Multivariate regression models for Twin data

Analysis of Twin Data in Health Science · Session IV

Prof. Wagner Hugo Bonat

Ômega Data Science | Online School of Data Science

Motivation

Heredity and variation

- Genetic epidemiology is impelled by three basic questions:
  1. Why isn't everyone the same?
  2. Why are children like their parents?
  3. Why aren't children from the same parents all alike?
- **Main goal: Isolate/Separate sources of variation!**

- Variation is everywhere!

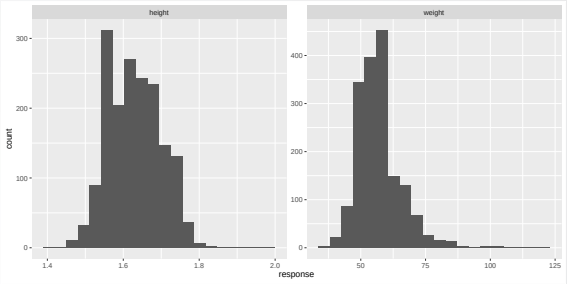


Figura 1. Histogram of height and weight.

Motivating dataset: Anthropometric measures

- Anthropometric measurements (weight and height).
- 861 twin pairs: 327 DZ (dizygotic) and 534 MZ (monozygotic).
- Bivariate continuous traits.
- Covariates: age and group.
- Available as an example in the OpenMx package (Neale, et al., 2016).
- Easy access from the `mglm4twin` package.

Figura 2. Photo by Pixabay.

Motivating dataset: Anthropometric measures

- The dataset

```
library(mglm4twin)
data(anthro)
glimpse(anthro)
```

```
## Rows: 1,722
## Columns: 6
## $ weight    <int> 62, 55, 66, 73, 51, 44, 52, 57, 54, 54, 58, 57, ~
## $ height    <dbl> 1.6499, 1.6299, 1.6599, 1.7000, 1.7300, 1.5698, ~
## $ age       <int> 24, 24, 20, 20, 20, 20, 26, 26, 20, 20, 22, 22, ~
## $ Group     <fct> DZ, DZ, DZ, DZ, DZ, DZ, DZ, DZ, DZ, DZ, DZ, DZ, ~
## $ Twin      <int> 535, 535, 536, 536, 537, 537, 538, 538, 539, 539~
## $ Twin_pair <int> 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, ~
```

Graphing and Quantifying Familial Resemblance

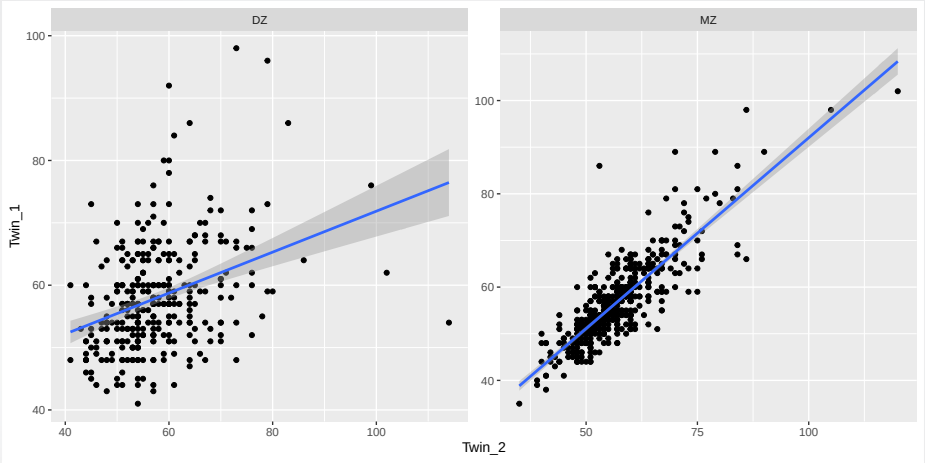


Figura 3. Dispersion diagram by zygosity · Trait weight.

