Brazilian Journal of Biometrics (2022), **40**, 381–392
doi:10.28951/bjb.v40i4.622

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Marginal logistic regression with a cure fraction in a cluster design: An application in dental traumatology

Enrico Antônio Colosimo,*[,1] Eduardo Fernandes e Silva,*[,2] and Juliana Vilela Bastos[3]

[1]Department of Statistics, Federal University of Minas Gerais, Belo Horizonte, Brazil
[2]Federal University of Minas Gerais, Belo Horizonte, Brazil
[3]Department of Odontology, Federal University of Minas Gerais, Belo Horizonte, Brazil
*Corresponding author. Email: enricoc57@gmail.com;

(Received: June 1,2022; Revised: December 14,2022; Accepted: December 14,2022; Published: December 30,2022)

Abstract

Logistic regression model is the first option to deal with binary outcomes in cross-sectional health studies. However, some conditions, such as the presence of a cure fraction, characterized when an unknown portion of the population is no longer at risk of developing the event of interest, can lead to the non-adequacy of the model. Therefore, the presence of a cure fraction requires an extension in the standard form of the logistic regression model or the use of an alternative one. The present work aims to identify risk factors for the presence of External Inflammatory Root Resorption (EIRR) using a real application. The data set consisted in replanted permanent teeth referred to treatment at the Dental Trauma Clinic of the School of Dentistry from the Federal University of Minas Gerais (DTC-SD-UFMG) after emergency care at the Metropolitan Hospital Odilon Beherns in Belo Horizonte, Brazil. A logistic regression type model is considered to study the association between clinical and radiographic factors and the presence/absence of EIRR, measured radiographically at the first patient appointment at DTC-SD-UFMG. Considering that EIRR is only expected in those cases where the root canal become infected following pulp necrosis, those teeth whose pulp healing is favorable are not at risk of developing EIRR. However, pulpal status usually can only be defined in the long term, such that information is not available at the time of data collection, characterizing the presence of a latent cure fraction. Moreover, in the present sample some patients contributed with more than one replanted tooth, forming clusters of correlated measurements. In the present work we followed the methodology proposed by Hall & Zhang (2004) in which they used an adaption of the EM (expectation-maximization) algorithm, called ES (Expectation-Solution) algorithm combined with GEE (Generalized Estimation Equations) to accommodate the cluster (individual) multivariate response in a logistic cure fraction model.

Keywords: Cluster, healing factor, logistic Regression, ZIB model.

1. Introduction

Cross-sectional studies are frequently used in clinical health research as they are useful for measuring the prevalence of health/disease conditions, understand their determinants and describe features of a population. The present research consisted in a cross-sectional study conducted at the Federal University of Minas Gerais (Belo Horizonte- Brazil) aiming to identify risk factors for the presence of External Inflammatory Root Resorption (EIRR) in permanent teeth reimplanted after traumatic avulsion. Patients received emergency care at the Metropolitan Hospital of Belo Horizonte Odilon Bherens and were referred to sequential treatment at the Dental Trauma Clinic at the School of dentistry of UFMG (DTC-SD-UFMG). Tooth avulsion implies the total displacement of the tooth out of its socket with rupture of the dental pulp neurovascular supply, at the apical foramen, as well as the damage of all periodontal fibers that joints the tooth to the surrounding bone. Replantation of avulsed permanent teeth is the treatment of choice, but the long-term prognosis shows great variability because it is affected by several factors related to the immediate management of the avulsed tooth as well as to the emergency and sequential treatment (Coste *et al.,* 2020). External root resorption is the most serious and frequent sequel after replantation of permanent teeth (ibid.). Two progressive forms have been described being both triggered by severance of the periodontal ligament and damage to the cementoblast layer as a result of extended extra-alveolar maintenance of the avulsed tooth in unsuitable storage conditions prior to replantation (Andersen *et al.,* 1992). Progressive External Inflammatory Root Resorption (EIRR) develops in the presence of root canal infection following pulp necrosis. Therefore, teeth with favorable pulpal healing will not develop EIRR. Considering that pulpal status usually can only be confirmed in the long term, and such information was not available at the time of the outcome measure, performed at the initial visit at the DTC-SD-UFMG, such clinical condition is a latent cure factor. That is, a portion of the study population was not susceptible to develop the outcome, in this case the EIRR.

