

# Extensões de Modelos Mistas

Viviana Giampaoli

December, 2017

I Encontro de Modelagem Estatística-UEM

# Table of contents

- 1 Introduction
- 2 Mixed beta regression model
  - Beta distribution
  - Random intercept model
  - Residual analysis
  - Application I: Ophthalmology
  - Empirical Best Prediction (EBP)
  - Application II: Newborn weight data
  - Application III: Periodontal treatment
- 3 Weibull regression mixed model
  - Application III: Lung cancer
- 4 Conclusions
- 5 Future works
- 6 References

# Introduction

## Challenges

- Specification
- Estimation
- Prediction
- Analysis of residuals

# Mixed beta regression model

Beta regression models are a class of models used frequently to model response variables in the interval  $(0, 1)$ . Although they are used to model clustered and longitudinal data, the prediction of random effects is limited, and residual analysis has not been implemented. A random intercept beta regression model is proposed for the complete analysis of this type of data structure. We propose some type of residuals and formulate a methodology to obtain the best prediction of random effects.

# Application I: Ophthalmology

We analyse the data from a prospective study in ophthalmology reported by [Meyers et al (1992)], in which intraocular gas ( $C_3F_8$ ) was used in complex retinal surgeries to provide an internal tamponade for retinal breaks in the eye. The concentration levels of  $C_3F_8$  used were 25%, 20%, and 15%.

The patients were followed up 3 to 15 times over a 3-month period. Let  $y_{ij}$  be the observed proportion of remaining gas volume relative to the initial volume of gas injected in the eye undergoing surgery for patient  $i$  at time  $j$  on follow-up day  $t_{ij}$ .

## Application II: Newborn weight data

The mixed beta model with logit function was fitted in Zerbeto(2014) to analyse a real database in which the dependent variable is the weight of newborn and, among the explanatory variables there are the amount of weight gain (*taxapeso*) during pregnancy, abortion (*aborto*), sex of the baby (*sexorn*) and gestational age (*igrn*); the clusters of this model were defined according to urinary tract infection status.

## Application III: Periodontal treatment

The analyzed database is composed of clinical examination data from 40 patients who were treated with periodontal treatment and the information was collected at four different times: before the start of treatment, 3,6 and 12 months after treatment. The nature of the response variable anti-oxLDL is to assume positive values restricted by an upper limit that can be approximated by the maximum value observed in the sample.

# Beta distribution

Suppose  $y$  is a response variable that follows a beta distribution. [Ferrari and Cribari Neto (2004)] proposed a parameterisation of its density that is indexed by the mean,  $\mu$ , and also by the precision parameter,  $\phi$ ,

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad (1)$$

where  $0 < y < 1$ ,  $0 < \mu < 1$ , and  $\phi > 0$ . The mean and variance of  $y$  are, respectively, given by  $E(y) = \mu$  and  $\text{Var}(y) = \mu(1-\mu)/(1+\phi)$ .

Alternatively, the beta distribution can be parameterised in terms of the mean,  $\mu$ , and the dispersion parameter,  $\sigma$ ,

$$f(y; \mu, \sigma) = \frac{\Gamma((1 - \sigma^2)/\sigma^2)}{\Gamma(\mu((1 - \sigma^2)/\sigma^2))\Gamma((1 - \mu)((1 - \sigma^2)/\sigma^2))} y^{\mu((1 - \sigma^2)/\sigma^2) - 1} \\ \times (1 - y)^{(1 - \mu)((1 - \sigma^2)/\sigma^2) - 1}, \quad (2)$$

with  $0 < y < 1$ ,  $0 < \mu < 1$ , and  $0 < \sigma < 1$ . In this parameterisation, the mean of  $y$  is  $E(y) = \mu$ , and the variance of  $y$  is  $\text{Var}(y) = \sigma^2 \mu(1 - \mu)$ .

# Random intercept model

Olga Cecilia Usuga Manco-Grupo de Investigación INCAS

Departamento de Ingeniería Industrial-Facultad de Ingeniería-Universidad de Antioquia

Let  $y_{ij}$  be the response value for cluster  $i$  at time  $t_{ij}$ ,  $i = 1, 2, \dots, N$ ,  $j = 1, 2, \dots, n_i$ , and let  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^T$  be the  $n_i$  observations within cluster  $i$ . In the beta random intercept model, it is assumed that the conditional distribution of  $y_{ij}$  given  $\gamma_i = (\gamma_{i1}, \gamma_{i2})^T$  follows a distribution beta with a density determined by Equation (2). Given  $\gamma_i$ , the repeated observations,  $y_{i1}, y_{i2}, \dots, y_{in_i}$ , are independent, and the  $\gamma_{i1}$  e  $\gamma_{i2}$  are independent and identically distributed normal random variables.

