characterized by

$$\vartheta(\boldsymbol{\mu}) = \mu(1 - \mu).$$

- Deals with binary, binomial and bounded outcomes.
- Extended binomial variance function,

$$\vartheta(\boldsymbol{\mu}; p, q) = \mu^p(1 - \mu)^q.$$

- Extra flexibility to deal with binomial and bounded outcomes.

Dispersion functions

- Poisson-Tweedie family (dispersion function), characterized by

  $$\vartheta(\boldsymbol{\mu}; p) = \mu + \mu^{p}$$

  Hermite ($p = 0$), Neyman Type A ($p = 1$), negative binomial ($p = 2$).
- Special form in (1) $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\mu}) + \mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}(\tau_0 \boldsymbol{I})\mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}$.
- Deals with under, equi, overdispersed, heavy-tailed and zero-inflated count outcomes.

Covariance Generalized Linear Models

- Change the identity matrix in (1) to a non-diagonal matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$.

$$\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}(\boldsymbol{\Omega}(\boldsymbol{\tau}))\mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}.$$

- Similar to *working correlation* matrix (GEE) (Liang and Zeger, 1987).
- Model $\boldsymbol{\Omega}(\boldsymbol{\tau})$ as a linear combination of known matrices.

$$h(\boldsymbol{\Omega}(\boldsymbol{\tau})) = \tau_0 Z_0 + \cdots + \tau_D Z_D$$

  where h is a covariance link function (Pourahmadi, 2011).
- Z_d with $d = 0, \ldots, D$ are known matrices.
- $\boldsymbol{\tau} = (\tau_0, \cdots, \tau_D)$ is a $D \times 1$ dispersion parameter vector.
- Identity, inverse, exponential-matrix, modified Cholesky decomposition, etc.

Multivariate Covariance Generalized Linear Models

- Let $\boldsymbol{Y}_{N \times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ be an outcome matrix.
- Let $\boldsymbol{M}_{N \times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ be an expected value matrix.
- Let $\boldsymbol{\Sigma}_r = \mathrm{V}_r(\boldsymbol{\mu}; p)^{\frac{1}{2}} \boldsymbol{\Omega}_r(\boldsymbol{\tau}) \mathrm{V}_r(\boldsymbol{\mu}; p)^{\frac{1}{2}}$ be the covariance matrix within outcomes.
- Let $\boldsymbol{\Sigma}_b$ be the correlation matrix between outcomes.
- We define the McGLM by,

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1 \boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R \boldsymbol{\beta}_R)\} \\ \mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b \end{aligned}$$

  where
  $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \ldots, \tilde{\boldsymbol{\Sigma}}_R^T)$.
- Generalized Kronecker product proposed by Martinez-Beneito (2013).
- $\tilde{\boldsymbol{\Sigma}}_r$ is the lower triangular matrix of the Cholesky decomposition of $\boldsymbol{\Sigma}_r$.
- Bdiag denotes a block diagonal matrix.

Special cases: Linear covariance models

- Double generalized linear models.
- Linear mixed models.
- Moving average models.
- Exchangeable or compound symmetry and unstructured models (popular in longitudinal data analysis).
- Conditional autoregressive models (time series, spatial and space-time data).
- Models in quantitative genetic and phylogenetic.
- Models for Twin and family data.
- Many more . . .
- McGLMs provide multivariate extensions for all these modelling strategies.

Implementation in R

- Package `mcglm` available at `github` and CRAN.
  1. `https://cran.r-project.org/web/packages/mcglm/index.html`.
  2. `https://github.com/wbonat/mcglm`.
- Main mathematical tool: Linear Algebra.
- Computational expansive tasks: Cholesky decomposition and matrix multiplication.
- Main dependences:
  1. Matrix package.
  2. Version ($>$ 0.4.0) RcppArmadillo.

Implementation in R

- Installation

```
library(devtools)
install_github("wbonat/mcglm")
install.packages("mcglm")
```

- Main function `mcglm`.

```
require(mcglm)
args(mcglm)

## function (linear_pred, matrix_pred, link, variance, covariance,
##     offset, Ntrial, power_fixed, data, control_initial = "automatic",
##     contrasts = NULL, control_algorithm = list())
## NULL
```

- Link functions: `logit`, `probit`, `cauchit`, `cloglog`, `loglog`, `identity`, `log`, `sqrt`, `inverse`.
- Variance functions: `constant`, `tweedie`, `poisson_tweedie`, `binomialP` and `binomialPQ`.
- Covariance link functions: `identity`, `inverse` and `exponential-matrix`.

