UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

MCGLM - A Python Library

Jean Carlos Faoot Maia

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Preamble

Statistics

- Statistics has built tools to extract information about data and the system that generated it.
- Randomness is an assumption and one of the biggest challenges.
- Statistical models stand out in dealing with the inherent randomness in data, supporting the statistical analysis.

Introduction to Statistical Models

Statistical Models

Statistical models aim to associate a dependent variable y with a group of independent variables x_i, $i >= 1$.

Statistical Inference provides methods to assess consistent model parameters alongside hypothesis testing of covariates.

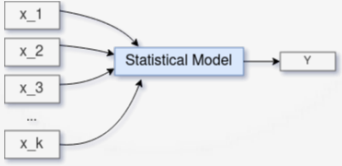


Figura: Standard diagram of statistical models.

Nature may produce distinguish kinds of data.

- Response variables can be independent or dependent events, such as time series or spatial.
- Response variables can assume real values, positive real values, integer values, or bounded values.

- The early statistical models cover independent data exclusively.
- Over the years, many statistical publications aimed to expand the boundaries of models.
- Nowadays, some models can cover almost all kinds of data as an unified solution.

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Statistical Models, History

Linear Regression

Dated to the early nineteenth century, Linear Regression is one of the foremost statistical models.

Statistical Models, History

Linear Regression

Dated to the early nineteenth century, Linear Regression is one of the foremost statistical models.

A vector of observation $\boldsymbol{y}$ with n components are independent realizations of a random variable $\boldsymbol{Y}$; the vector $\boldsymbol{\mu}$ defines its mean parameters. Linear regression combines a systematic and a random part.

Dated to the early nineteenth century, Linear Regression is one of the foremost statistical models.

A vector of observation $\boldsymbol{y}$ with n components are independent realizations of a random variable $\boldsymbol{Y}$; the vector $\boldsymbol{\mu}$ defines its mean parameters. Linear regression combines a systematic and a random part.
The systematic part of the model specifies the vector $\boldsymbol{\mu}$ by employing a linear operation between regression coefficients $\boldsymbol{\beta}$, and a design matrix $\boldsymbol{X}$. The mathematical notation for the systematic part:

$$\boldsymbol{\mu} = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

For the random part, we assume independence and constant variance of errors as a Gaussian distribution with mean 0 and constant variance σ^2.

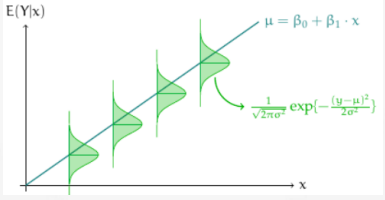


Figura: Graphical explanation of Linear Regression.

Statistical Models, History

Generalized Linear Models

In 1972, Nelder and Wedderburn went a step further in unifying the theory of statistical modeling and, in particular, regression models, publishing their article on Generalized Linear Models (GLM). GLMs universalize linear regression by allowing fitting outcome variables from exponential family.

Statistical Models, History

Generalized Linear Models

In 1972, Nelder and Wedderburn went a step further in unifying the theory of statistical modeling and, in particular, regression models, publishing their article on Generalized Linear Models (GLM). GLMs universalize linear regression by allowing fitting outcome variables from exponential family.

Exponential Family

Exponential Family is a group of distribution models in which either probability density function or probability function assume the form:

$$\mathrm{f}_y(y;\theta,\phi) = \exp\{(\mathrm{y}\theta - \mathrm{b}(\theta))/\mathrm{a}(\phi) + \mathrm{c}(\mathrm{y},\phi)\},$$

for some specific functions $a(\phi)$, $b(\theta)$ and $c(y,\phi)$. The first and second moments are specified as follows:

$$\mathrm{E}(Y) = b'(\theta). \qquad \mathrm{Var}(Y) = b''(\theta)a(\phi).$$

Statistical Models, History

Generalized Linear Models - Exponential Family

Distribution	Support	Cases
Gaussian	Real Numbers	General Symmetric Distributions.
Binomial(Bernoulli)	Bounded Data	Probability/Odds of an event.
Poisson	Integer Positive	Positive Count Distributions.
Gamma	Real Positive Numbers	Positive Asymmetric Distributions.
Inverse Gaussian	Real Positive Numbers	Positive Asymmetric Distributions.

Tabela: Some models of exponential family, data support, and common cases.

Statistical Models, History

Generalized Linear Models

Generalized Linear Models rely on a three-step specification.

- A linear Predictor via design matrix;
- Link Function which maps $\boldsymbol{\eta}$ to $\boldsymbol{\mu}$;
- An exponential family model or a variance function.