We will assume the following model

$$\begin{aligned} y_{ij} \mid \gamma_{i1}, \gamma_{i2} &\stackrel{\text{ind}}{\sim} \text{Be}(\mu_{ij}, \sigma_{ij}), \\ \gamma_{i1} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_1^2), \\ \gamma_{i2} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_2^2), \end{aligned} \tag{3}$$

where  $\lambda_1$  and  $\lambda_2$  are the standard deviation of random effects.

$$\begin{aligned} g_1(\mu_{ij}) &= \eta_{ij1} = \mathbf{x}_{ij1}^T \boldsymbol{\beta}_1 + \gamma_{i1}, \\ g_2(\sigma_{ij}) &= \eta_{ij2} = \mathbf{x}_{ij2}^T \boldsymbol{\beta}_2 + \gamma_{i2}, \end{aligned} \tag{4}$$

where  $\mathbf{x}_{ij1} = (x_{ij11}, x_{ij21}, \dots, x_{ijp_11})^T$  and  $\mathbf{x}_{ij2} = (x_{ij12}, x_{ij22}, \dots, x_{ijp_22})^T$  contain values of explanatory variables,  $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \dots, \beta_{p_11})^T$  and  $\boldsymbol{\beta}_2 = (\beta_{12}, \beta_{22}, \dots, \beta_{p_22})^T$  are the fixed parameter vectors, and  $\gamma_i$  is the random intercept vector. The link functions  $g_1 : (0, 1) \rightarrow \mathbb{R}$  and  $g_2 : (0, 1) \rightarrow \mathbb{R}$  are strictly monotonic and twice differentiable.

## Residual analysis

We define three types of residuals that accomodate the extra source of variability in the proposed model: randomised quantile residuals (see [Dunn and Smyth(1996)]), standardised conditional residuals, standardised marginal residuals.

To assess the overall adequacy of this random intercept beta regression model for the data, we proposed the randomised quantile residual, which is given by

$$r_{qij} = \Phi^{-1} (F(y_{ij}; \hat{\mu}_{ij}, \hat{\sigma}_{ij})) , \quad (5)$$

where  $\Phi(\cdot)$  denotes the cumulative distribution function of the standard normal and  $F(y_{ij}; \hat{\mu}_{ij}, \hat{\sigma}_{ij})$  denotes the cumulative distribution function of  $\text{Be}(\hat{\mu}_{ij}, \hat{\sigma}_{ij})$ .

The standardised conditional residual is defined as

$$r_{cij} = \frac{y_{ij} - \hat{E}(y_{ij} | \gamma_{i1}, \gamma_{i2})}{\sqrt{\widehat{\text{Var}}(y_{ij} | \gamma_{i1}, \gamma_{i2})}}, \quad (6)$$

where  $\hat{E}(y_{ij} | \gamma_{i1}, \gamma_{i2}) = \hat{\mu}_{ij}$  and  $\widehat{\text{Var}}(y_{ij} | \gamma_{i1}, \gamma_{i2}) = \hat{\sigma}_{ij}^2 \hat{\mu}_{ij}(1 - \hat{\mu}_{ij})$ , with  $\hat{\mu}_{ij} = g_1^{-1}(\mathbf{x}_{ij1}^T \hat{\beta}_1 + \hat{\gamma}_{i1})$  and  $\hat{\sigma}_{ij} = g_2^{-1}(\mathbf{x}_{ij2}^T \hat{\beta}_2 + \hat{\gamma}_{i2})$ ;  $\hat{\beta}_1$  and  $\hat{\beta}_2$  denoting the maximum likelihood estimators of  $\beta_1$  and  $\beta_2$ ; and  $\hat{\gamma}_{i1}$  and  $\hat{\gamma}_{i2}$  denoting the BP of  $\gamma_{i1}$  and  $\gamma_{i2}$ , respectively.

Presence of outlying observations  $r_{cij}$  vs observations.

The standardised marginal residual is given by

$$r_{mij} = \frac{y_{ij} - \hat{E}(y_{ij})}{\sqrt{\widehat{\text{Var}}(y_{ij})}}, \quad (7)$$

Linearity of effects  $r_{mij}$  vs explanatory variables

# Application I: Ophthalmology

We consider a random intercept beta regression model with the following specification:

$$\text{logit}(\mu_{ij}) = \beta_{11} + \beta_{21}\log(t_{ij}) + \beta_{31}\log^2(t_{ij}) + \beta_{41}x_{ij} + \gamma_{i1},$$
$$\text{logit}(\sigma_{ij}) = \beta_{12} + \beta_{22}\log(t_{ij}) + \beta_{32}\log^2(t_{ij}) + \beta_{42}x_{ij} + \gamma_{i2},$$

with  $i = 1, \dots, 29$  and  $\gamma_{i1} \sim N(0, \lambda_1)$  and  $\gamma_{i2} \sim N(0, \lambda_2)$ , where  $t_{ij}$  is the time covariate of the days after surgery and  $x_{ij}$  is the covariate of the standardised gas concentration level taking values 1(25%), 0(20%), or -1(15%).