The logistic regression model is the most common approach to estimate the associations between a binary outcome and exposures measured at one specific time point. However, when an unknown portion of the population is no longer at risk of developing the event of interest, the conventional logistic model became not adequate due to the presence of such latent cure factor, requiring an extension of it or the use of an alternative model. Moreover, in the present sample, many patients contributed with more than one traumatized tooth, characterizing the presence of clusters. It means that the observations (teeth) within each cluster (individual) are correlated and it is necessary to take into account in the statistical inference.

According to Diop *et al.* (2011), the cure fraction problem in binary response can be treated as a Zero-Inflated Binomial (ZIB) model. Hall (2000) described the ZIB model for the adjustment of binary data in cases where there are zero inflated for the binomial distribution and its extension including random effects. Hall & Zhang (2004) presented an alternative approach to deal with ZIB models involving more than one measurement for each individual: flexibilizing the step of maximizing in the EM (Expectation-Maximization) algorithm, including GEE (Generalized Estimation Equations), named the ES (Expectation-Solution) algorithm. In the currrent work we present an extension of the logistic model in order to include cure factor and repeated measures in the same patient. Outline of the paper is as follow. The model for binary response under cure fraction is presented for just one measure for each individual in Section 2. Section 3 extends the model for including clusters of observations. Results of a restricted set of Monte Carlo simulations appear in Section 4 showing small properties of the model estimates. The proposed methodology is illustrated with a real data set from the DTC-SD-UFMG in Section 5. In Section 6 we present some final remarks and future work.

2. Binary response with cure factor: A simple random sample

Let's consider Y as a binary variable and W is a latent binary random variable assuming 1 if the event of interest occurs with probability p_1 and 0 otherwise, with probability $1 - p_1$, which $0 < p_1 < 1$. If $W = 1$, the binary random variable Y is degenerated at 0. Conditional on $W = 0$, the random variable of interest follows a Bernoulli distribution with probability π_1 of success, that is, $Y = 1$, in which $0 < \pi_1 < 1$.

Therefore Y is defined as the following mixture random variable

$$Y = \begin{cases} 0, & \text{with probability} \quad p_1 \\ \text{Bernoulli}(\pi_1), & \text{with probability} \quad (1 - p_1). \end{cases}$$

In section 2.1 we consider the inference in the logistic regression model with a cure factor for a simple random sample of the population, including covariates. As in generalized linear models, link functions are used to connect the mean of the distribution with the linear predictor. According to Yamaguchi *et al.* (1992), one of the main advantages of using a standard mixture model is the possibility of including covariates through a link function in both parts of the model structure. In this way, it makes possible to determine covariates associated to the components of the mixture, which is of practical interest for the researchers.

2.1 Inference in logistic model with cure factor: An EM algorithm approach

Let $\mathbf{Y} = (y_1, ..., y_n)'$ be a random sample of size n. Consider $\mathbf{X}$ and $\mathbf{Z}$ regression structures containing covariates with dimensions $n \times (p+1)$ and $n \times (q+1)$ respectively. Let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors of dimensions $(p+1) \times 1$ and $(q+1) \times 1$, respectively. Considering the logistic function, we have:

$$\pi_1(x) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})} \tag{1}$$

and

$$p_1(z) = \frac{\exp(\mathbf{Z}\boldsymbol{\gamma})}{1 + \exp(\mathbf{Z}\boldsymbol{\gamma})} \tag{2}$$

The observed log-likelihood function is:

$$l_{obs}(\boldsymbol{\beta}, \boldsymbol{\gamma}; Y) = \sum_{i=1}^{n} \left(I_{(0)}(y_i) \log \left[\exp(\mathbf{Z}_i\boldsymbol{\gamma}) + \frac{1}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \right] \right.$$
$$\left. - \log(1 + \exp(\mathbf{Z}_i\boldsymbol{\gamma})) + \left(1 - I_{(0)}(y_i)\right) (\mathbf{X}\boldsymbol{\beta} - \log(1 + \exp(\mathbf{X}_i\boldsymbol{\beta}))) \right), \tag{3}$$

where $I_{(0)}(y_i)$ is a indicator function for $y_i = 0$.