Multiple traits

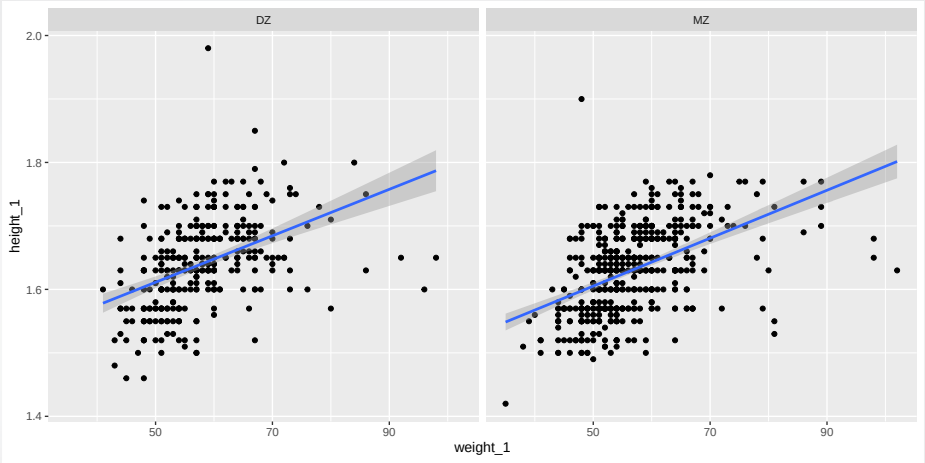


Figura 4. Dispersion diagram by zygosity · Weight vs Height.

Building and Fitting Models

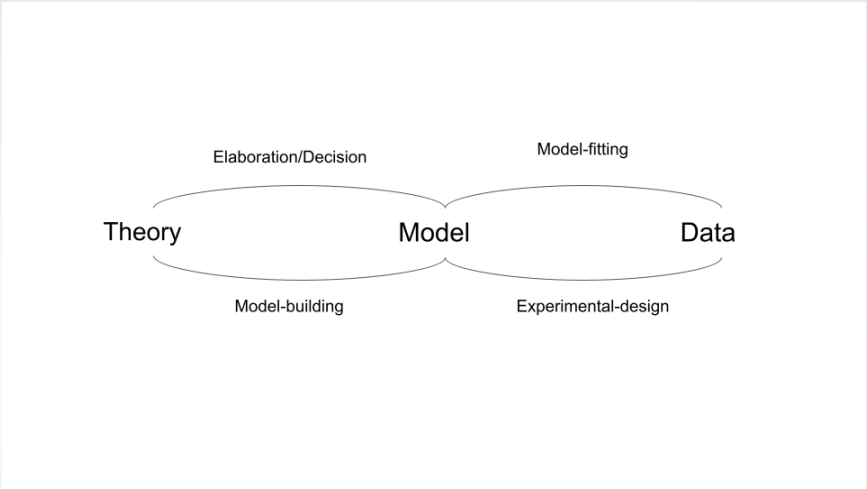


Figura 5. Diagram of the interrelationship between theory, model and empirical observation. Adapted from Neale and Maes (1992).

Challenges for model-building in Twin data analyses

- Decompose sources of variation
  1. Genetic Effects.
  2. Environmental Effects.
  3. Genotype-Environment Interaction.
- Traits types
  1. Binary and binomial data.
  2. Bounded data and continuous proportions.
  3. Under-, equi- and over-dispersed count data.
  4. Semi-continuous data (continuous + mass at zero).
  5. Symmetric and assymetric continuous data.
- Multiple traits of mixed types.

Figura 6. Photo by Magda Ehlers from Pexels.

Importance and statistical approaches

- Multivariate twin and family studies are important tools to:
  1. Determine traits inheritance;
  2. Determine the influence of genetic and environmental effects on traits.
- Statistical challenge:
  1. Model the covariance structure to take into account the genetic and environmental structures induced by the twin and family designs.
- Orthodox approaches:
  1. Structural equation modelling (SEM); 2. Linear mixed models (LMM).
- Main limitations of SEM and LMM:
  1. Both deal only with Gaussian (symmetric) data;
  2. Standard computational implementations are difficult to adapt for the analysis of twin and family data.

Multivariate generalized linear models

Multivariate generalized linear models (mglm): What is it?