Distribution	Canonical link function	Variance function
Binomial	Logit	$\mu(\mu - 1)$
Normal	Identity	1
Poisson	Log	μ
Gamma	Reciprocal	μ^2
Inverse Gaussian	Reciprocal^2	μ^3

Tabela: Candidate distribution models, its usual link and variance functions.

Statistical Models, History

Generalized Linear Models - Estimation

- The Maximum Likelihood Estimation (MLE) leverages the underlying distribution and downstream the optimal regression and dispersion parameters by maximizing the product of general likelihood or sum of its log.
- As long as the log-likelihood function may not have an analytical solution, we apply numerical methods for MLE. By default, GLMs apply the Fisher Scoring method, a.k.a iterative weighted least squares.

Statistical Models, History

Generalized Linear Models - Two moment specification

An unusual GLM specification is via two moments. To develop foundational knowledge for upcoming MCGLM, we deeply dig into GLM in that fashion.

Statistical Models, History

Generalized Linear Models - Two moment specification

An unusual GLM specification is via two moments. To develop foundational knowledge for upcoming MCGLM, we deeply dig into GLM in that fashion.

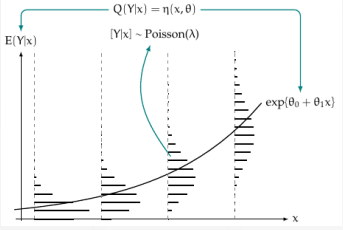
Let Y be a N x 1 response vector, X an N x k design matrix and β a k x 1 regression parameter vector. A mathematical notation for GLM might be specified as:

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}). \\ \mathrm{Var}(\mathbf{Y}) &= \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}(\boldsymbol{\tau}_0 \boldsymbol{I})\mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}. \end{aligned}$$

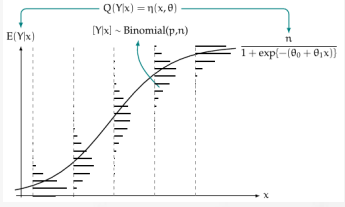
where g is the link function, $V(\mu; p) = diag(\vartheta(\mu; p))$, is a diagonal matrix whose main entries are given by the variance function $\vartheta(; p)$ applied elementwise to the vector μ. The I denotes the N x N identity, whereas τ_0 and p are the dispersion and power parameters.

Statistical Models, History

Generalized Linear Models - Poisson and Binomial Models.



(a) A Poisson regression model.



(b) A Binomial regression model.

Figura: Two examples of GLM realizations.

Statistical Models, History

Quasi-likelihood

In 1974, one of the authors for GLM, Wedderburn, wrote an iconic paper "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method". Wedderburn proposed a near likelihood estimation, the quasi-likelihood, which doesn't rely on a distribution model. Yet asymptotically, the convergence is guaranteed.

Most of the time, it is unreasonable to assume statisticians know the probability model upfront. Hence, the quasi-likelihood function gained attention for upcoming endeavors.

Quasi-score is an estimating equation for quasi-likelihood functions.

Generalized Estimating Equations (GEE) take advantage of quasi-likelihood functions.

Statistical Models, History

Generalized Estimating Equations

In 1986, Liang and Zeger published the paper "Longitudinal data analysis using generalized linear models", presenting the Generalized Estimating Equation (GEE), which universalizes the GLMs by allowing longitudinal data analysis.

For example, the severity of respiratory disease along with children's nutritional status, age, and family income might be observed once every three months for 18 months. The dependence of the outcome variable, severity of disease, on the covariates is of interest.

This paper introduces estimating equations that give consistent estimates of the regression parameters and of their variances under weak assumptions about the joint distribution. The dispersion parameters remain nuisance.

A correlation matrix establishes the dependence between components of the response variable analyzed.

A list of usual dependences structures for GEEs: independent, autoregressive, exchangeable, unstructured, stationary-M, M-dependent or non-stationary.

Therefore, A GEE has four components:

- Linear Predictor;
- Link function;
- Variance function;
- Correlation Matrix(Dependence Structure).

Copulas

- Copulas can calculate any joint distribution with its marginals, alongside the vital statistical properties of these family distributions.
- Copulas can model the dependence of several random variables. Sklar (1959).
- Estimation by either the classical Maximum Likelihood or a Bayesian-based inference.

Statistical Models, History

Copulas and Mixed Models

Copulas

- Copulas can calculate any joint distribution with its marginals, alongside the vital statistical properties of these family distributions.
- Copulas can model the dependence of several random variables. Sklar (1959).
- Estimation by either the classical Maximum Likelihood or a Bayesian-based inference.