The selected random intercept beta regression model was fit to these data, with the following results:

$$\begin{aligned}\text{logit}(\mu_{ij}) &= \beta_{11} + \beta_{31} \log^2(t_{ij}) + \beta_{41} x_{ij} + \gamma_{i1}, \\ \text{logit}(\sigma_{ij}) &= \beta_{12} + \gamma_{i2}.\end{aligned}\tag{8}$$

	Parameter	Estimate	s.e.	p-value
$\mu$	$\beta_{11}$	1.673	0.120	2.00e-16
	$\beta_{31}$	-0.262	0.015	2.00e-16
	$\beta_{41}$	0.314	0.080	9.31e-05
$\sigma$	$\beta_{12}$	-1.166	0.119	2.00e-16
	$\lambda_1$	1.109	0.093	
$\sigma$	$\lambda_2$	0.322	0.156	

The Figure shows the half-normal probability plot with a simulated envelope for randomised quantiles, standardised conditional and standardised marginal residuals.

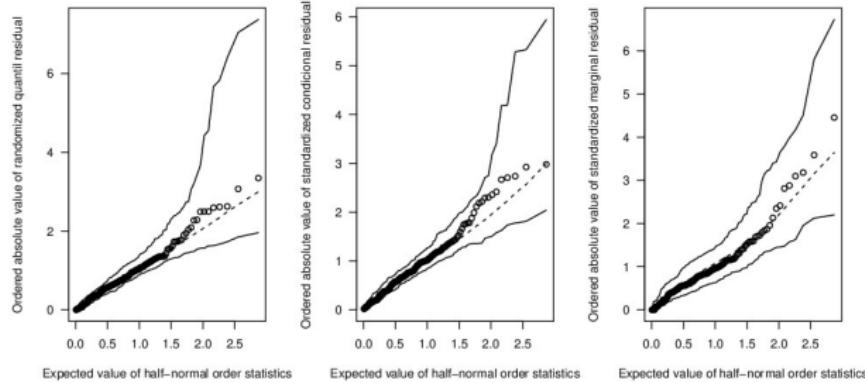


Figure : Residuals

# Empirical Best Prediction (EBP)

Ana Paula Zerbeto-IME-USP

Empirical Bayes predictors  $\hat{\varsigma} = E(\varsigma|y)$ . That minimizes the mean-squared error of prediction (MSEP)  $E(\varsigma' - \varsigma)^2$  for predictor  $\varsigma'$  of  $\varsigma$  over the joint distribution of  $(\varsigma)$  and the responses. It can be calculated as follows

$$\varsigma = \frac{E(\varsigma(\beta, \alpha_i) \cdot \exp(\phi^{-1}S_i(\beta, \alpha_i)))}{E(\exp(\phi^{-1}S_i(\beta, \alpha_i)))}, \quad (9)$$

with  $S_i(\beta, \alpha_i) = \sum_{j=1}^{n_i} [y_{ij} h(\mathbf{x}_{ij}^t \beta + \alpha_i) - (\mathbf{x}_{ij}^t \beta + \alpha_i)]$ .

$\alpha_i = \sigma \xi$  with  $\xi \sim \mathcal{N}(0, 1)$ .

[Jiang and Lahiri (2001)]

It proposed an extension for the prediction of random effects for **individuals who did not belong to the fit data base.**

The mixed beta regression models are used to analyse data with hierarchical structure and that take values in a restricted and known interval like rates, proportions, fractions and others. A specific case of this model is the beta regression models with random intercept exposed in [Usunga (2013)] and they can be written as:

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \alpha_i, \quad (10)$$

, in which:

- $i=1,\dots,q$  is the index of cluster;
- $j=1,\dots,n_i$  is the index of unit in each cluster;
- $y_{ij}$  is the  $j$ th unit of  $i$ th cluster;
- $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$  is a vector with the  $n_i$  observations of  $i$ th cluster;

- $\alpha_1, \dots, \alpha_q$  are random variables i.i.d.  $N(0, \gamma^2)$ ;
- $x_{ij} = (1, x_{ij1}, \dots, x_{ijp'})'$  is a vector of covariates, which are assumed fixed and known;
- $\beta = (\beta_0, \beta_1, \dots, \beta_{p'})'$  is a vector of unknown regression coefficients;
- $\alpha_i$  is the random intercept of  $i$ th cluster;
- $g(\cdot)$  is a strictly monotonic and twice differentiable link function.