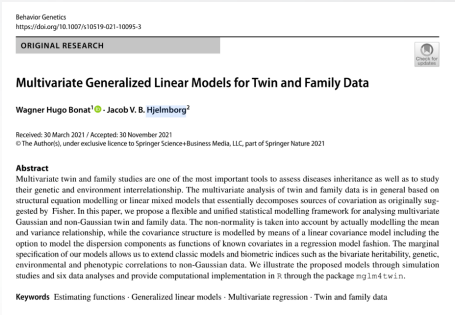
- Flexible statistical modelling framework to deal with multivariate traits.
- Tailored for twin and family data by Bonat and Hjelmborg (2022).
- The mglm approach deals with:
  1. Binary and binomial data;
  2. Bounded data and continuous proportions;
  3. Under-, equi- and over-dispersed count data.
  4. Semi-continuous data (continuous + mass at zero);
  5. Symmetric and assymetric continuous data.
  6. Combination of all the previous mentioned data.
- Estimation and inference based on estimating functions (Bonat and Jorgense, 2016).
- Computational implementation available through the `mglm4twin` package.

Multivariate generalized linear models (mglm): Non-standard features

- Extend standard measures of genetic studies such as:
  1. Bivariate heritability, environmentability and common environmentability;
  2. Genetic, environmental and phenotypic correlations,

to non-Gaussian traits.

- Provide a flexible framework for modelling the dispersion parameters as functions of potential covariates.
- Provide software implementation in `R`.



Behavior Genetics
https://doi.org/10.1007/s10519-021-10095-3

ORIGINAL RESEARCH

Multivariate Generalized Linear Models for Twin and Family Data

Wagner Hugo Bonat[1] · Jacob V. B. Hjelmborg[2]

Received: 30 March 2021 / Accepted: 30 November 2021

Abstract
Multivariate twin and family studies are one of the most important tools to assess diseases inheritance as well as to study their genetic and environment interrelationship. The multivariate analysis of twin and family data is in general based on structural equation modelling or linear mixed models that essentially decomposes sources of covariation as originally suggested by Fisher. In this paper, we propose a flexible and unified statistical modelling framework for analysing multivariate Gaussian and non-Gaussian twin and family data. The non-normality is taken into account by actually modelling the mean and variance relationship, while the covariance structure is modelled by means of a linear covariance model including the option to model the dispersion components as functions of known covariates in a regression model fashion. The marginal specification of our models allows us to extend classic models and biometric indices such as the bivariate heritability, genetic, environmental and phenotypic correlations to non-Gaussian data. We illustrate the proposed models through simulation studies and six data analyses and provide computational implementation in `R` through the package `mglmtwin`.

Keywords Estimating functions · Generalized linear models · Multivariate regression · Twin and family data

Multivariate generalized linear models for twin data

Generalized linear models for twin data

- Let Y_i be a 2×1 random vector of the ith twin pair for $i = 1, \dots, n$.
- Let $\mathrm{x}_i = (x_{i1}, \dots, x_{ik})^\top$ denote a $2 \times k$ design matrix.
- Let β be a $k \times 1$ parameter vector.
- Consider (y_i, x_i), where $y_i's$ are iid realizations of Y_i according to an **unspecified** bivariate distribution, whose expectation and variance are given by

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = g^{-1}(\mathrm{x}_i^\top \beta) \\ \mathrm{var}(Y_i) &= \Sigma_i = \mathrm{V}(\mu_i; p)^{\frac{1}{2}} \Omega \mathrm{V}(\mu_i; p)^{\frac{1}{2}}. \end{aligned} \tag{1}$$

- g some suitable link function.
- $\mathrm{V}(\mu_i; p) = \mathrm{diag}(\vartheta(\mu_i; p))$, where $\vartheta(\mu_i; p)$ describes the mean and variance relation and p is the power parameter (to be estimated).
- Ω is a 2×2 dispersion matrix.

Generalized linear models for twin data

- The models decompose the covariance matrix into two components.

$$\text{var}(Y_i) = \Sigma_i = \text{V}(\mu_i; p)^{\frac{1}{2}} \Omega \text{V}(\mu_i; p)^{\frac{1}{2}}$$

- $\text{V}(\mu_i; p)$ deals with non-Gaussianity.
- Variance/dispersion functions
  1. $\vartheta(\mu; p) = \mu^p$ characterizes the Tweedie distribution deals with continuous and semi-continuous data. Gaussian ($p = 0$), Gamma ($p = 2$) and inverse Gaussian ($p = 3$).
  2. $\vartheta(\mu; p) = \mu + \tau\mu^p$ characterizes the Poisson-Tweedie distribution deals with count data. Neyman-type A ($p = 1$), negative binomial ($p = 2$) and PIG ($p = 3$).
  3. $\vartheta(\mu; p) = \mu^p(1 - \mu)^p$ generalization of binomial variance function deals with binary, binomial and bounded data.
- p is an index that identifies the distribution.
- In practice, we estimate p which works as an automatic model selection.

Generalized linear models for twin data

- Ω models the dependence between twin pair.
- Polygenic ACDE model has the components

$$\mathrm{A} = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}, \quad \mathrm{C} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathrm{D} = \begin{bmatrix} 1 & d \\ d & 1 \end{bmatrix} \quad \text{and} \quad \mathrm{E} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

- Dispersion matrix is modelled by

$$\Omega = \tau_A \mathrm{A} + \tau_C \mathrm{C} + \tau_D \mathrm{D} + \tau_E \mathrm{E}, \tag{2}$$

  where $a = 1$ and $d = 1$ for MZ twins and $a = \frac{1}{2}$ and $d = \frac{1}{4}$ for DZ twins.
- Plugging Eq.(2) in Eq.(1), we have a flexible class of models to deal with twin data.
- But, still only one trait.

Multivariate GLMs for twin data

- Let Y_{ir} be the 2×1 response vector of the rth trait for $r = 1, \ldots, R$.
- Let $\mathrm{x}_{ir} = (x_{ir1}, \ldots, x_{irk})^\top$ be the $2 \times k_r$ design matrix.
- Let β_r be the $k_r \times 1$ parameter vectors.
- Let $\mathrm{Y}_i = (Y_{i1}^\top, \ldots, Y_{iR}^\top)^\top$ denote the $2R \times 1$ stacked vector of response variables.
- Multivariate GLMs for twin data

$$\begin{aligned} \mathrm{E}(\mathrm{Y}_i) &= \mu_i = (g_1^{-1}(\mathrm{x}_{i1}^\top \beta_1), \ldots, g_R^{-1}(\mathrm{x}_{iR}^\top \beta_R)) \\ \mathrm{var}(\mathrm{Y}_i) &= \Sigma_i = \mathrm{V}(\mu_i; p)^{\frac{1}{2}} \Omega \mathrm{V}(\mu_i; p)^{\frac{1}{2}}. \end{aligned} \tag{3}$$

- $\mathrm{V}(\mu_i; p) = \mathrm{diag}(\vartheta_1(\mu_1; p_1), \ldots, \vartheta_R(\mu_R; p_R))$,
- $p = (p_1, \ldots, p_R)$ is an $R \times 1$ vector of power parameters.
- Ω is a $2R \times 2R$ dispersion matrix.

Multivariate GLMs for twin data

- Specification of Ω is crucial.
- Let $\nabla_{rr'}$ denote an $R \times R$ matrix, whose entries $r = r'$ and $r' = r$ are equal to 1 and 0 elsewhere, for $r = 1, \ldots, R$ and $r' \leq r$.

$$\begin{aligned} \Omega &= \tau_{A_{rr'}} \{\nabla_{rr'} \otimes \mathrm{A}\} + \tau_{C_{rr'}} \{\nabla_{rr'} \otimes \mathrm{C}\} \\ &+ \tau_{D_{rr'}} \{\nabla_{rr'} \otimes \mathrm{D}\} + \tau_{E_{rr'}} \{\nabla_{rr'} \otimes \mathrm{E}\}, \end{aligned} \tag{4}$$

where $\tau_{A_{rr'}}, \tau_{C_{rr'}}, \tau_{D_{rr'}}$ and $\tau_{E_{rr'}}$ are dispersion parameters associated with the additive genetic, common environment, dominance genetic and unique environment effects.

Dispersion matrix

- Bivariate case

$$\begin{aligned}\Omega = \quad & \tau_{A_{11}} \begin{bmatrix} \mathrm{A} & 0 \\ 0 & 0 \end{bmatrix} + \tau_{A_{22}} \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{A} \end{bmatrix} + \tau_{A_{12}} \begin{bmatrix} 0 & \mathrm{A} \\ \mathrm{A} & 0 \end{bmatrix} + \\ & \tau_{C_{11}} \begin{bmatrix} \mathrm{C} & 0 \\ 0 & 0 \end{bmatrix} + \tau_{C_{22}} \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{C} \end{bmatrix} + \tau_{C_{12}} \begin{bmatrix} 0 & \mathrm{C} \\ \mathrm{C} & 0 \end{bmatrix} + \\ & \tau_{E_{11}} \begin{bmatrix} \mathrm{E} & 0 \\ 0 & 0 \end{bmatrix} + \tau_{E_{22}} \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{E} \end{bmatrix} + \tau_{E_{12}} \begin{bmatrix} 0 & \mathrm{E} \\ \mathrm{E} & 0 \end{bmatrix}.\end{aligned}$$

- Note: ACDE model is unidentifiable.

Measures of interest

- Broad sense bivariate heritability, common environmentality and environmentality:

$$
\begin{aligned}
h_{rr'} &= \frac{\tau_{A_{rr'}} + \tau_{D_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}}, \\
c_{rr'} &= \frac{\tau_{C_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}} \quad \text{and} \\
e_{rr'} &= \frac{\tau_{E_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}}.
\end{aligned}
$$

Measures of interest

- Genetic, common environmental and environmental correlations:

$$
\begin{aligned}
r_{G_{rr'}} &= \frac{\tau_{A_{rr'}} + \tau_{D_{rr'}}}{\sqrt{\tau_{A_{rr}} + \tau_{D_{rr}}}\sqrt{\tau_{A_{r'r'}} + \tau_{D_{r'r'}}}}, \\
r_{C_{rr'}} &= \frac{\tau_{C_{rr'}}}{\sqrt{\tau_{C_{rr}}}\sqrt{\tau_{C_{r'r'}}}} \quad \text{and} \quad r_{E_{rr'}} = \frac{\tau_{E_{rr'}}}{\sqrt{\tau_{E_{rr}}}\sqrt{\tau_{E_{r'r'}}}}.
\end{aligned}
$$

- Phenotypic correlation

$$
r_{P_{rr'}} = \frac{\tau_{Prr'}}{\sqrt{\tau_{Prr}}\sqrt{\tau_{Pr'r'}}},
$$

where $\tau_{Prr'} = \tau_{Arr'} + \tau_{Crr'} + \tau_{Drr'} + \tau_{Err'}$.

Modelling the dispersion parameters

- Interest to model the component τ_A in a regression model fashion.
- Let z_i be a $(2 \times q)$ design matrix.

$$z_i = \begin{bmatrix} 1 & z_{i11} & \dots & z_{i1q} \\ 1 & z_{i21} & \dots & z_{i2q} \end{bmatrix}.$$

- Let $\tau_A = (\tau_A(0), \tau_A(1), \dots, \tau_A(q))$ denote the $(q \times 1)$ vector of dispersion parameters.
- Dispersion components associated to the additive genetic effect

$$\tau_{A(0)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \circ \mathrm{A} + \tau_{A(1)} \begin{bmatrix} z_{i11} \\ z_{i21} \end{bmatrix} \circ \mathrm{A} + \dots \tau_{A(q)} \begin{bmatrix} z_{i1q} \\ z_{i2q} \end{bmatrix} \circ \mathrm{A}, \tag{5}$$

  where $\circ$ denotes the Hadamard product.
- All dispersion parameters can be modelled as in Eq.(5) and the model remains a linear covariance model.