Random Effect Models

- It was introduced by Ronald Fisher back in 1950s to study dependent data. Two level of random components: Fixed Effect and Random Effects.
- Breslow and Clayton (1993) penalized quasi-likelihood.
- Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Statistical Models, History

MCGLM

MCGLM, 2015. Bonat, Jørgensen; A brand new family of Statistical Models: Multivariate Covariance Generalized Linear Models.

MCGLM universalizes the GLM by allowing multivariate and non-independent fitting, such as longitudinal and spatial dependence. This dependence is specified by means of dependence matrices and covariance link functions.

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Project Goals

- This project aims to implement mcglm in Python. Thus far, only R users could access the algorithm.
- The main Python library for statistical models, the statsmodels, implements GEE, Mixed Models and Copulas.
- This brand new-Python library inherits the statsmodels basic methods and interface, delivering a new API that the library itself can wrap.
- The mcglm library leverages numpy, scipy, statsmodels and csr_matrix;
- The numpy uses BLAS and LAPACK, such as R packages.
- The mcglm library implements some dependencies matrices as moving averages, mixed models, and the independent case through methods: mc_ma(), mc_mixed() and mc_id().

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Multivariate Covariance Generalized Linear Models

Mathematical Notation

Let $\mathbf{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ be a response matrix, and $\mathbf{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ denote the corresponding matrix of expected values. Let $\boldsymbol{\Sigma}_r$ the $N \times N$ variance-covariance matrix of the outcome r, for $r = 1, \ldots, R$. Similarly, let $\boldsymbol{\Sigma}_b$ be a $R \times R$ correlation matrix inter responses. Let $\boldsymbol{X}_r$ denote an $N \times k_r$ design matrix, $\boldsymbol{\beta}_r$ a $k_r \times 1$ regression parameter vector, and $\mathrm{g_r}(\cdot)$ the link function of the outcome r. The two-moment specification of the MCGLM model goes as follow:

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

where the C leverages the generalized Kronecker product $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \ldots, \tilde{\boldsymbol{\Sigma}}_R^T)$. The matrix $\tilde{\boldsymbol{\Sigma}}_r$ denotes a low triangular Cholesky decomposition of $\boldsymbol{\Sigma}_r$. The operator Bdiag denotes a block diagonal matrix and $\boldsymbol{I}$ denotes an $N \times N$ identity matrix.

Multivariate Covariance Generalized Linear Models

Mathematical Notation

The matrix $\boldsymbol{\Sigma}_r$ is defined by:

$$\boldsymbol{\Sigma}_r = \mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}}(\boldsymbol{\Omega}(\boldsymbol{\tau}_r))\mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}},$$

where $\mathrm{V}(\boldsymbol{\mu}_r; p_r) = \mathrm{diag}(\vartheta(\boldsymbol{\mu}_r; p_r))$ is a diagonal matrix, whose main entries are the variance function applied on the expected values $\boldsymbol{\mu_r}$. Each variance function establishes its unique marginal distributions on MCGLM. Furthermore, MCGLM leverages linear matrix predictor with covariance link function for dispersion matrix definition:

$$\boldsymbol{\Omega}(\boldsymbol{\tau}_r) = h^{-1}(\tau_{r0}Z_{r0} + \cdots + \tau_{rD}Z_{rD}), \tag{1}$$

where $\mathrm{h}(\cdot)$ is the covariance link function and matrices Z_r specifies the dependence inner response. The paper develops an intuition about structures that can be fitted through linear matrix predictor.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

The five-step specification of MCGLM:

- Linear Predictor;
- Link function;
- Variance function;
- Z matrices for the dependence specification;
- Covariance link function.

Multivariate Covariance Generalized Linear Models

Estimation and inference

- MCGLM implements the second moment assumptions, grounded on two moments for the estimation: the mean and the variance.
- There are two optimization algorithms and two estimating equations.
- The estimation produces two groups of parameters: regression and dispersion.

Multivariate Covariance Generalized Linear Models

Estimation and inference

The MCGLM fits via estimating equations through second-moment assumptions. Therefore, the learning process retrieves two groups of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_R^\top)^\top$ e $\boldsymbol{\lambda} = (\rho_1, \ldots, \rho_{R(R-1)/2}, p_1, \ldots, p_R, \boldsymbol{\tau}_1^\top, \ldots, \boldsymbol{\tau}_R^\top)^\top$ denotes $K \times 1$ e $Q \times 1$ vectors with the regression and dispersion parameters.
Let $\mathcal{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_R^\top)^\top$ and $\mathcal{M} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_R^\top)^\top$ be stacked vectors of outcome matrix $\mathbf{Y}_{N\times R}$ and expected values $\mathbf{M}_{N\times R}$, respectively. For regression parameters, the function `Quasi-score`.