A useful maximization strategy is to use the EM algorithm (Dempster *et al.,* 1977). Let's consider w_i =1 when Y_i takes the value 0 and w_i =0, if Y_i follows a Bernoulli distribution with a probability of success π_1. Considering $W = (w_1, w_2, ..., w_n)^T$ as the missing data, the log-likelihood function for the complete data (Y, W) is:

$$
\begin{aligned}
l_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \gamma, w) &= \log \prod_{i=1}^{n} P(Y_i = \gamma_i, W_i = w_i) \\
&= \sum_{i=1}^{n} [w_i \mathbf{Z}_i \boldsymbol{\gamma} - \log(1 + \exp(\mathbf{Z}_i \boldsymbol{\gamma}))] \\
&+ \sum_{i=1}^{n} (1 - w_i) \left[\left(1 - I_{(0)}(\gamma_i)\right) \mathbf{X}_i \boldsymbol{\beta} - \log(1 + \exp(\mathbf{X}_i \boldsymbol{\beta})) \right] \\
&= l_c(\boldsymbol{\gamma}; w) + l_c(\boldsymbol{\beta}; \gamma, w) \qquad (4)
\end{aligned}
$$

The log-likelihood (4) factored into two parts, one that depends only on $\boldsymbol{\gamma}$ and the other on $\boldsymbol{\beta}$. This factorization, which does not occur in the observed log-likelihood (2) function, facilitates the treatment of more than one measurement for the same individual, which is the main goal of the present paper. The log-likelihood function (4) is a particular case of the ZIB function, described by Hall (2000).

Let $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$. The EM algorithm starts with an initial guess $\boldsymbol{\lambda}^{(0)}$ and then switches between the steps of expectation and maximization. The steps are the following.

1. Start $\boldsymbol{\lambda}^{(0)}$ and $j = 0$
2. **Step E**: estimate w_i as the conditional average. That is $\hat{w}_i^{(j)} = E\left[w_i \mid \gamma_i, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right]$.

   Expectations is obtained by using Bayes' theorem. *j-th* element of $\hat{w}_i^{(j)}$ is given by:

$$
\hat{w}_i^{(j)} = I_0(\gamma_i) \left[1 + \exp\left(-\mathbf{Z}_i \boldsymbol{\gamma}^{(j)}\right) \left(\frac{1}{1 + \exp\left(\mathbf{X}_i \boldsymbol{\beta}^{(j)}\right)} \right) \right]^{-1}
$$

3. **Step M to $\boldsymbol{\gamma}$**: the estimate of $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(j+1)}$ is obtained by maximizing $l_c(\boldsymbol{\gamma}; w^{(j)})$ in (4). Hall (ibid.) defined that it is a binomial regression with weights and the vector response $W^{(j)}$.
4. **Step M to $\boldsymbol{\beta}$**: the estimate of $\boldsymbol{\beta} = \boldsymbol{\beta}^{(j+1)}$ is obtained by maximizing $l_c(\boldsymbol{\beta}; w^{(j)})$ in 4. Hall (ibid.) defined the expression as a logistic regression with weights $(1 - w_i^{(j)})$, $i : 1, ..., n$ and the response vector Y_i
5. Repeat the **E** and **M** steps until convergence is achieved.

The convergence criterion used to stop the iterative algorithm EM is:

$$
\max_r \left(\frac{|\lambda_r^{j+1} - \lambda_r^j|}{|\lambda_r^j| + \delta_1} \right) < \delta_2,
$$

where δ_1 and δ_2 are pre-specified constants and $max_r A_r$ represents the highest value of A_r for $r = 1, \ldots, p$.

The usual information matrix proposed by Louis (1982) has a difficult derivation because it requires finding the distribution of the latent variable conditioned to the observed data. As an alternative, we used the information matrix proposed by Sy & Taylor (2001), considering by $l_{obs}(.)$ and l_c the likelihood functions based on the observed and complete data, respectively. That is

$$
-\frac{\partial^2 l_{obs}}{\partial \gamma_s \partial \gamma_b} = -\sum_{i=1}^{n} \mathbf{Z}_{is} \left(\frac{\partial w_i^*}{\partial \gamma_b} - \frac{\partial p_{1i}}{\partial \gamma_b} \right),
$$

$$-\frac{\partial^2 l_{obs}}{\partial\beta_r\partial\beta_l} = -\sum_{i=1}^{n}\mathbf{X}_{ir}\left(\frac{\partial\left(1-w_i^*\right)}{\partial\beta_l}\right)\left[y_i-\pi_{1i}\right],$$

$$-\frac{\partial^2 l_{obs}}{\partial\gamma_s\partial\beta a} = -\sum_{i=1}^{n}\mathbf{Z}_{is}\left(\frac{\partial w_i^*}{\partial\beta_a}\right),$$

where w_i^* represents the value of w_i evalueted at $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Wald statistic is used for inference on β and γ.