Linear covariance structures

Tabela 1. Linear covariance structures available.

Functions	Description
`mc_id()`	Identity matrix.
`mc_ns()`	Unstructured model.
`mc_dglm()`	Double generalized linear models.
`mc_mixed()`	Linear mixed models (formula similar to lme4).
`mc_ma()`	Moving average models of order p.
`mc_rw()`	CAR models for times series.
`mc_car()`	CAR models for space data.
`mc_dist()`	Distance based models.
`mc_twin()`	ACE, ADE, AE, and CE models for twin.

- The users can use any list of symmetric matrices.
- Combine pre-specified structures with new ones.

Methods

Tabela 2. Methods available for objects of mcglm class.

Functions	Description
`print()`	Simple printed display of model features.
`summary()`	Standard regression output.
`fitted()`	Fitted values for observed data.
`residuals()`	Pearson, raw and standardized residuals.
`coef()`	Coefficient estimates.
`vcov()`	Variance-covariance matrix of coefficient estimates.
`confint()`	Confidence intervals.
`anova()`	Analysis of variance tables for fitted models.
`plot()`	Diagnostic plots of Pearson residuals and algorithm check.

Extra features

Tabela 3. Extra features for objects of mcglm class.

Functions	Description
gof()	Measures of goodness-of-fit.
mc_sic()	SIC for regression parameters.
mc_sic_covariance()	SIC for dispersion parameters.
mc_bias_correct_std()	Bias-corrected std.
mc_robust_std	Robust std.
mc_conditional_test	Conditional hypotheses tests.
mc_compute_rho	Compute autocorrelation estimates.
mc_initial_values	Initial values for mcglm.

- GOF's implemented: plogLik, pAIC, pKLIC, pBIC, ESS, GOSHO and RJC.

Mixed response variables

- Experiment carried out in a vegetation house with soybeans.
- Two plants by plot with three levels of the factor amount of water in the soil (`water`).
- Five levels of potassium fertilization (`pot`).
- The plots were arranged in five blocks (`block`).
- Three response variables are of the interest:
  1. Grain yield - continuous.
  2. Number of seeds - counting.
  3. Number of viable peas per plant - binomial.
- Main data analysis goal: assess the effect of the covariates `water` and `pot`.
- Data set available in the `mcglm` package (`soya`).

R code

```
# Loading mcglm package
require(mcglm)
# Loading the data set
data("soya")
# Linear predictor
form.grain <- grain ~ block + water * pot
form.seed <- seeds ~ block + water * pot
soya$viablepeasP <- soya$viablepeas / soya$totalpeas
form.peas <- viablepeasP ~ block + water * pot
# Matrix linear predictor (identity - independent observations)
Z0_ex4 <- mc_id(soya)
# Univariate fit
fit.grain <- mcglm(linear_pred = c(form.grain), matrix_pred = list(Z0_ex4),
                   data = soya)


## Automatic initial values selected.


fit.seed <- mcglm(linear_pred = c(form.seed), matrix_pred = list(Z0_ex4),
                  link = "log", variance = "poisson_tweedie",
                  power_fixed = TRUE, data = soya)


## Automatic initial values selected.


fit.peas <- mcglm(linear_pred = c(form.peas), matrix_pred = list(Z0_ex4),
                  link = "logit", variance = "binomialP",
                  Ntrial = list(soya$totalpeas), data = soya)


## Automatic initial values selected.
```

R code

- Multivariate fit

```
# Multivariate fit
fit.joint <- mcglm(linear_pred = c(form.grain, form.seed, form.peas),
                   matrix_pred = list(Z0_ex4, Z0_ex4, Z0_ex4),
                   link = c("identity","log", "logit"),
                   variance = c("constant", "poisson_tweedie", "binomialP"),
                   Ntrial = list(NULL, NULL, soya$totalpeas), data = soya)

## Automatic initial values selected.
```

- Comparing univariate and multivariate fits

```
rbind(gof(list(fit.grain, fit.seed, fit.peas)), gof(fit.joint))

##   plogLik Df   pAIC    pKLIC      pBIC
## 1 -339.54 60 799.08 833.0497 1004.0460
## 2 -319.73 63 765.46 847.7457  980.6743
```