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M}),$$

where $\boldsymbol{D} = \nabla_{\boldsymbol{\beta}}\mathcal{M}$ is a matrix $NR \times K$, and $\nabla_{\boldsymbol{\beta}}$ the gradient operator. Matrices *sensitivity* and *variability* $K \times K$ are $\psi_{\boldsymbol{\beta}}$ described by:

$$\mathrm{S}_{\boldsymbol{\beta}} = \mathrm{E}(\nabla_{\boldsymbol{\beta}}\psi_{\boldsymbol{\beta}}) = -\boldsymbol{D}^\top \boldsymbol{C}^{-1}\boldsymbol{D} \quad \text{e} \quad \mathrm{V}_{\boldsymbol{\beta}} = \mathrm{Var}(\psi_{\boldsymbol{\beta}}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}\boldsymbol{D}.$$

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Multivariate Covariance Generalized Linear Models

Estimation and inference

The `Pearson Estimating Equation`, defined by components:

$$\psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}(W_{\boldsymbol{\lambda}_i}(\boldsymbol{r}^\top \boldsymbol{r} - \boldsymbol{C})) \quad \text{, para} \quad i = 1, \ldots, Q,$$

where $W_{\boldsymbol{\lambda}_i} = -\partial \boldsymbol{C}^{-1}/\partial \boldsymbol{\lambda}_i$ and $\boldsymbol{r} = \mathcal{Y} - \mathcal{M}$ were adapted to dispersion parameters. The entry (i, j) for $Q \times Q$ of sensitivity matrix of $\psi_{\boldsymbol{\lambda}}$ is given by:

$$\mathrm{S}_{\boldsymbol{\lambda}_{ij}} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\lambda}_i}\psi_{\boldsymbol{\lambda}_j}\right) = -\mathrm{tr}\left(W_{\boldsymbol{\lambda}_i}\boldsymbol{C}W_{\boldsymbol{\lambda}_j}\boldsymbol{C}\right).$$

The entry (i, j) of $Q \times Q$ variability matrix $\psi_{\boldsymbol{\lambda}}$ is given by:

$$\mathrm{V}_{\boldsymbol{\lambda}_{ij}} = \mathrm{Cov}(\psi_{\boldsymbol{\lambda}_i}, \psi_{\boldsymbol{\lambda}_j}) = 2\mathrm{tr}(W_{\boldsymbol{\lambda}_i}\boldsymbol{C}W_{\boldsymbol{\lambda}_j}\boldsymbol{C}) + \sum_{l=1}^{NR} k_l^{(4)}(W_{\boldsymbol{\lambda}_i})_{ll}(W_{\boldsymbol{\lambda}_j})_{ll},$$

where $k_l^{(4)}$ denotes the fourth cumulant of $\mathcal{Y}_l$.

Multivariate Covariance Generalized Linear Models

Estimation and inference

The model implements the algorithm modified *Chaser* to solve the system equations $\psi_{\boldsymbol{\beta}} = \mathbf{0}$ and $\psi_{\boldsymbol{\lambda}} = \mathbf{0}$, defined by:

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha\mathrm{S}_{\boldsymbol{\lambda}}^{-1}\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}). \end{aligned} \tag{2}$$

The term *tuning* controls the second-moment fitting and adjustment of the second-moment $\boldsymbol{\lambda}$.

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\top}, \hat{\boldsymbol{\lambda}}^{\top})^{\top}$ a estimator for $\boldsymbol{\theta}$, the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is:

$$\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta}, \mathrm{J}_{\boldsymbol{\theta}}^{-1}),$$

where $\mathrm{J}_{\boldsymbol{\theta}}^{-1}$ is the inverse of Godambe matrix,

$$\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-T},$$

where $\mathrm{S}_{\boldsymbol{\theta}}^{-T} = (\mathrm{S}_{\boldsymbol{\theta}}^{-1})^{T}$.

Multivariate Covariance Generalized Linear Models

Model components

The link function $g(\cdot)$ are usual picks of GLM. To cite a few examples: `logit`, `log` and `identity`.

The covariance link function $h(\cdot)$ are usual picks of covariance models. To cite a few examples: `identity`, `inverse` and `exponential-matrix`.

A user incites dependency onto the model through dependence matrices `Z`. Many of the classical statistical models are replicable by setting `Z` matrices. To cite a few, `mixed models`, `moving averages`, and `compound symmetry`. For in-deep details, see the base article.

Multivariate Covariance Generalized Linear Models

Model components

The variance function is fundamental to the MCGLM, as it defines the marginal distribution of a response variable.

- *Power* specialized in handling continuous data and defines the `Tweedie` family of distribution models. This family has its emblematic cases: *Gaussian* (p = 0), *Gamma* (p = 2) and *Inverse Gaussian* (p = 3).
- *Extended binomial* is a common choice for analyzing limited data.
- *Poisson-Tweedie* is flexible to capture notable models, such as: *Hermite* (p = 0), *Neyman Type A* (p = 1), *Negative Binomial* (p = 2) and *Gaussian Poisson-inverse* (p = 3).

Function name	Formula
`power/Tweedie`	μ^p
`binomial`	$\mu^p(1-\mu)^p$
`Poisson-Tweedie`	$\mu + \mu^p$

Tabela: Table with candidate variance functions

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Multivariate Covariance Generalized Linear Models

Matrix Linear Predictor - Examples

The compound symmetry or exchangeable structure.

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Unstructured.

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \tau_2 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + \tau_3 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Moving Average, 2-window.

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \tau_2 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Multivariate Covariance Generalized Linear Models

Matrix Linear Predictor - Feasible Models

- linear regression.
- quasi likelihood models.
- double generalized linear models.
- linear mixed model.
- moving average models.
- exchangeable or compound symmetry.
- 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.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Multivariate Covariance Generalized Linear Models

Measurements of goodness-of-fit and model comparison

Gaussian Pseudo Log Likelihood (plogLik):

$$\text{plogLik}(\boldsymbol{\theta}) = -\frac{NR}{2}\log(2\pi) - \frac{1}{2}\log|\hat{\boldsymbol{C}}| - (\mathcal{Y} - \hat{\mathcal{M}})^{\top}\hat{\boldsymbol{C}}^{-1}(\mathcal{Y} - \hat{\mathcal{M}}).$$

The pseudo *Akaike* information criterion pAIC is given by

$$\text{pAIC}(\boldsymbol{\theta}) = 2(P+Q) - 2\text{plogLik}(\boldsymbol{\theta}).$$

The pseudo *Bayesian* information criterion pBIC is given by

$$\text{pBIC}(\boldsymbol{\theta}) = logNR(P+Q) - 2\text{plogLik}(\boldsymbol{\theta}).$$

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Python Implementation

Classes - UML Diagram

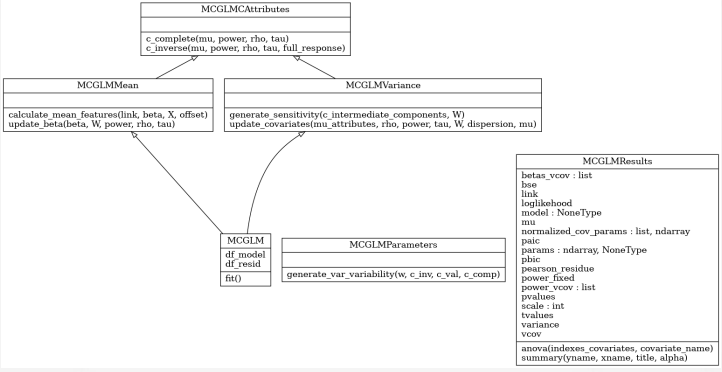


Figura: UML diagram

Python Implementation

Packages connection

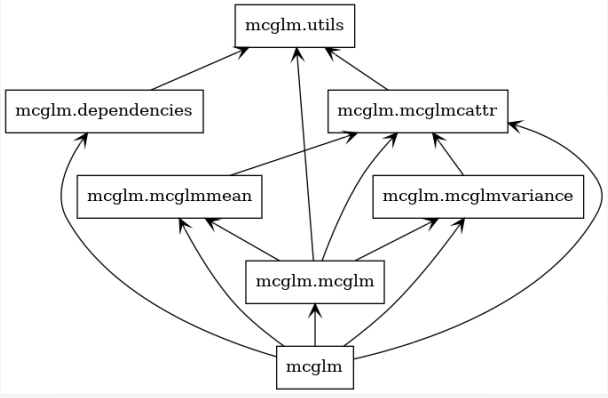


Figura: mcglm classes

Python Implementation

Library usage

```
from mcglm import MCGLM, mc_id

model = MCGLM(endog=y, exog=X, z=[mc_id(X)], link="log", variance="tweedie",
power=2)
mcglmresults = model.fit()
mcglmresults.summary()

model = MCGLM(endog=[y_1, y_2], exog=[X, X], z=[[mc_id(X)], [mc_id(X)]],
link=["log", "log"], variance=["tweedie", "tweedie"], power=[1, 2])
mcglmresults = model.fit()
mcglmresults.summary()
```

Python Implementation

Unit Testing

Unit testing evaluates the accuracy of MCGLM's by tallying with outcomes from specified inputs.

```
========================================= 26 passed in 2.80s ==========================================
jean@pop-os:~/dev/Github/mcglm$ poetry run coverage report
Name                                Stmts   Miss  Cover
-------------------------------------------------
mcglm/__init__.py                       7      0   100%
mcglm/dependencies.py                  43      3    93%
mcglm/mcglm.py                        473     75    84%
mcglm/mcglmcattr.py                   252     21    92%
mcglm/mcglmmean.py                     55      0   100%
mcglm/mcglmvariance.py                 36      0   100%
mcglm/utils.py                         20      0   100%
tests/__init__.py                       0      0   100%
tests/test_dependencies.py             32      0   100%
tests/test_mcglm.py                   283      0   100%
-------------------------------------------------
TOTAL                                1201     99    92%
```

Figura: 92% of unit testing coverage

Python Implementation

Available Functions

- Link Functions: *logit*, *identity*, *log*, *probit*, *cauchy*, *cloglog*, *loglog*, *negative binomial*, *reciprocal*.
- Variance Functions: *constant*, *tweedie*, *binomialP*, *binomialPQ*, *geom_tweedie*, *poisson_tweedie*.
- Covariance Link Functions: identity.

Python Implementation

Variance Function - options

Variance Function	Formula
constant	1
tweedie	μ^p
binomialP	$\mu(1-\mu)$
binomialPQ	$\mu^p(1-\mu)^q$
poisson_tweedie	$\mu + \mu^p$
geom_tweedie	$\mu^2 + \mu^p$

Tabela: Table with Variance function of mcglm

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Simulations with mcglm

Assessing trustworthiness of estimates

- This simulation aims to assess the trustworthiness of mcglm estimates. We evaluate bias, consistency and coverage rate through two variance functions: `BinomialP` and `tweedie` on four different values of `power`.

Simulations with mcglm

Assessing trustworthiness of estimates

- This simulation aims to assess the trustworthiness of mcglm estimates. We evaluate bias, consistency and coverage rate through two variance functions: `BinomialP` and `tweedie` on four different values of `power`.
- Power values selected are 1.01, 1.5, 2, and 3, which emulate different kinds of distributions: near `Poisson`, Compound `Poisson/Gamma`, `Gamma`, and `Inverse Gaussian`.

Simulations with mcglm

Assessing trustworthiness of estimates

- This simulation aims to assess the trustworthiness of mcglm estimates. We evaluate bias, consistency and coverage rate through two variance functions: `BinomialP` and `tweedie` on four different values of `power`.
- Power values selected are 1.01, 1.5, 2, and 3, which emulate different kinds of distributions: near `Poisson`, Compound `Poisson/Gamma`, `Gamma`, and `Inverse Gaussian`.
- For each power, we evaluate three different dispersion parameter. They were generated by relative percentages of the coefficient of variation, set as 15%, 50%, and 80%.

Simulations with mcglm

Assessing trustworthiness of estimates

- This simulation aims to assess the trustworthiness of mcglm estimates. We evaluate bias, consistency and coverage rate through two variance functions: `BinomialP` and `tweedie` on four different values of `power`.
- Power values selected are 1.01, 1.5, 2, and 3, which emulate different kinds of distributions: near `Poisson`, Compound `Poisson/Gamma`, `Gamma`, and `Inverse Gaussian`.
- For each power, we evaluate three different dispersion parameter. They were generated by relative percentages of the coefficient of variation, set as 15%, 50%, and 80%.
- We define two covariates. A sequence between -1 and 1, conditioned by the sample size, and a two-level categorical variable, randomly chosen.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Simulations with mcglm

Assessing trustworthiness of estimates

- This simulation aims to assess the trustworthiness of mcglm estimates. We evaluate bias, consistency and coverage rate through two variance functions: `BinomialP` and `tweedie` on four different values of `power`.
- Power values selected are 1.01, 1.5, 2, and 3, which emulate different kinds of distributions: near `Poisson`, Compound `Poisson/Gamma`, `Gamma`, and `Inverse Gaussian`.
- For each power, we evaluate three different dispersion parameter. They were generated by relative percentages of the coefficient of variation, set as 15%, 50%, and 80%.
- We define two covariates. A sequence between -1 and 1, conditioned by the sample size, and a two-level categorical variable, randomly chosen.
- By emulating those five scenarios, we significantly cover the fitting of independent outcomes due to wide-variety traits of data.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Simulations with mcglm

Assessing trustworthiness of estimates

We leverage four different sample sizes (250; 500; 750; 1000). For each group, we ran 1000 simulations.

Simulations with mcglm

Assessing trustworthiness of estimates

We leverage four different sample sizes (250; 500; 750; 1000). For each group, we ran 1000 simulations.

Variance	Power	Dispersion	Regression	Model
BinomialP	1	1	(0.8; 1; -1.5)	Binomial
tweedie	1.01	(1.5; 15; 40)	(2; 0.8; -1.5)	C.Poisson-Gamma
tweedie	1.5	(0.2; 2; 5.5)	(2; 0.8; -1.5)	C.Poisson-Gamma
tweedie	2	(0.023; 0,25; 0.65)	(2; 0.8; -1.5)	Gamma
tweedie	3	(0.0003; 0.0034; 0.0083)	(2; 0.8; -1.5)	Inverse Gaussian

Tabela: Summary with study simulation proposal.

Simulations with mcglm

Binomial

The figure below presents the approximation to the regression and dispersion parameters via *boxplot* graphs, and the parameter coverage analysis via line graphs.

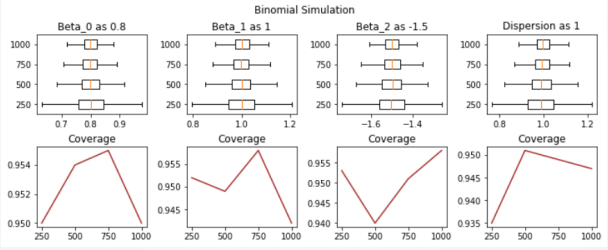


Figura: Report with binomial simulations.

It is possible to verify the consistent and progressive approximation to the real parameters, under the increase of the sample size, and the excellent coverage of the confidence intervals, around 95% for most of the estimates.

Simulations with mcglm

Power set on 1.01

We assess the trustworthiness of the Compound Poisson-Gamma simulation with a power set of 1.01. Dispersion parameters evaluated are 1.5, 15, and 40. From the graphs, it is possible to verify the consistent and progressive akin to the actual parameters, under the sample size increase, and the excellent coverage of the confidence intervals, around 95% for most of the estimates.

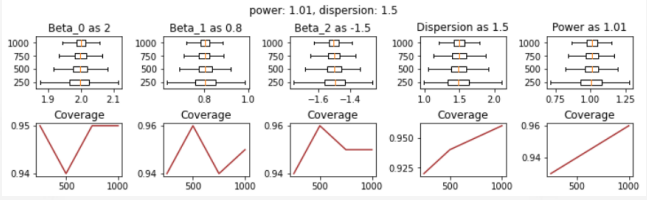


Figura: Report with simulations for dispersion 1.5

Simulations with mcglm

Power set on 1.01

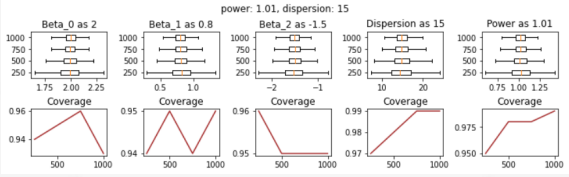


Figura: Report with simulations for dispersion 15

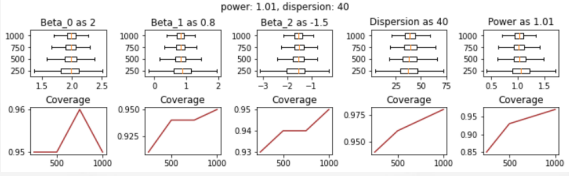


Figura: Report with simulations for dispersion 40

Simulations with mcglm

Power set on 1.5

We assess the trustworthiness of the Compound Poisson-Gamma simulation with a power set of 1.5. The dispersion parameters evaluated are 0.2, 1.8, and 5.2. From the graphs, it is possible to verify the consistent and progressive akin to the actual parameters under the sample size increase and the excellent coverage of the confidence intervals, around 95% for most of the estimates.

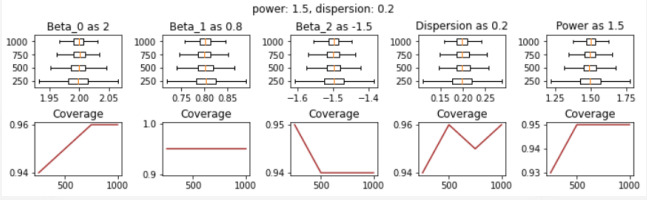


Figura: Report with simulations for dispersion 0.2

Simulations with mcglm

Power set on 1.5

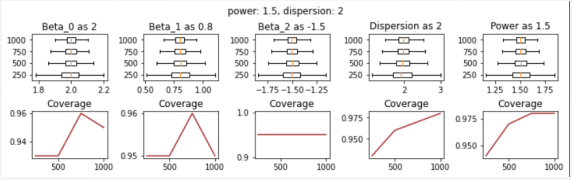


Figura: Report with simulations for dispersion 2

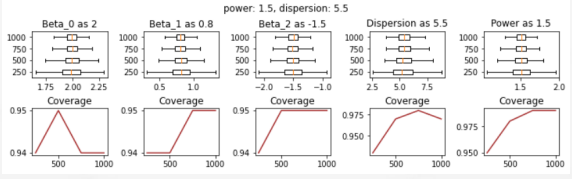


Figura: Report with simulations for dispersion 5.5

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Simulations with mcglm

Power set on 2

We assess the trustworthiness of the Compound Poisson-Gamma simulation with a power set as 2. The dispersion parameters evaluated are 0.025, 0.25, and 0.6. From the graphs, it is possible to verify the consistent and progressive akin to the actual parameters under the sample size increase and the excellent coverage of the confidence intervals, around 95% for most of the estimates.

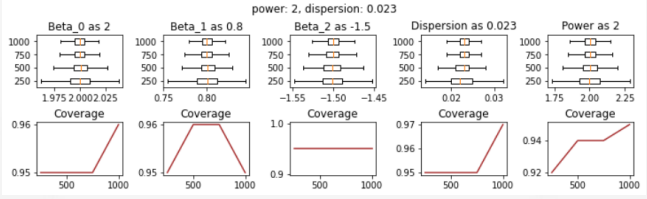


Figura: Report with simulations for dispersion 0.023

Simulations with mcglm

Power set on 2

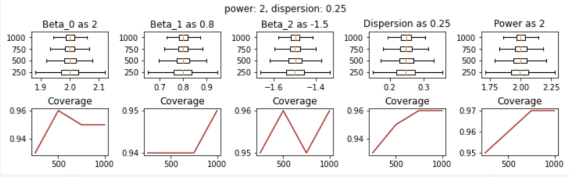


Figura: Report with simulations for dispersion 0.25

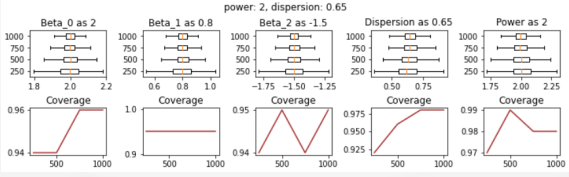


Figura: Report with simulations for dispersion 0.65

Simulations with mcglm

Power set on 3

We assess the trustworthiness of the Compound Poisson-Gamma simulation with a power set as 3. The dispersion parameters evaluated are 0.0003, 0.0034, and 0.0083. From the graphs, it is possible to verify the consistency and progress akin to the actual parameters on the highest variance case.

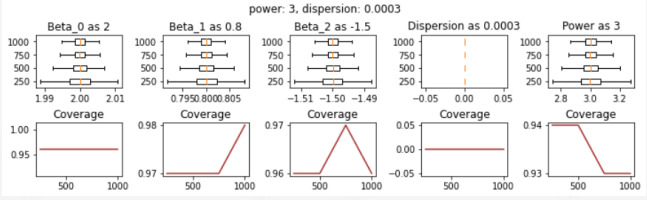


Figura: Report with simulations for dispersion 0.0003

Simulations with mcglm

Power set on 3

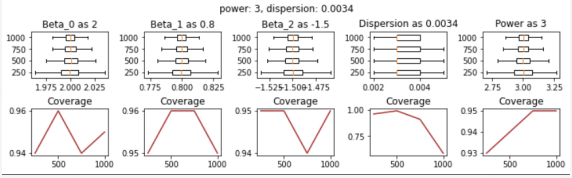


Figura: Report with simulations for dispersion 0.0034

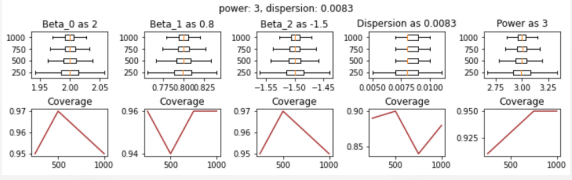


Figura: Report with simulations for dispersion 0.0083

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

To demonstrate the library usage on two actual examples.

`https://github.com/jeancmaia/mcglm/blob/main/nbks/qualification_examples.ipynb`

Outline

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Future work

Wishlist

- To implement more covariance link functions;
- To implement new methods to specify dependency;
- To implement the method "predict";
- To deploy as a new endpoint at the statsmodels library.

Outline

The next steps

Tasks to be accomplished

- Final polishments on the medium story and the article to a scientific journal(journal of statistical software);
- To apply the final suggestions of my board of examiners.