In order to be identifiable the mixture model needs to follow some regularity conditions. The following subsection presents some regularity conditions presented by Diop *et al.* (2011).

2.2 Regularity and identifiability conditions

A common problem with the presence of the cure factor is the impossibility of differentiating between susceptible and cured individuals when the observed value of Y is 0. Diop *et al.* (ibid.) present four regularity conditions necessary to ensure identifiability and asymptotic results. The condition that differs from traditional regularity ones for the logistic regression models, is the following. It is necessary the existence of a continuous covariate V that is in $\mathbf{X}$, but not in $\mathbf{Z}$. If β_V and γ_V denote the coefficients of V in the two linear predictors, then $\beta_V \neq 0$ and $\gamma_V = 0$. Likewise, we can assume that V is in $\mathbf{Z}$ and not in $\mathbf{X}$. According to the authors, this condition imposes restrictions on the choice of predictors because they cannot be equal, but it prevents the parameters from being interchangeable.

Follmann & Lambert (1991) showed that finite mixtures of logistic regressions are identifiable as long as the number of combinations of observed values of the covariates is large. According to Diop *et al.* (2011), the established condition is enough to achieve this condition and guarantee identifiability. Kelley & Anderson (2008) showed that if the focus is the comparison of two groups, it is possible to include a continuous covariate in one and only one linear predictor, even if not significant, that the model in question becomes identifiable.

3. Binary response with cure factor: a cluster sample

Situations with more than one measurement on the same individual may occur in practice, especially in odontology. It can be said that each individual represents a cluster of correlated measurements. Ignoring this fact can lead to wrong conclusions, as the standard errors of the estimates tend to be poorly estimated. In addition, a standard analysis would discard information, as it would be necessary to randomly select an individual measurement in order to use the methodology presented in Section 2.

Hall (2000) presented the mixed zero-inflated models as a possible solution. However, the direct interpretation is not at the population level, which is not of practical interest to the researcher. Hall & Zhang (2004) presented two solutions: the first is to change the **M** step of the **EM** algorithm to accommodate more than one measurement on the same individual, including equations very similar to the GEE, called the **ES** algorithm. The second option proposed to change the GEE equations to accommodate zero inflation. According to the authors, from simulations studies performed by them, the first solution is more efficient and was adopted in this work.

3.1 ES algoritm

Let's consider K clusters and $\boldsymbol{\gamma_i}$ the response vector of dimension $n_i \times 1$ for the *i-th* individual or cluster, $i = 1, ..., K$ and γ_{ij} is the jth; $j = 1, \ldots, n_i$, response associated to the ith individual. Let's considered that Y_{ij} follows a degenerate distribution at 0 with probability p_{ij} if cure occurred and a Bernoulli distribution with probability $1 - p_{ij}$, in which $0 < p_{ij} < 1$, with $j = 1, ..., n_i$.

Consider $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ and $\mathbf{W}_i = (w_{i1}, ..., w_{in_i})$, where $w_{ij} = 0$, if the random variable of interest follows the Bernoulli distribution, and $w_{ij} = 1$, when $Y_{ij} = 0$, $i = 1, ..., K$, $j = 1, ..., n_i$. Let's considered the same link functions (1) and (2). The log likelihood function for the complete data $(\mathbf{Y}, \mathbf{W})$ is the following

$$l_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \gamma, w) = \sum_{i,j} \left[w_{ij} \mathbf{Z}_{ij} \boldsymbol{\gamma}' - \log\left(1 + \exp\left(\mathbf{Z}_{ij}\boldsymbol{\gamma}'\right)\right)\right] + \sum_{i,j} (1 - w_{ij}) \left[\left(1 - I_{(0)}(\gamma_{ij})\right) \mathbf{X}_{ij}\boldsymbol{\beta}' - \log\left(1 + \exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}'\right)\right)\right]. \quad (5)$$

Following the same steps as in Section 2, Hall & Zhang (2004) changed the form of the M step of the EM algorithm, by an S step, proposing the ES algorithm. E step consists of estimating w_{ij} in the *h-th* iteraction by

$$\hat{w_{ij}}^{(h)} = I_0(\gamma_{ij}) \left[1 + \exp\left(-\mathbf{Z}_i \boldsymbol{\gamma}'^{(h)}\right)\left(\frac{1}{1 + \exp\left(\mathbf{X}_i \boldsymbol{\beta}'^{(h)}\right)}\right)\right]^{-1} \quad (6)$$

$I_0(\gamma_{ij})$ is the indicator for $\gamma_{ij} = 0$.