Estimation and Inference

Estimation and Inference

- Estimation and inference is carried out using an estimating function approach.
- Fitting procedure adapted from Bonat and Jørgensen, 2016.
- Quasi-score estimating functions for regression parameters.
- Pearson estimating functions for power and dispersion parameters.
- Computational implementation in `R` through the `mglm4twin`.
- Available on github `https://github.com/wbonat/mglm4twin`.

Data analyses

Data set 1: Bivariate dichotomous data

- Bronchopulmonary dysplasia (BPD) and respiratory distress syndrome (RDS) on preterm infants.
- Both diseases are lung related and expected to have a genetic component.
- Dataset consists of 200 twin-pairs being 137 DZ and 63 MZ.
- Covariates: birth weight (BW), gestation age (GA) and gender (1: male and 0: female).
- Bivariate ACE model and its special cases as well as their univariate counterparts.

Data set 1: Results

- Multivariate AE model provides the best balance between goodness-of-fit and complexity.

Tabela 1. Pseudo log-likelihood (pll) values along with the pseudo *Akaike* (pAIC), Bayesian ($pBIC$) informatic criterion and degrees of freedom (df) by fitted models - Bivariate binary twin data.

Criterion	Models							
	Univariate				Multivariate			
	ACE	CE	AE	E	ACE	CE	AE	E
pll	−323.03	−334.97	−324.22	−359.99	−311.86	−323.82	−312.90	−347.59
pAIC	686.06	705.94	684.44	751.98	669.72	687.64	665.80	729.18
pBIC	779.75	742.04	748.56	774.86	777.47	781.33	759.49	808.82
df	20	18	18	16	23	20	20	17

Data set 1: Results

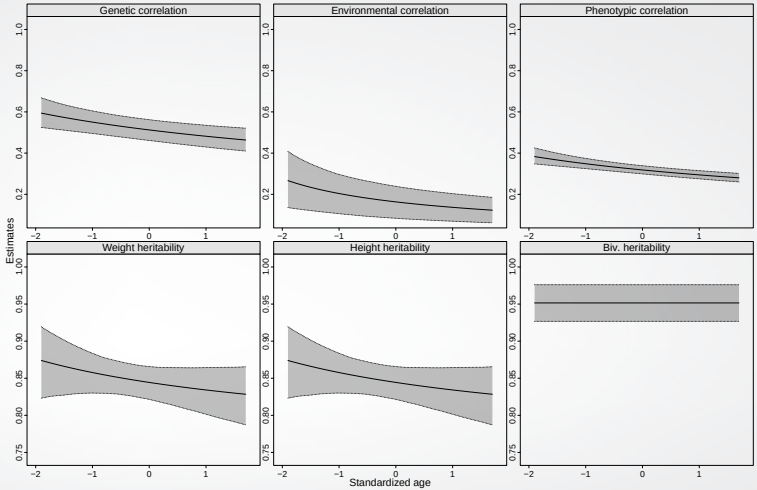
Tabela 2. Parameter estimates and standard errors for some genetic and environment indices of interest - Bivariate binary twin data.

Indices	Traits		
	BPD	RDS	BPD × RDS
Heritability	0.70(0.15)	0.43(0.07)	0.98(0.18)
Environmentality	0.30(0.15)	0.57(0.07)	0.02(0.18)
Genetic correlation	––	––	0.45(0.11)
Environment correlation	––	––	0.01(0.11)

Data set 2: Bivariate continuous data

- Traits of interest: body weight and height measures.
- Data set consists of 861~(327 DZ and 534 MZ) twin-pairs.
- These traits are well-known to be highly genetic determined and correlated.
- Covariate age was used for modelling both the mean and covariance structures.
- Responses and covariates were standardized (mean zero and variance one).

Data set 2: Results



Discussion

Main results

- Flexible multivariate statistical modelling framework to deal with twin data.
- Second-moment assumptions.
- General framework for estimation based on estimating function.
- Efficient algorithms for estimation.
- Asymptotic theory (Godambe Information).
- General software implementation in `R` (`mglm4twin` package).
- Extension of the main genetic measures to non-Gaussian data.
- Future research topics
  1. Weighted estimating functions.
  2. Time to event data.
- Check the webpage for more examples mglm4twin.