Link function	$g(\mu)$
Logit	$\log\left(\frac{\mu}{1-\mu}\right)$
Complementary log-log	$\log(-\log(1-\mu))$
Cauchy	$\tan(\pi(\mu - 0.5))$
Probit	$\Phi^{-1}(\mu)$
Log-log	$-\log(-\log(\mu))$

Table : Usual link functions for the model (10).

The main goals of this work are: to analyse if there are difference in the performance of the empirical best predictor for beta regression models with random intercept for the link functions logit, complementary log-log and Cauchy and; to study the suitability of this link functions in the fit and in the prediction of the model in study for a database in which the dependent variable is the weight of newborn.

Note that the simulation studies for the probit link function was not done, that is due to the logit and probit functions being near linearly related in the interval [0.1, 0.9] and that is why it is not trivial to discriminate between the goodness of fit of this two functions ([McCullagh and Nelder (1989)]).

At the simulation studies, the data were generated according beta model considering different link functions  $g(\cdot)$ , one covariate, true values of parameters  $\beta_0 = 0.5$ ,  $\beta_1 = 1$ ,  $\phi = 3$  and  $n_i = 5$  for all  $i = 1, 2, \dots, q$ . From this database it was randomly selected one proportion ( $p$ ) of the clusters to compose the **fit database (FDB)**, and it was used to calculate the estimates of parameters of model; the remaining of clusters were named **prediction database (PDB)** and, it allows to evaluate the performance of EBP for data that did not were used to estimate the parameters of model. The values of  $q$ ,  $p$  and  $\gamma$  vary according to the scenario.

Simulation scenario	$q$	$p$	$\gamma$
1	10	0.7	1
2	10	0.7	1.5
3	20	0.7	1
4	10	0.9	1

Table : Scenarios of simulation studies.

The topics analysed were: quality of estimation of  $\beta_0$ ,  $\beta_1$ , and  $\phi = 3$  using Square Root of the Mean Square Error (SRMSE) and Relative Bias (RB) of estimates

$$MSE = \frac{\sum_{k=1}^N \sum_{i=1}^q \sum_{j=1}^{n_i} (\hat{\zeta}_{ijk} - \zeta_{ijk})^2}{N \sum_{i=1}^q n_i}$$

$$RB = \frac{1}{N \sum_{i=1}^q n_i} \sum_{k=1}^N \sum_{i=1}^q \sum_{j=1}^{n_i} \left( \frac{\hat{\zeta}_{ijk} - \zeta_{ijk}}{\zeta_{ijk}} \right)$$

performance of EBP of  $\mu_{ij}$  using SRMSE and RB of predicted values.

Scenario	SRMSE( $\hat{\beta}_0$ )			SRMSE( $\hat{\beta}_1$ )			SRMSE( $\hat{\phi}$ )		
	logit	c-log-log	Cauchy	logit	c-log-log	Cauchy	logit	c-log-log	Cauchy
1	0.7251	0.4076	1.0866	0.9877	0.6402	1.9993	4.4152	4.0731	4.6635
2	0.8156	0.5027	1.1914	0.9310	0.7111	1.8262	4.5022	4.3350	4.7429
3	0.5358	0.3208	0.8579	0.7876	0.6204	2.2263	4.4868	4.1830	4.7978
4	0.6598	0.3755	1.0185	0.9485	0.6349	2.0994	4.4779	4.2322	4.1727

Table : SRMSE

	Logit		C-log-log		Cauchy	
Scenario	FDB	PDB	FDB	PDB	FDB	PDB
1	0.1979	0.2190	0.2138	0.2460	0.2072	0.2319
2	0.2568	0.2834	0.2788	0.3176	0.2546	0.2866
3	0.1983	0.2078	0.2173	0.2271	0.2103	0.2216
4	0.1987	0.2137	0.2154	0.2382	0.2093	0.2297

Table : SRMSE of EBP of  $\mu_{ij}$  for FDB and PDB.

	Logit		C-log-log		Cauchy	
Scenario	FDB	PDB	FDB	PDB	FDB	PDB
1	0.0512	0.0640	0.0205	0.0462	0.0615	0.0523
2	0.2794	0.3429	0.2686	0.4058	0.2071	0.2257
3	0.0542	0.0483	0.0348	0.0241	0.0674	0.0564
4	0.0525	0.0558	0.0270	0.0356	0.0627	0.0505

Table : RB of EBP of  $\mu_{ij}$  for FDB and PDB.

In majority of the scenarios the logit was the link function with the lower SRMSE.