Hall & Zhang (ibid.) described the S step by using the following GEE for the maximization of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{p_{1i}}(\boldsymbol{\gamma})^T}{\partial \boldsymbol{\gamma}}\right] \left[\boldsymbol{A_i}^{1/2} \left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] \mathbf{R}(\boldsymbol{\delta}) \boldsymbol{A_i}^{1/2} \left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right]\right]^{-1} \left[\boldsymbol{w_i}^{(h)} - \boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] = 0 \quad (7)$$

and

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\zeta_i}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}}\right] \left[\boldsymbol{D_i}^{1/2} \left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] \mathbf{P}(\boldsymbol{\rho}) \boldsymbol{D_i}^{1/2} \left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right]\right]^{-1} \boldsymbol{W_i}^{(j)} \left[\boldsymbol{\gamma_i} - \boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] = 0 \quad (8)$$

where $\mathbf{R}(\boldsymbol{\delta})$ and $\mathbf{P}(\boldsymbol{\rho})$ are working correlation matrices, which must be specified, $\boldsymbol{A_i}(\boldsymbol{p_{1i}}) = \text{diag}\left[p_{1i1}(1 - p_{i1}), ..., p_{1in_i}(1 - p_{1in_i})\right]$ and $\boldsymbol{W_i}^{(h)} = \text{diag}\left[\left(1 - w_{i1}^{(h)}\right), ..., \left(1 - w_{in_i}^{(h)}\right)\right]$ and $\boldsymbol{D_i}(\zeta_i) = \text{diag}\left[\pi_{1i1}(1 - \pi_{i1}), ..., \pi_{1in_i}(1 - \pi_{1in_i})\right]$. In expressions 7 and 8, $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$ are correlation parameters and must be estimated.

Correlation parameters are estimated at each iteration by the moment estimators for δ and ρ. In special, that is the interest for the data set analysis in Section 5, for the compound simetry structure, we have the following estimators

$$\hat{\delta} = \sum_{i=1}^{K} \sum_{s<t} \frac{(w_{is} - p_{1is})(w_{it} - p_{1it})}{\sqrt{p_{1is}p_{1it}(1 - p_{1is})(1 - p_{1it})}}, \quad (9)$$

and

$$\hat{\rho} = \sum_{i=1}^{K} \sum_{s<t} \frac{(1-w_{is})(1-w_{it})(y_{is}-\pi_{1is})(y_{it}-\pi_{1it})}{\sqrt{\pi_{1is}\pi_{1it}(1-\pi_{1is})(1-\pi_{1it})}} \tag{10}$$

The estimation steps are the following (Xu (2013)).

1. Let h = 0. In addition to a guess for β and γ, it is necessary a guess for the correlation parameters δ and ρ. Usually, it is used δ=0 and ρ=0, the independent case.
2. **E** step: compute the expectation of the latent variable conditioned to the observed data, by using the expression (6).
3. **S** step: Find the solution in terms of β and γ of the equations (??) and (??).
4. Using the values β^{h+1}, γ^{h+1} and $\hat{w_{ij}}^{(h)}$, update the value of δ, by using expression 9.
5. Using the values β^{h+1}, γ^{h+1} and $\hat{w_{ij}}^{(h)}$, update the value of ρ, by using expression 10.
6. Repeat steps 2-5 until convergence. Use the same stop rule criterion for the EM algorithm established in Section 2.

Let $V_{\gamma i}$ and $V_{\beta i}$ equal to $\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\gamma)\right]\mathbf{R}(\delta)\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\gamma)\right]$ and $\boldsymbol{D_i}^{1/2}\left[\zeta_i(\beta)\right]\mathbf{P}(\rho)\boldsymbol{D_i}^{1/2}\left[\zeta_i(\beta)\right]$, respectively. Xu (ibid.) defined the sandwich variance estimator for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ as being: $\boldsymbol{B^{-1}M_1B^{-1}}$, such that