R code: Output

- Multivariate hypotheses tests

```
manova.mcglm(fit.joint)

##     Effects Df Hotelling.Lawley Qui.square      p_value
## 1 Intercept  3          124.415   9331.107 0.000000e+00
## 2     block 12            0.531     39.840 7.645264e-05
## 3     water  6            0.201     15.107 1.944296e-02
## 4       pot 12            3.339    250.435 1.110176e-46
## 5 water:pot 24            1.540    115.536 6.007058e-14
```

- Correlation estimates

```
summary(fit.joint, verbose = TRUE, print = "Correlation")

## Correlation matrix:
##   Parameters  Estimates Std.error   Z value
## 1      rho12 0.63766799 0.1417181 4.4995521
## 2      rho13 0.07034875 0.1156794 0.6081352
## 3      rho23 0.08882827 0.1152445 0.7707810
##
## Algorithm: chaser
## Correction: TRUE
## Number iterations: 10
```

Further examples

- Flexible Tweedie regression models (Bonat and Kokonendji, 2017).
  1. Smoothing time series of rainfall in Curitiba (semi-continuous).
  2. Income dynamics in Australia (right-skewed)
  3. Gain in weight of rats (symmetric).
- Extended Poisson-Tweedie: properties and regression models (Bonat, et. al. 2017).
  1. Respiratory disease morbidity among children in Curitiba (overdispersed).
  2. Cotton bolls greenhouse experiment (underdispersed).
  3. Radiation-induced chromosome aberration counts (zero-inflated).
  4. Customers profile (equidispersed).
- Multivariate covariance generalized linear models (Bonat and Jørgensen (2016)).
  1. Australian health survey (multivariate count).
  2. Respiratory physiotherapy on premature (outcomes of mixed types).
  3. Venezuelan rainfall data (space-time).

Further examples

- Modelling the covariance structure in marginal multivariate count models: Hunting in Bioko Island (Bonat, et. al. 2017).
- Modelling mixed types of outcomes in additive genetic models (Bonat, 2017).
- Multiple Response Variables Regression Models in `R`: The `mcglm` Package (Bonat, 2017).
  1. Gaussian mixed model.
  2. Longitudinal Tweedie model.
  3. Spatial areal data analysis.
  4. Mixed response variables.
  5. Bivariate count along with repeated measures.
  6. Multivariate Tweedie model for spatial areal data.
- For more, see `www.leg.ufpr.br/papercompanions`

Discussion

- Coming soon
  1. Residual analysis (improvements).
  2. Diagnostics (Leverage, DFBETA's and Cook's distance).
  3. Penalized estimating functions (high dimensional data and splines).
  4. Prediction (time, space and space-time).
  5. Special module for Twin data analysis.
  6. Include parameter constraints to fit structural equation models (SEM).
  7. Improve package documentation and output.
  8. Examples gallery.

Main References

Bonat, W. H. ; Jørgensen, B.,*Multivariate Covariance Generalized Linear Models*. Journal of the Royal Statistical Society: Series C (Applied Statistics) 65(5): 649–675, 2016.

Bonat, W. H. ; Kokonendji, C., C. *Flexible Tweedie regression models for continuous data*. Journal of Statistical Computation and Simulation 87(11): 2138–2152, 2017.

Bonat, W. H. ; Jørgensen, B. ; Kokonendji, C., C. ; Hinde, J. ; Demt́rio, C. G. *Extended Poisson-Tweedie: properties and regression models*. Statistical Modelling, 2017.

Bonat, W. H. ; Olivero, J. ; Grande-Vega, M. ; Fárfa, M. A. ; Fa, J. E. *Modelling the covariance structure in marginal multivariate count models*. Journal of Agricultural Biological and Environmental Statistics, 2017.

Bonat, W. H. *Modelling mixed types of outcomes in additive genetic models*. International Journal of Biostatistics, 2017.

Bonat, W. H. *Multiple regression models in R: The mcglm package*. Journal of Statistical Software, 2018.

Contact

- Work in progress …
- Name: Wagner Hugo Bonat
- e-mail: wbonat@ufpr.br
- Webpages
  1. https://cran.r-project.org/web/packages/mcglm
  2. https://github.com/wbonat/mcglm
  3. www.leg.ufpr.br/papercompanions
  4. http://www.leg.ufpr.br/∼wagner
- Thank you !