The results shown that, in general, there are no significant differences in the behavior of EBP for the link functions considered.

Comparing the behavior of EBP in the fit and prediction databases, it is possible to see that logit is also the link function with better results in the sense of to have the lower percentage increase at SRMSE of PDB compared to that observed in FDB.

## Application II: Newborn weight data

$$z_{ij} = \frac{(y_{ij} - a)}{(b - a)} \quad (11)$$

$$g(\mu_{ij}) = \beta_0 + \beta_1 \text{igrn}_{ij} + \beta_2 \text{aborto}_{ij} + \beta_3 \text{sexorn}_{ij} + \beta_4 \text{taxapeso}_{ij} + \alpha_i, \quad (12)$$

with:

- $j$  Baby/Pregnant;
- $i$  urinary tract infection status.

Parameter	Associated variable	Estimate	SE	t value	p-value
$\beta_0$	-	-11.4495	1.411	-8.114	< 0.001
$\beta_1$	igrn	0.2952	0.036	8.285	< 0.001
$\beta_3$	sexorn	0.0948	0.071	1.336	0.185
$\gamma$	-	0.0779			

Table : Estimates of parameters, standard errors, t values and p-values of final model using complementary log-log link function.

Parameter	Associated variable	Estimate	SE	t value	p-value
$\beta_0$	-	-18.092	2.478	-7.301	< 0.001
$\beta_1$	igrn	0.477	0.062	7.709	< 0.001
$\beta_2$	aborto	-0.215	0.137	-1.577	0.187
$\beta_3$	sexorn	0.142	0.133	1.070	0.288
$\beta_4$	taxapeso	0.603	0.339	1.730	0.087
$\gamma$	-	0.115			

Table : Estimates of parameters, standard errors, t values and p-values of final model using logit link function.

Parameter	Associated variable	Estimate	SE	t value	p-value
$\beta_0$	-	-17.053	2.969	-5.743	< 0.001
$\beta_1$	igrn	0.450	0.075	6.015	< 0.001
$\beta_2$	aborto	-0.281	0.158	-1.778	0.079
$\beta_3$	sexorn	0.075	0.152	0.492	0.624
$\beta_4$	taxapeso	0.633	0.395	1.604	0.112
$\gamma$	-	0.022			

**Table :** Estimates of parameters, standard errors, t values and p-values of final model using Cauchy link function.

The values of AIC of final fitted models for the link functions logit, complementary log-log and Cauchy were, respectively, equal to -125.3, -133.6 and -113.6 indicating that the complementary log-log appears to be the most appropriate.

## Application III: Periodontal treatment

The analyzed database is composed of clinical examination data from 40 patients who were treated with periodontal treatment.

- ① *anti-oxLDL* : units of oxidized anti-LDL antibodies;
- ② *sex*: sex of the patient;
- ③ *age* : age (in years);
- ④ *race* : race declared by the patient;
- ⑤ *BMI* : body mass index (BMI);
- ⑥ *HDL* : HDL concentration (mg/dL);
- ⑦ *LDL* : LDL concentration (mg/dL);
- ⑧ *PCR* : c-reactive protein concentration (em mg/mL);
- ⑨ *PB>4mm* : percentage of periodontal pockets with depth greater than 4mm.

In this case we chose to split the data, randomly, into two bases: the fit database and the prediction database. The first one consists of 70% of the patients and the second one for the remaining 30% of the 40 patients. The final model is presented in equation (13).

$$\log \left( \frac{\mu_{ij}}{1 - \mu_{ij}} \right) = \beta_0 + \beta_1 m.3_{ij} + \beta_2 m.6_{ij} + \beta_3 m.12_{ij} + \beta_4 HDL_{ij} + \beta_5 LDL_{ij} + \alpha_i, \quad (13)$$

where:

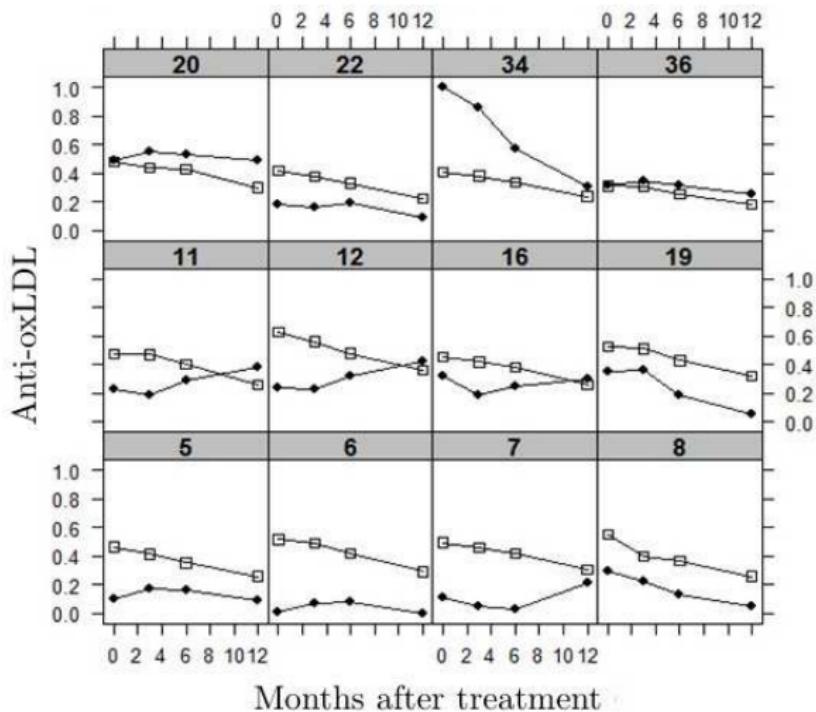
- $i$  denotes the patient;
- $j=0$  indicates that the measurements occurred before the beginning of treatment;
- $j=3$  indicates that the measurements occurred 3 months after the treatment;
- $j=6$  indicates that the measurements occurred 6 months after the treatment;
- $j=12$  indicates that the measurements occurred 12 months after the treatment.

- $y_{ij}$  is the number of antibodies against oxidized LDL of the  $i$ th patient in the  $j$ th month post-treatment;
- $m_{3ij}$  is the indicator variable that assumes value 1 if  $j = 3$  and assumes value 0 otherwise, and in the same way, the variables  $m_{6ij}$  e  $m_{12ij}$ ;
- $\alpha_i$  is the random intercept of  $i$ th patient;
- $\alpha_1, \alpha_2, \dots, \alpha_{40}$  are independently and identically distributed  $N(0, \sigma_\alpha^2)$ ;
- $Y_{ij} | \alpha_i$  are conditionally independent and are distributed as  $Beta(\mu_{ij}, \phi)$  and;
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma_\alpha$  e  $\phi$  are parameters of the model.

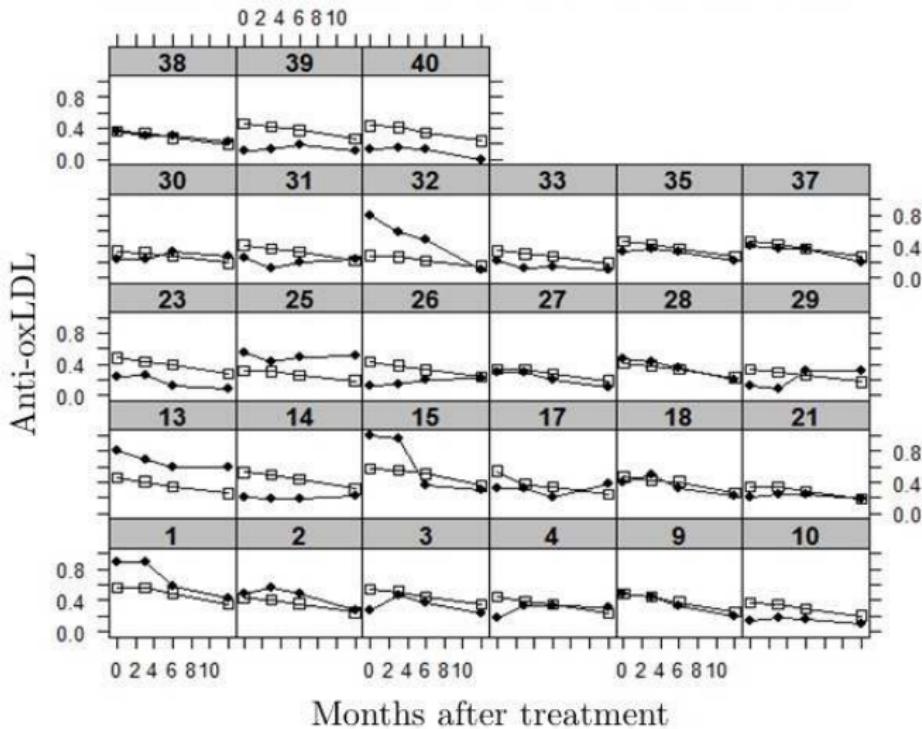
The Table presents the information related to the parameters of the model.