$$\boldsymbol{B} = \begin{pmatrix} \sum_{i=1}^{K} \frac{\partial p_{1i}(\gamma)^T}{\partial \gamma} V_{\gamma i}^{-1} \frac{\partial p_{1i}(\gamma)}{\partial \gamma^T} & 0 \\ 0 & \sum_{i=1}^{K} \left[\frac{\partial \zeta_i(\beta)^T}{\partial \beta}\right] \mathrm{diag}(1-\hat{\boldsymbol{w}}_i)^{1/2} V_{\beta i}^{-1} \mathrm{diag}(1-\hat{\boldsymbol{w}}_i) \frac{\partial \zeta_i(\beta)}{\partial \beta^T} \end{pmatrix}$$

and
$\boldsymbol{M_1} = \boldsymbol{MM}^T$
with

$$\boldsymbol{M} = \begin{pmatrix} \sum_{i=1}^{K} \frac{\partial p_{1i}(\gamma)^T}{\partial \gamma} V_{\gamma i}^{-1} (\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{p}}_{1i}) \\ \sum_{i=1}^{K} \frac{\partial \zeta_i(\beta)^T}{\partial \beta} V_{\beta i}^{-1} \mathrm{diag}(1-\hat{\boldsymbol{w}}_i) \left[\boldsymbol{y}_i - \zeta_i(\beta)\right] \end{pmatrix}.$$

Some extensions of this approach might also be found in the paper by Hall & Zhang (2004). They presented a strategy to obtain greater efficiency in the estimation process, replacing the equations in (7) and (8) by some combined estimation equations. Xu (2013) proposed an application of the **ES** algorithm in the context of a ZIBN (Zero-Inflated Negative Binomial) model.

4. Simulation study

Monte Carlo simulations were performed, using EM algoritm in the independent case and ES algoritm in the cluster correlated one. Two scenarios were considered taking X_1 as a standard normal and X_2 as a Bernoulli(p) with p = 0.5 and the vector of parameters equal to $\boldsymbol{\beta} = (\beta_0 = 1, \beta_1 = 2, \beta_2 = 1)'$. For the linear predictor of cure, Z_1 was generated from a standard normal and the real parameter vector equal to $\boldsymbol{\gamma} = (\gamma_0 = -2.0, \gamma_1 = -2.5)'$ to get a cure fraction of approximately 25% and $(\gamma_0 = 2.0, \gamma_1 = 2,5)'$ for 75%, respectively. In the results, VR(%) indicates the mean of the relative bias and PC(%) indicates the proportion of the confidence intervals that contained the true value of

the parameter. Mean standard error is the mean of the standard errors of each Monte Carlo simulation and the column Standard deviation estimates is the true standard error obtained by calculation the standard deviation of the estimates of the parameters of the Monte Carlo repetitions.

4.1 Independent sample- EM Algorithm

Table 1 presents the results of the 200 simulations for n = 10000 for the independent case of Section 2. Results obtained are very close to the real ones, specially mean standard errors are very similar to the true ones. Increasing the cure fraction survival estimates presented a small bias, around 10%.

Table 1. Monte Carlo simulations by the EM algorithm - cure fraction 25% and 75%.

Cure fraction	Par	Mean	Mean Standard Error	Standard deviation of estimates	VR(%)	PC(%)
25%	β_0	1,012	0,198	0,203	1,20%	95,4%
	β_1	2,026	0,213	0,200	1,30%	96,3%
	β_2	1,029	0,247	0,256	2,90%	93,0%
	γ_0	-2,037	0,281	0,290	1,85%	94,4%
	γ_1	-2,550	0,297	0,313	2,00%	94,4%
75%	β_0	1,116	0,510	0,518	11,60%	95,8%
	β_1	2,208	0,532	0,543	10,40%	97,2%
	β_2	1,086	0,544	0,539	8,60%	96,7%
	γ_0	2,015	0,177	0,183	0,75%	94,5%
	γ_1	2,519	0,256	0,251	0,76%	94,9%

4.2 Sample in clusters - ES algorithm

.

Cluster data were generated using the function *rbin* from the package *SimCorMultRes* by Touloumis (2016). The data generation process used the NORTA method, proposed by Cario & Nelson (1997) to generate values of marginal distributions from a specified correlation matrix. More details can be found in Li & Hammond (1975) and in the SimCorMultRes package documentation.

The simulation was conducted assuming the exchangeable correlation structure. This choice is justified by the practical application, as it was the adequate modeling correlation structure. The simulation was conducted considering cluster sizes of 250 and 100, with 4 measurements each. The correlation matrix to generate the cure and the observed response are identical, with a 4x4 dimension and the correlation value equal to 0.50.