**Table :** Estimates of the parameters, errors, t-values e *p*-values of fitted model for response variable **anti-oxLDL** of periodontitis database

Parameter	Variable	Estimate	Error	t-value	<i>p</i> -value
$\beta_0$	-	-1.989	0.407	-4.885	< 0.001
$\beta_1$	$m_3$	-0.129	0.160	-0.806	0.423
$\beta_2$	$m_6$	-0.359	0.162	-2.219	0.029
$\beta_3$	$m_{12}$	-0.873	0.168	-5.184	< 0.001
$\beta_4$	$HDL$	0.014	0.008	1.797	0.076
$\beta_5$	$LDL$	0.008	0.002	4.098	< 0.001
$\sigma_\alpha$	-	0.536			



**Figure :** Observed values (●) and EBP (□) of **anti-oxLDL** for the patients of PDB of periodontitis database



**Figure :** Observed values (●) and EBP (□) of **anti-oxLDL** for the patients of FDB of periodontitis database

## Application IV: Lung cancer

The data presents the survival time ( $y$ ) in days for 167 patients who were diagnosed with advanced lung cancer from the North Central Cancer Treatment Group, these patients are clustered within 17 medical institutions ([Loprinzi et al. (1994)]). The minimum and the maximum of the response variable are 5 and 1022 days respectively and contains 28.14% of censored observations. The predictors considered were: age in years ( $x_1$ ), sex ( $x_2$ , 0 for female and 1 for male), ECOG performance score ( $x_3$ ), Karnofsky performance score rated by physician ( $x_4$ , 0 for bad and 100 for good), Karnofsky performance score rated by patient ( $x_5$ ), calories consumed at meals ( $x_6$ ) and weight loss in last six months ( $x_7$ ).

# Weibull regression mixed model

*Freddy Hernández Barajas-Escuela de Estadística Facultad de Ciencias  
Universidad Nacional de Colombia sede Medellín  
For WEI3 parameterization ([Hernandez(2013)])*

$$f_Y(y | \mu, \sigma) = \frac{\sigma}{\kappa} \left(\frac{y}{\kappa}\right)^{\sigma-1} \exp\left[-\left(\frac{y}{\kappa}\right)^\sigma\right], \quad \mu > 0, \quad \sigma > 0,$$

$$E(Y) = \mu,$$

$$\text{Var}(Y) = \mu^2 \left[ \Gamma\left(\frac{2}{\sigma} + 1\right) \Gamma^{-2} \left(\frac{1}{\sigma} + 1\right) - 1 \right],$$

$$F(y) = 1 - \exp\left[-\left(\frac{y}{\kappa}\right)^\sigma\right], \quad \text{cumulative distribution function}$$

$$S(y) = \exp\left[-\left(\frac{y}{\kappa}\right)^\sigma\right], \quad \text{survival}$$

$$h(y) = \frac{\sigma}{\kappa} \left(\frac{y}{\kappa}\right)^{\sigma-1}, \quad \text{hazard function}$$

Where  $\kappa = \mu/\Gamma(1/\sigma + 1)$  and  $\Gamma(\cdot)$  is the Gamma function. In this parameterization the  $\mu$  parameter matches with the mean value  $E(Y)$ .

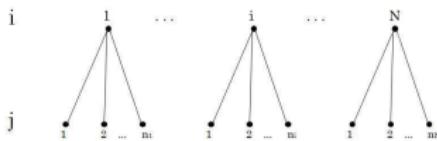


Figure : Hierarchical structure for the model with two levels.

The model structure is as follows:

$$\begin{aligned}
 y_{ij} \mid u_{1i}, u_{2i}, \delta_{ij} &\stackrel{ind}{\sim} WEI3(\mu_{ij}, \sigma_{ij}), \\
 \log(\mu_{ij}) &= \mathbf{X}_{1i,j}^T \boldsymbol{\beta}_1 + u_{1i}, \\
 \log(\sigma_{ij}) &= \mathbf{X}_{2i,j}^T \boldsymbol{\beta}_2 + u_{2i}, \\
 u_{1i} &\stackrel{iid}{\sim} N(0, \tau_1^2), \\
 u_{2i} &\stackrel{iid}{\sim} N(0, \tau_2^2), \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, n_i,
 \end{aligned} \tag{14}$$

The indicator  $\delta_{ij} = 0$  if  $y_{ij}$  is censored and  $\delta_{ij} = 1$  otherwise.

For the  $i$ th cluster, in vectorial form, the vector parameters  $\mu_i$  and  $\sigma_i$  have the following structure:

$$\begin{aligned}\log(\mu_i) &= \mathbf{X}_{1i} \boldsymbol{\beta}_1 + \mathbf{1}_{n_i} u_{1i}, \\ \log(\sigma_i) &= \mathbf{X}_{2i} \boldsymbol{\beta}_2 + \mathbf{1}_{n_i} u_{2i},\end{aligned}$$

where

$$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{in_i})^T,$$

$$\log(\boldsymbol{\mu}_i) = (\log(\mu_{i1}), \log(\mu_{i2}), \dots, \log(\mu_{in_i}))^T,$$

$$\boldsymbol{\sigma}_i = (\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{in_i})^T,$$

$$\log(\boldsymbol{\sigma}_i) = (\log(\sigma_{i1}), \log(\sigma_{i2}), \dots, \log(\sigma_{in_i}))^T,$$

with  $\mathbf{1}_{n_i}$  as 1's vector of order  $n_i \times 1$ .