Table 2 presents the results of the Monte Carlo study with 200 repetitions and a cure fraction of 25%. The average of the estimates are very close to true values. As expected the standard deviations decrease as the sample size increases. Increasing the cure ratio, the algorithm took longer to converge, approximately 39% of the simulation runs required more than 50 iterations to reach convergence.

The mean of the standard errors is different from the standard deviation of the estimates, the true value. The problem is also found in Xu (2013), because the sandwich variance treats the latent variable as known, ignoring the uncertainty about its estimate.

One way to get around with this difficulty is using bootstrap resample method at the cluster level to estimate standard error and the confidence interval. The group identification is randomly selected with replacement, loading all 4 measurements to form the database for each bootstrap repetition. 200 bootstrap repetitions were considered for each generated database, calculating the deviation of the estimates and the confidence interval with quantiles of order 2.5% and 97.5%.

Table 2. Monte Carlo simulations by the ES algorithm with cure fraction 25%

K	Par	Mean	Mean Standard Error	Standard deviation of estimates	VR(%)
250	β_0	1,049	0,168	0,321	4,90%
	β_1	2,076	0,181	0,289	3,38%
	β_2	1,248	0,251	0,349	24,80%
	γ_0	-2,047	0,115	0,392	2,35%
	γ_1	-2,502	0,154	0,403	0,08%
100	β_0	1,371	0,304	0,792	37,10%
	β_1	2,378	0,332	0,863	18,90%
	β_2	1,352	0,446	1,230	35,2%
	γ_0	-1,939	0,182	0,464	-3,5%
	γ_1	-2,546	0,246	0,502	1,96%

Table 3 presents the results for K = 250, showing the mean of the standard deviations and point estimates obtained by the bootstrap method, and PC(%). The results are in agreement with the true values. The mean deviation of the bootstrap estimates is close to the standard deviation of the estimates and the confidence intervals are close to the nominal value and symmetrical around the true value.

Table 3. Bootstrap results. Cure fraction 25% and K = 250 groups

Par	Average bootstrap deviation	Standard deviation of estimates	Average point estimate	PC(%)
β_0	0,318	0,321	1,03	91,3%
β_1	0,336	0,289	2,12	94,8%
β_2	0,390	0,349	1,25	93,1%
γ_0	0,384	0,392	-2,14	91,3%
γ_1	0,361	0,403	-2,61	91,3%

5. Real data application

The logistic model with a cure factor was applied to a real data set from the DTC-SD-UFMG briefly described in Section 1. The sample is formed by 90 patients and 104 reimplanted teeth after a traumatic avulsion, referred to treatment at DTC-DS-UFMG) after an emergency care at the Metropolitan Hospital Odilon Beherns in Belo Horizonte, Brazil. The number of replanted teeth per patient is presented in Table 4.

Table 4. Number of patients x Number of replanted teeth

Number of patients	Number of teeth
78	1
10	2
2	3

Clinical and radiographic data were collected at the first appointment at DTC-DS-UFMG. The main goal was to identify risk factors related to the presence or absence of EIRR. Among the following covariates:

- tooth storage medium after avulsion and before reimplantation, categorized as (1) dry storage or (2) wet storage (saliva, water, saline and milk);

- extra-alveolar period (PerEO), that is, the time in minutes that the tooth remained out of its soket after avulsion;
- stage of root development: stage of root formation of the avulsed tooth categorized according to the diameter of the apical foramen;
- time in days between the date of replantation and the beginning of treatment at DTC DS FUMG, called infection time.
- Prescription of systemic antibiotic therapy after reimplantation (such information was not available in 26.7% of the sample)
- patient gender.

Although, in the present sample the number of individuals with more than one traumatized tooth was small, this is a frequent feature in the data sets of the DTC-SD-UFMG. Therefore, the first step in the analysis was to identify the most appropriate model for this data set. The exchangeable correlation is the most appropriate structure for cluster analysis. Others structures avalilable, such as the first order Auto-Regressive cannot be used since there is no temporal component in a cross-sectional study. The unstructured case is not feasible due to the high number of correlation parameters to be estimated and the unbalanced design.

Table 5 presents the final model, with 90% bootstrap confidence intervals for the odds ratio. Bootstrap method was used in the application since it was observed in the simulations that the sandwich variance estimates do not eliminate the effect of the latent variable and therefore it is biased. Final model was obtained by using a backward stepwise method.

Table 5. Final fitted model estimates.

Predictor	Parameter	Estimative	DP Boot	OR	IC Boot OR (90%)
	β_0	-13,98	7,02	-	-
EIRR	β_1 (Log Time to Infection)	4,32	2,02	2,19	[1,17; 3,56]
	γ_0	1,87	1,87	-	-
Cure	γ_1 (Antibiotic-No)	-2,18	0,87	0,11	[0,02; 0,53]
	γ_2 (Log PerEO)	-0,51	0,31	0,91	[0,80; 0,99]

In terms of the cure model, antibiotic prescription and extra alveolar period (PerEO) were important markers. An increase of 20% of the PerEO decreased the chance of cure by approximately 9%. Patients who did not have an antibiotic prescription after reimplantation had approximately 89% less chance of cure. When considering the main response (EIRR), only infection time was a significant marker. An increasing of 20% in the time of infection increased the chance of EIRR by 220%. Correlation parameter estimate associated to the symmetry compound structure was -0.23 (se = 9.82) for the observed response model and -0.18 (se=0.21) for the cure one.

The regularity conditions were satisfied according to the points raised in Section 2.2. That is, the linear predictor for the observed response is not the same as the one for the cure portion of the model.

6. Conclusions and future work

The present study treated binary response with cure fraction for correlated data adopting a GEE model using the ES algorithm version to take into account the cluster structure. A Monte Carlo simulation evaluated the small sample properties of the proposed estimators and a real data illustrated the results.

Additional studies are important to assess situations in which there are substantial gains in considering such methodology including clusters. Using bootstrap has advantages and disadvantages.

Main advantage makes possible to identify factors associated with the occurrence EIRR with an easy computational implementation. The disadvantages and limitations are the computational time due to the amount of repetitions needed and some possible numerical algorithm non-convergence.

No studies were found in the literature that take into account the fact that pulpal healing is a latent cure factor for the outcome of interest. The limitation of the analysis is related to small sample sizes. Numerical problems might occur in simulation studies and real data set applications under small sample sizes. A real challenge for future works is related to remove the effect of the latent variable in the estimation of the sandwich variance presented in Xu (2013). Finally but not least, another important challenge is obtaining goodness of fit statistics or a residual analysis to check the adequacy of the model. As far as we konwn, residual analysis is available only for the logistic model with cure fraction without a cluster structure, the independent case.

References

1. Andersen, M, Lund, A, Andreasen, J. & Andreasen, F. In vitro solubility of human pulp tissue in calcium hydroxide and sodium hypochlorite. *Dental Traumatology* **8,** 104–108 (1992).
2. Cario, M. C. & Nelson, B. L. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix* tech. rep. (Citeseer, 1997).
3. Coste, S. C., e Silva, E. F., Santos, L. C. M., Ferreira, D. A. B., de Souza Côrtes, M. I., Colosimo, E. A. & Bastos, J. V. Survival of replanted permanent teeth after traumatic avulsion. *Journal of Endodontics* **46,** 370–375 (2020).
4. Dempster, A. P. *et al.* Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22 (1977).
5. Diop, A. *et al.* Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic journal of statistics* **5,** 460–483 (2011).
6. Follmann, D. A. & Lambert, D. Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference* **27,** 375–381 (1991).
7. Hall, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56,** 1030–1039 (2000).
8. Hall, D. B. & Zhang, Z. Marginal models for zero inflated clustered data. *Statistical Modelling* **4,** 161–180 (2004).
9. Kelley, M. E. & Anderson, S. J. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in medicine* **27,** 3674–3688 (2008).
10. Li, S. T. & Hammond, J. L. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics,* 557–561 (1975).
11. Louis, T. A. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 226–233 (1982).
12. Sy, J. & Taylor, J. Standard errors for the Cox proportional hazards cure model. *Mathematical and computer modelling* **33,** 1237–1251 (2001).
13. Touloumis, A. Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package. *The R Journal* **8,** 79–91. https://journal.r-project.org/archive/2016/RJ-2016-034/index.html (2016).
14. Xu, S. Generalized estimating equation based zero-inflated models with application to examining the relationship between dental caries and fluoride exposures. (2013).

15. Yamaguchi, M. *et al.* Preliminary criteria for classification of adult Still's disease. *The Journal of rheumatology* **19,** 424 (1992).