We initially considered the following regression model for the parameters  $\mu$  and  $\sigma$ :

$$y_{ij} \mid u_{1i}, u_{2i}, \delta_{ij} \stackrel{ind}{\sim} WEI3(\mu_{ij}, \sigma_{ij}),$$

$$\log(\mu_{ij}) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 + \beta_{14}x_4 + \beta_{15}x_5 + \beta_{16}x_6 + \beta_{17}x_7 + u_{1i},$$

$$\log(\sigma_{ij}) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \beta_{23}x_3 + \beta_{24}x_4 + \beta_{25}x_5 + \beta_{26}x_6 + \beta_{27}x_7 + u_{2i},$$

with  $i = 1, 2, \dots, 17$ ,  $j = 1, 2, \dots, n_i$ ,  $u_{1i} \sim N(0, \tau_1^2)$  and  $u_{2i} \sim N(0, \tau_2^2)$ .

$$\begin{aligned} y_{ij} \mid u_{1i}, u_{2i}, \delta_{ij} &\stackrel{ind}{\sim} WEI3(\mu_{ij}, \sigma_{ij}), \\ \log(\mu_{ij}) &= \beta_{10} + \beta_{15}x_5 + u_{1i}, \\ \log(\sigma_{ij}) &= \beta_{20} + \beta_{21}x_1 + u_{2i}. \end{aligned} \tag{15}$$

The proposed model finds jointly the fixed effects and the variance components whereas GAMLSS estimates fixed effects conditional on the estimated covariance components. From this table we can see that the estimated parameters from proposed model and GAMLSS are close and that the standard errors slightly smaller for GAMLSS.

**Table :** Parameter estimates and standard errors (in parentheses) for the lung data using the proposed model and GAMLSS.

Parameter	Proposed model	GAMLSS
$\beta_{10}$	4.783(0.358)	4.755(0.273)
$\beta_{15}$	0.014(0.004)	0.015(0.003)
$\beta_{20}$	1.723(0.539)	1.669(0.403)
$\beta_{21}$	-0.021(0.008)	-0.020(0.006)
$\tau_1$	0.254(0.050)	0.111(n.a.)
$\tau_2$	0.279(0.078)	0.198(n.a.)
$I(\theta)$	-834.580	-822.764

n.a. means not available

# Conclusions

- In the beta model there were differences between the fitted models for each of the three link functions considered, but the predictions calculated by EBP were very similar.
- Multilevel Weibull Model can be used to analyze datasets in which the response variable is related to durations time with or without right censored observations.
- Flexibility.
- Require large computational development.

# Future works

- Determination of the most appropriate residuals to identify additional explanatory variables that contribute significantly to the model.
- Confidence intervals for predicted values.
- Influence analysis.

Acknowledgements: FAPESP, CNPq, CAPES and the Graduate Program in Statistics of the Institute of Mathematics and Statistics, University of São Paulo (IME- USP)

# References I

-  Dunn, P. K., Smyth, G. K., 1996. Randomised quantile residuals. *J. Comput. Graph. Statist.* 5, 236–244.
-  Ferrari, S., & Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*.
-  Hernandez, F.B. (2013). Modelos multiníveis Weibull com efeitos aleatórios. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil.
-  Jiang, J. and Lahiri, P., 2001. Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, 2, 217–243.

## References II

-  Loprinzi, C., Laurie, J., Wieand, H., Krook, J., Novotny, P., Kugler, J., Bartel, J., Law, M., Bateman, M., Klatt, N., 1994. Prospective evaluation of prognostic variables from patient-completed questionnaires. *Journal of Clinical Oncology* 12, 601–607.
-  McCullagh, P., and Nelder, J. A. (1989). Generalized linear models. London, England, Chapman and Hall.
-  Meyers, S., Ambler. J., Tan, M., Werner, J., Huang, S. Variation of perfluoropropane disappearance after vitrectomy. *Retina*, 12, 359–363.
-  Usuga, O.C. (2013). Modelos de regressão beta com efeitos aleatórios normais e não normais para dados longitudinais. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil.

## References III

-  Zerbeto, A.P. (2014). Melhor preditor empírico aplicado aos modelos beta mistos. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil.