Computational Statistics and Data Analysis 55 (2011) 993–1007

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Regression models for grouped survival data: Estimation and sensitivity analysis

Elizabeth M. Hashimoto [a], Edwin M.M. Ortega [a,*], Gilberto A. Paula [b], Mauricio L. Barreto [c]

[a] Departamento de Ciências Exatas, Universidade de São Paulo, Brazil
[b] Departamento de Estatística, Universidade de São Paulo, Brazil
[c] Instituto de Saúde Coletiva, Universidade Federal da Bahia, Brazil

ARTICLE INFO

Article history:
Received 26 January 2010
Received in revised form 9 July 2010
Accepted 8 August 2010
Available online 14 August 2010

Keywords:
Censored data
Grouped survival data
Link function
Regression model
Sensitivity analysis

ABSTRACT

In this study, regression models are evaluated for grouped survival data when the effect of censoring time is considered in the model and the regression structure is modeled through four link functions. The methodology for grouped survival data is based on life tables, and the times are grouped in k intervals so that ties are eliminated. Thus, the data modeling is performed by considering the discrete models of lifetime regression. The model parameters are estimated by using the maximum likelihood and jackknife methods. To detect influential observations in the proposed models, diagnostic measures based on case deletion, which are denominated global influence, and influence measures based on small perturbations in the data or in the model, referred to as local influence, are used. In addition to those measures, the local influence and the total influential estimate are also employed. Various simulation studies are performed and compared to the performance of the four link functions of the regression models for grouped survival data for different parameter settings, sample sizes and numbers of intervals. Finally, a data set is analyzed by using the proposed regression models.

1. Introduction

Grouped survival data are obtained from studies in which all sample units are evaluated at the same moment, which implies the occurrence of an excessive number of tied lifetimes. Hence, to eliminate the presence of ties, the lifetimes are grouped into intervals so that only the information concerning the intervals in which the individuals failed or were censored is available. According to Aranda-Ordaz (1983), the presence of many tied times is a particularly problematic point in model fitting, since the verification of the model's assumptions is based on continuous data. The first ideas concerning the treatment of grouped survival data arose in the 1980s from articles by Cox (1972) and Kalbfleisch and Prentice (1973), which presented semi-parametric models for analyzing this type of data. Some more recent applications on grouped survival data can be found in the literature. For instance, Hertz-Piccioto and Rockhill (1997) studied the efficiency of partial likelihood approximations in the presence of ties; Lam and Ip (2003) reported group-based modeling; and Yu et al. (2004) proposed models with a cure fraction for grouped survival data under the parametric approach in the absence of covariates. According to Heitjan (1991), inferential analyses have rarely been performed, particularly because of the difficulty of the computational calculation due to the fact that the response variable is changed into intervals. Therefore, this paper presents a modified regression model for grouped data in which the regression structure is modeled by using four link functions (logit, complementary log–log, log–log and probit). Also, the likelihood function is modified so as to include failing individuals, individuals at risk and censored individuals.

* Corresponding address: Departamento de Ciências Exatas, USP, Av. Pádua Dias 11, Caixa Postal 9, 13418-900 Piracicaba, São Paulo, Brazil. Tel.: +55 11 19 34294127; fax: +55 11 19 34294468.
E-mail address: edwin@esalq.usp.br (E.M.M. Ortega).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.08.004

We consider a frequentist analysis of a regression model for grouped survival data. The inferential part was carried out using the asymptotic distribution of the maximum likelihood estimators, which, with a small sample, can present some difficulties. As an alternative to classic analysis, we explore the use of the jackknife estimator for regression models. In this case, it is not necessary to use the asymptotic distribution of the maximum likelihood estimators.

After modeling, it is important to check the assumptions in the model as well as to conduct a sensitivity study to detect influential or outlying observations that can distort the results. Numerous approaches have been proposed in the literature to detect influential or outlying observations. An efficient way to detect influential observations was proposed by Cook (1986). He suggested that more confidence can be placed in a model which is relatively stable under small modifications. The best-known perturbation schemes are based on case deletion, introduced by Cook (1977), in which the effect of completely removing cases from the analysis is studied. This reasoning forms the basis for our global influence introduced in Section 3.1, and in performing such an analysis, it will be possible to determine which subjects might be influential for the analysis (see, for example, Cook and Weisberg, 1982; Xie and Wei, 2007).

On the other hand, when using case deletion, all information from a single subject is deleted at once. Therefore, it is hard to tell whether that subject has any influence on a specific aspect of the model. A solution for the earlier problem can be found in a quite different paradigm, by taking a local influence approach where one again investigates how the results of an analysis change under small perturbations in the model and where these perturbations can be specific interpretations. Also, some authors have investigated the assessment of local influence in survival analysis models. For instance, Pettitt and Bin Daud (1989) investigated local influence in proportional hazard regression models; Escobar and Meeker (1992) adapted local influence methods to regression analysis with censoring; Ortega et al. (2003) considered the problem of assessing local influence in generalized log-gamma regression models with censored observations; Ortega et al. (2006) derived curvature calculations under various perturbation schemes in exponentiated Weibull regression models with censored data; Carrasco et al. (2008) investigated local influence in log-Weibull modified regression models with censored data; Silva et al. (2008) adapted global and local influence methods in log-Burr XII regression models with censored data; Ortega et al. (2009) investigated local influence in generalized log-gamma regression models with a cure fraction; and Hashimoto et al. (2010) derived the appropriate matrices for assessing local influences in the log-exponentiated Weibull regression model for interval-censored data. We have developed a similar methodology to detect influential subjects in regression models for grouped survival data.

The article is organized as follows. In Section 2, we describe the regression model for grouped survival data using one of the four link functions as well as the methods of estimation by maximum likelihood and jackknife employed in this study. In Section 3, we discuss the measurements used for analyzing sensitivity. In Section 4, we perform various simulation studies, and compare them to the performance of the four link functions, for different parameter settings, sample sizes and numbers of intervals. In Section 5, we present an application of the proposed method on a real data set. Finally, we present our main conclusions in Section 6.

2. Regression models for grouped survival data

T is defined as a random variable representing an individual's failure time and C as another random variable that is independent of T and represents the censoring time associated with this individual. The time observed for such individuals is $t = \min(T, C)$, and δ is a censoring indicator associated with the individual, with $\delta = 1$ if $T \leq C$ and $\delta = 0$ if $T > C$. If the data set observed, in addition to censoring, has an excessive number of ties, the observed times $t = \min(T, C)$ are grouped into k intervals to eliminate the ties and allow the intervals to be considered as discrete lifetimes so that discrete survival analysis techniques can be used. However, Chalita (1997) points out that the large number of intervals causes estimation bias. Consider $(T_i : 1 \leq i \leq n)$ to be a random sample of size n and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ be a p-dimensional vector of explanatory variables associated with the ith individual. Also consider that the lifetimes T_i are grouped into k mutually exclusive intervals, $I_j = [a_{j-1}, a_j), j = 1, \ldots, k$, where a_j are constructed using the methodology of life tables with $0 = a_0 < a_1 < \cdots < a_k = \infty$, (Lawless, 2003). The probability of failure of the ith individual, conditional on the explanatory variable $\mathbf{x}_i$ and survival at the beginning of the jth interval, is given by

$$p_j(\mathbf{x}_i) = P\left(T_i < a_j | T_i \geq a_{j-1}, \mathbf{x}_i\right). \tag{1}$$

In terms of the survival function $S(.)$, expression (1) is written as

$$p_j(\mathbf{x}_i) = P\left(a_{j-1} \leq T_i < a_j | T_i \geq a_{j-1}, \mathbf{x}_i\right) = 1 - \frac{S\left(a_j|\mathbf{x}_i\right)}{S\left(a_{j-1}|\mathbf{x}_i\right)}. \tag{2}$$

The regression structure represented through probability $p_j(\mathbf{x}_i)$ can be modeled by using different link functions. In this study, in addition using the logit and complementary log–log links, which are usual in grouped survival data analysis, we propose use of the log–log and probit links, as specified below.

- Logit link:

$$p_j(\mathbf{x}_i) = \exp\left(\eta_{ij}\right) / \left[1 + \exp\left(\eta_{ij}\right)\right]. \tag{3}$$

- Complementary log–log link:

$$p_j(\mathbf{x}_i) = 1 - \exp\left[-\exp\left(\eta_{ij}\right)\right]. \tag{4}$$

- Log–log link:

$$p_j(\mathbf{x}_i) = \exp\left[-\exp\left(-\eta_{ij}\right)\right]. \tag{5}$$

- Probit link:

$$p_j(\mathbf{x}_i) = \Phi(\eta_{ij}), \tag{6}$$

in which $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

Note that $\eta_{ij} = \gamma_j + \mathbf{x}_i^T\boldsymbol{\beta}$, where γ_j is the effect of the interval in the model, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of the parameters associated with each explanatory variable, for all $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

2.1. Estimation by maximum likelihood

The likelihood function is written in terms of the ith individual's failure probability in the jth interval (I_j), given that the individual survived in a_{j-1} as defined in (2). Hence, this likelihood function can be obtained by considering explanatory variables $\mathbf{x}_i$ such that the contribution from the ith individual in the jth interval for the likelihood function is given by the following.

- If the ith individual failed in the jth interval, its contribution to the likelihood function is given by $p_j(\mathbf{x}_i) = 1 - S(a_j|\mathbf{x}_i)/S(a_{j-1}|\mathbf{x}_i)$.
- If the ith individual survived (that is, it is at risk) in the jth interval, its contribution to the likelihood function is given by $1 - p_j(\mathbf{x}_i) = S(a_j|\mathbf{x}_i)/S(a_{j-1}|\mathbf{x}_i)$.
- If the ith individual is censored in the jth interval, its contribution to the likelihood function is given by $1 - p_j(\mathbf{x}_i) = S(c_i|\mathbf{x}_i)/S(a_{j-1}|\mathbf{x}_i)$, in which $c_i \in I_j$.

Thus, the likelihood function, considering the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \gamma^T)^T$, is given by

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{k}\left\{\prod_{i\in F_j}\left[p_j(\mathbf{x}_i)\right]\prod_{i\in R_j}\left[1 - p_j(\mathbf{x}_i)\right]\prod_{i\in C_j}\left[S_i(c_i|\mathbf{x}_i)/S_i(a_{j-1}|\mathbf{x}_i)\right]\right\}, \tag{7}$$

in which F_j denotes the set of failing individuals in the jth interval; R_j denotes the set of surviving individuals in the jth interval; and C_j denotes the set of censored individuals in the jth interval. The use of expression (7) is complicated in practice, since the times of c_i are unknown if the data were grouped into intervals. In this situation, an alternative is to associate a uniform distribution in interval $[a_{j-1}, a_j)$ to censoring time c_i and approximate the failing rate as being constant in the jth interval (see, for example, Thompson, 1977). After algebraic manipulation (see details in Hashimoto, 2008), the contribution of censored individuals in the jth interval is given by

$$\frac{S(c_i|\mathbf{x}_i)}{S(a_{j-1}|\mathbf{x}_i)} = \left[\frac{S(a_j|\mathbf{x}_i)}{S(a_{j-1}|\mathbf{x}_i)}\right]^{1/2}. \tag{8}$$

Now, substituting Eq. (2) into (8), we have

$$\frac{S(c_i|\mathbf{x}_i)}{S(a_{j-1}|\mathbf{x}_i)} = \left[1 - p_j(\mathbf{x}_i)\right]^{1/2}. \tag{9}$$

According to Colosimo and Giolo (2006), the interpretation of the quantity 1/2 is that observations for which censoring occurred in interval I_j are treated as if they were at risk during half of the interval under consideration.

Finally, by replacing Eq. (9) in (7), the logarithm of the likelihood function when considering the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \gamma^T)^T$, is given by

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{k}\left\{\sum_{i\in F_j}\log\left[p_j(\mathbf{x}_i)\right] + \sum_{i\in R_j}\log\left[1 - p_j(\mathbf{x}_i)\right] + \frac{1}{2}\sum_{i\in C_j}\log\left[1 - p_j(\mathbf{x}_i)\right]\right\}. \tag{10}$$

The maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\theta}}$ of the model parameters in $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ can be obtained by maximizing the log-likelihood function (10). For grouped survival data, Lawless (2003) suggested the quantity $(n_j - d_j)/n_j$ as an initial estimation of parameters γ, in which n_j is the number of individuals at risk in the jth interval and d_j is the number of failures in the jth interval. We used the Ox matrix programming language (MaxBFGS function) (see Doornik, 2007) to compute these estimates. Under conditions that are fulfilled for parameter vector $\boldsymbol{\theta}$ in the interior of the parameter space, but not on the boundary, the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is the multivariate normal $N_{p+k}(0, K(\boldsymbol{\theta})^{-1})$, where $K(\boldsymbol{\theta})$ is the

information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the $(p+k)\times(p+k)$ observed information matrix $-\ddot{\mathbf{L}}_{\theta\theta}$, and then the asymptotic inference for parameter vector $\boldsymbol{\theta}$ can be based on the normal approximation $N_{p+k}(0, -\ddot{\mathbf{L}}_{\theta\theta}^{-1})$ for $\widehat{\boldsymbol{\theta}}$. The elements of the observed information matrix

$$-\ddot{\mathbf{L}}_{\theta\theta} = \begin{pmatrix} \mathbf{L}_{\beta\beta} & \mathbf{L}_{\beta\gamma} \\ . & \mathbf{L}_{\gamma\gamma} \end{pmatrix}$$

are given in the Appendix.

The asymptotic multivariate normal distribution $N_{p+k}(0, -\ddot{\mathbf{L}}_{\theta\theta}^{-1})$ can be used to construct approximate confidence regions for some parameters in $\boldsymbol{\theta}$ and for the hazard and survival functions. In fact, a $100(1-\alpha)\%$ asymptotic confidence interval for each parameter θ_v is given by

$$ACI_v = \left(\hat{\theta}_j - z_{\alpha/2}\sqrt{-\ddot{\hat{L}}^{v,v}}, \hat{\theta}_v + z_{\alpha/2}\sqrt{-\ddot{\hat{L}}^{v,v}}\right),$$

where $-\ddot{\hat{L}}^{v,v}$ denotes the vth diagonal element of the inverse of the estimated observed information matrix $-\ddot{\hat{\mathbf{L}}}_{\theta\theta}^{-1}$ and $z_{\alpha/2}$ is the quantile $1-\alpha/2$ of the standard normal distribution, for $v = 1, \ldots, p+k$. The asymptotic normality is also useful for testing goodness of fit of some sub-models and for comparing some special sub-models using the likelihood ratio (LR) statistic.

Information criteria are based on decision theory, which penalizes models with a large number of parameters. In this study, some of these criteria were used based on AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and CAIC (Consistent Akaike Information Criterion) statistics, in which p is the number of parameters in the model and n is the sample size. The computations were done using the subroutine in the matrix programming language available in Ox (see Doornik, 2007).

2.2. Jackknife estimator

The idea of jackknifing is to transform the problem of estimating any population parameter into a problem of estimating a population mean. So, a mean value is estimated through this method, but from an unusual standpoint. An important method of implementing the jackknife method is given by Lipsitz et al. (1990), who suggest an alternative robust estimator of the covariance matrix based on a jackknife for analyzing data from repeated measures studies. In this paper, we use this method as an alternative to estimate the population parameters.

Suppose that $T_1, \ldots, T_n$ is a random sample of n values and $\bar{T} = \sum_{i=1}^{n} \frac{T_i}{n}$ is the sample mean used to estimate the population mean.

The sample mean calculated with the lth observation missing is

$$\bar{T}_{-l} = \frac{\sum_{i=1}^{n} T_i - T_l}{n-1},$$

for which

$$T_l = n\bar{T} - (n-1)\bar{T}_{-l}. \tag{11}$$

In a general situation, we consider that θ is a parameter estimated by $\hat{E}(T_1, \ldots, T_n)$, and for ease of notation we drop $(T_1, \ldots, T_n)$. Finally, we calculate $\hat{E}_{-l}$, which is obtained with the T_l observation missing. It follows from Eq. (11) that pseudo-values can be calculated:

$$\hat{E}_l^* = n\hat{E} - (n-1)\hat{E}_{-l}, \quad l = 1, \ldots, n.$$

The average of the pseudo-values is the jackknife estimate of θ given by

$$\hat{E}^* = \frac{\sum_{l=1}^{n} \hat{E}_l^*}{n}.$$

Manly (1997) suggested that an approximate $100(1-\alpha)\%$ confidence interval for θ is given by $\hat{E}^* \pm t_{\alpha/2,n-1}s/\sqrt{n}$, where $t_{\alpha/2,n-1}$ is the upper $(1-\alpha/2)$ point of the t distribution with $(n-1)$ degrees of freedom, which has the effect of removing the bias of order $1/n$.

The jackknife estimate calculations for the regression model for grouped survival data are performed for γ_j $(j = 1, \ldots, k)$ and β_m $(m = 1, \ldots, p)$, and confidence intervals are calculated separately for each parameter.

3. Sensitivity analysis

In order to assess the sensitivity aspects of the MLEs, an analysis of global influence and normal curvatures of local influence (Cook, 1986) can be carried out for some common perturbation schemes.

3.1. Global influence

The first tool to perform sensitivity analysis, as stated before, operates by means of global influence starting from case deletion. Case deletion is a common approach to study the effect of dropping the ith observation from the data set. Case deletion for the model (1) is given by

$$p_j(\mathbf{x}_{(i)}) = P\left(T_{(i)} < a_j | T_{(i)} \geq a_{j-1}, \mathbf{x}_{(i)}\right), \quad i = 1, \ldots, n. \tag{12}$$

In the following, a quantity with subscript "(i)" means the original quantity with the ith observation deleted. For model (12), the log-likelihood function is denoted by $l_{(i)}(\boldsymbol{\theta})$.

Let $\hat{\boldsymbol{\theta}}_{(i)} = \left(\hat{\boldsymbol{\gamma}}_{(i)}^T, \hat{\boldsymbol{\beta}}_{(i)}^T\right)^T$ be the MLE of $\boldsymbol{\theta}$ from $l_{(i)}(\boldsymbol{\theta})$. To assess the influence of the ith observation on MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}^T, \hat{\boldsymbol{\beta}}^T)^T$, the basic idea is to compare the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$. If deletion of an observation seriously influences the estimates, more attention should be paid to that observation. Hence, if $\hat{\boldsymbol{\theta}}_{(i)}$ is far from $\hat{\boldsymbol{\theta}}$, then the case is regarded as an influential observation. A first measure of global influence is defined as the standardized norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ (the generalized Cook distance):

$$GD_i(\boldsymbol{\theta}) = \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right)^T \left\{-\ddot{\mathbf{L}}_{\theta\theta}\right\} \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right).$$

Another alternative is to assess the values $GD_i(\boldsymbol{\gamma})$ and $GD_i(\boldsymbol{\beta})$, which reveal the impact of the ith observation on the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively. Another popular measure of the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ is the likelihood displacement:

$$LD_i(\boldsymbol{\theta}) = 2\left\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\right\}.$$

Further, we can also compute $\hat{\beta}_m - \hat{\beta}_{(i)m}$ $(m = 1, \ldots, p)$ to detect the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. Alternative global influence measures are possible. We study the behavior of a test statistic, such as a Wald test for explanatory variable or censoring effect, under a case deletion scheme.

To avoid the direct model for estimation of all observations, we can use the following one-step approximation to reduce the burden:

$$\hat{\boldsymbol{\theta}}_{(i)} = \hat{\boldsymbol{\theta}} - \ddot{\mathbf{L}}_{\hat{\theta}\hat{\theta}}^{-1} \dot{l}_{(i)}(\hat{\boldsymbol{\theta}}),$$

where $\dot{l}_{(i)}(\hat{\boldsymbol{\theta}})$ is equal to $\frac{\partial l_{(i)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

We can also apply the techniques developed by Wang et al. (1996) to evaluate how the ith observation affects a set of parameter estimates. We define the following quantity as the influential estimate (IE) for individual i and for parameter vectors $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$:

$$IE(\boldsymbol{\theta})_i = \frac{1}{(p+k)} \sum_{v=1}^{p+k} \frac{|\hat{\theta}_v - \hat{\theta}_{(i)v}|}{SE(\hat{\theta}_v)}, \qquad IE(\boldsymbol{\gamma})_i = \frac{1}{k} \sum_{j=1}^{k} \frac{|\hat{\gamma}_j - \hat{\gamma}_{(i)j}|}{SE(\hat{\gamma}_j)} \quad \text{and}$$

$$IE(\boldsymbol{\beta})_i = \frac{1}{p} \sum_{m=1}^{p} \frac{|\hat{\beta}_m - \hat{\beta}_{(i)m}|}{SE(\hat{\beta}_m)},$$

where $\hat{\theta}_v$, $\hat{\theta}_{(i)v}$, $\hat{\gamma}_j$, $\hat{\gamma}_{(i)j}$, $\hat{\beta}_m$ and $\hat{\beta}_{(i)m}$ are the MLEs of the regression model for grouped survival data. The $IE(.)_i$ calculated for individual i can be interpreted as the average relative coefficient changes for a set of estimates. It is useful for assessing the effect of parameter estimates by exclusion of the ith observation. Therefore, a relatively large value of $IE(.)_i$ indicates a potential influential observation that might cause instability in model fitting.

3.2. Local influence

Another approach, suggested by Cook (1986), uses weights instead of removing observations. A local influence calculation can be carried out for model (7). If the likelihood displacement $LD(\boldsymbol{\omega}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}$ is used, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denotes the MLE under the perturbed model, the normal curvature for $\boldsymbol{\theta}$ at direction $\mathbf{d}$, $\|\mathbf{d}\| = 1$, is given by $C_{\mathbf{d}}(\boldsymbol{\theta}) = 2|\mathbf{d}^T \boldsymbol{\Delta}^T \left[\ddot{\mathbf{L}}_{\theta\theta}\right]^{-1} \boldsymbol{\Delta}\mathbf{d}|$, where $\boldsymbol{\Delta}$ is a $(p+k) \times n$ matrix that depends on the perturbation scheme and whose elements are given by $\Delta_{vi} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\theta_v \partial\omega_i$, $i = 1, \ldots, n$ and $v = 1, \ldots, p+k$, evaluated at $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$, where $\boldsymbol{\omega}_0$ is the no-perturbation vector. For the regression model for grouped survival data, the elements of $\ddot{\mathbf{L}}_{\theta\theta}$ are given in the Appendix. We can also calculate the normal curvatures $C_{\mathbf{d}}(\boldsymbol{\gamma})$ and $C_{\mathbf{d}}(\boldsymbol{\beta})$ to perform various index plots, for instance, the index plot of the eigenvector $\mathbf{d}_{max}$ corresponding to the largest eigenvalue $C_{\mathbf{d}_{max}}$ of the matrix $\mathbf{B} = -\boldsymbol{\Delta}^T \left[\ddot{\mathbf{L}}_{\theta\theta}\right]^{-1} \boldsymbol{\Delta}$, and the index plots of $C_{\mathbf{d}_i}(\boldsymbol{\gamma})$ and $C_{\mathbf{d}_i}(\boldsymbol{\beta})$, named total local influence, where $\mathbf{d}_i$ denotes an $n \times 1$ vector of zeros with one at the ith position. Thus, the curvature at

direction $\mathbf{d}_i$ takes the form $C_i = 2|\boldsymbol{\Delta}_i^T \left[\ddot{\mathbf{L}}_{\theta\theta}\right]^{-1} \boldsymbol{\Delta}_i|$, where $\boldsymbol{\Delta}_i^T$ denotes the ith row of $\boldsymbol{\Delta}$. It is usual to point out those cases such that $C_i \geq 2\bar{C}$, where $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$.

Next, we calculate, for three perturbation schemes, the matrix

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{vi})_{(p+k)\times n} = \left(\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\theta_v \partial\boldsymbol{\omega}_i}\right)_{(p+k)\times n}, \quad v = 1, \ldots, p+k \quad \text{and} \quad i = 1, \ldots, n,$$

by considering model (1) and log-likelihood (10). Consider the vector of weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$.

Case-weight perturbation

In this case, the log-likelihood function takes the form

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{j=1}^{k}\left\{\sum_{i\in F_j}\omega_i \log\left[p_j(\mathbf{x}_i)\right] + \sum_{i\in R_j}\omega_i \log\left[1-p_j(\mathbf{x}_i)\right] + \frac{1}{2}\sum_{i\in C_j}\omega_i \log\left[1-p_j(\mathbf{x}_i)\right]\right\},$$

$0 \leq \omega_i \leq 1$ and $\boldsymbol{\omega}_0 = (1, \ldots, 1)^T$.

We denote $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{\boldsymbol{\beta}}, \boldsymbol{\Delta}_{\boldsymbol{\gamma}})^T$. Then, the elements of vector $\boldsymbol{\Delta}_{\boldsymbol{\beta}}$ take the form

$$\Delta_{mi} = \begin{cases} -\sum_{l=1}^{j-1}\left\{\dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i)\right]_{\beta_m}}{1-\hat{p}_l(\mathbf{x}_i)}\right\} + \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\beta_m}}{\hat{p}_j(\mathbf{x}_i)} & \text{if } i \in F_j, \\ -\sum_{l=1}^{j-1}\left\{\dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i)\right]_{\beta_m}}{1-\hat{p}_l(\mathbf{x}_i)}\right\} - \dfrac{1}{2}\left\{\dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\beta_m}}{1-\hat{p}_j(\mathbf{x}_i)}\right\} & \text{if } i \in C_j. \end{cases}$$

On the other hand, the elements of vector $\boldsymbol{\Delta}_{\boldsymbol{\gamma}}$ can be shown to be given by

$$\Delta_{ji} = \begin{cases} \left\{\dfrac{1}{\hat{p}_j(\mathbf{x}_i)}\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\gamma_j}\right\} & \text{if } i \in F_j, \\ -\dfrac{1}{2}\left\{\dfrac{1}{1-\hat{p}_j(\mathbf{x}_i)}\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\gamma_j}\right\} & \text{if } i \in R_j, \\ -\left\{\dfrac{1}{1-\hat{p}_j(\mathbf{x}_i)}\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\gamma_j}\right\} & \text{if } i \in C_j, \end{cases}$$

where

$$\left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\beta_m} = \partial p(\mathbf{x}_i)/\partial\beta_m|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad \text{and} \quad \left[\dot{\hat{p}}_j(\mathbf{x}_i)\right]_{\gamma_j} = \partial p(\mathbf{x}_i)/\partial\gamma_j|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

for $m = 1, \ldots, p$ and $j = 1, \ldots, k$.

Explanatory variable perturbation

Consider now an additive perturbation on a particular continuous explanatory variable, namely X_t, by making $x_{it\omega} = x_{it} + \omega_i S_x$, where S_x is a scaled factor, $\omega_i \in \mathbf{R}$. This perturbation scheme leads to the following expressions for the log-likelihood function and for the elements of matrix $\boldsymbol{\Delta}$.

In this case, the log-likelihood function takes the form

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{j=1}^{k}\left\{\sum_{i\in F_j}\log\left[p_j(\mathbf{x}_i^*)\right] + \sum_{i\in S_j}\log\left[1-p_j(\mathbf{x}_i^*)\right] + \frac{1}{2}\sum_{i\in C_j}\log\left[1-p_j(\mathbf{x}_i^*)\right]\right\},$$

where $\mathbf{x}_i^{*T} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_t(x_{it} + \omega_i S_x) + \cdots + \beta_p x_{ip}$ and $\boldsymbol{\omega}_0 = (0, 0, \ldots, 0)^T$.

The elements of the matrix $\boldsymbol{\Delta}_{\boldsymbol{\beta}}$, for $m = 1, \ldots, p$ and $m \neq t$, take the form

$$\Delta_{mi} = \begin{cases} -\sum_{l=1}^{j-1}\left\{\dfrac{\left[\ddot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_m\omega}}{1-\hat{p}_l(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_m}\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_l(\mathbf{x}_i^*)\right]^2}\right\} + \left[\dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m\omega}}{\hat{p}_j(\mathbf{x}_i^*)} - \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[\hat{p}_j(\mathbf{x}_i^*)\right]^2}\right] & \text{if } i \in F_j, \\ -\sum_{l=1}^{j-1}\left\{\dfrac{\left[\ddot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_m\omega}}{1-\hat{p}_l(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_m}\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_l(\mathbf{x}_i^*)\right]^2}\right\} - \left\{\dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m\omega}}{1-\hat{p}_j(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_j(\mathbf{x}_i^*)\right]^2}\right\} & \text{if } i \in C_j. \end{cases}$$

The elements of the matrix $\mathbf{\Delta}_{\boldsymbol{\beta}}$, for $m = t$, are given by

$$\Delta_{ti} = \begin{cases} -\sum_{l=1}^{j-1} \left\{ \dfrac{\left[\ddot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_t\omega}}{1-\hat{p}_l(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_t}\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_l(\mathbf{x}_i^*)\right]^2} \right\} + \left\{ \dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_t\omega}}{\hat{p}_j(\mathbf{x}_i^*)} - \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_t}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[\hat{p}_j(\mathbf{x}_i^*)\right]^2} \right\} & \text{if } i \in F_j, \\ -\sum_{l=1}^{j-1} \left\{ \dfrac{\left[\ddot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_t\omega}}{1-\hat{p}_l(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\beta_t}\left[\dot{\hat{p}}_l(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_l(\mathbf{x}_i^*)\right]^2} \right\} - \left\{ \dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_t\omega}}{1-\hat{p}_j(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_t}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_j(\mathbf{x}_i^*)\right]^2} \right\} & \text{if } i \in C_j. \end{cases}$$

The elements of vector $\mathbf{\Delta}_{\boldsymbol{\gamma}}$, for $j = 1, \ldots, k$, can be expressed as

$$\Delta_{ji} = \begin{cases} \dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j\omega}}{\hat{p}_j(\mathbf{x}_i^*)} - \dfrac{\left[\dot{p}_j(\mathbf{x}_i^*)\right]_{\gamma_j}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[\hat{p}_j(\mathbf{x}_i^*)\right]^2} & \text{if } i \in F_j, \\ -\left\{ \dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j\omega}}{1-\hat{p}_j(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_j(\mathbf{x}_i^*)\right]^2} \right\} & \text{if } i \in R_j, \\ -\dfrac{1}{2}\left\{ \dfrac{\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j\omega}}{1-\hat{p}_j(\mathbf{x}_i^*)} + \dfrac{\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j}\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega}}{\left[1-\hat{p}_j(\mathbf{x}_i^*)\right]^2} \right\} & \text{if } i \in C_j, \end{cases}$$

where

$$\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m} = \partial p(\mathbf{x}_i^*)/\partial\beta_m|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}, \qquad \left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j} = \partial p(\mathbf{x}_i^*)/\partial\gamma_j|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0},$$
$$\left[\dot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\omega} = \partial p(\mathbf{x}_i^*)/\partial\omega_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}, \qquad \left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_m\omega} = \partial^2 p(\mathbf{x}_i^*)/\partial\beta_m\omega_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0},$$
$$\left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\beta_t\omega} = \partial^2 p(\mathbf{x}_i^*)/\partial\beta_t\omega_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0} \quad \text{and} \quad \left[\ddot{\hat{p}}_j(\mathbf{x}_i^*)\right]_{\gamma_j\omega} = \partial^2 p(\mathbf{x}_i^*)/\partial\gamma_j\omega_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0},$$

for $m = 1, \ldots, p$ and $j = 1, \ldots, k$.

4. Simulation study

We conducted Monte Carlo simulation studies to assess the finite sample behavior of the maximum likelihood estimators of β_1 and β_2 over the four link functions, considering different numbers of intervals. The sample size generated was $n = 50$. Following the method proposed by Colosimo et al. (2000), we consider the numbers of intervals $k = 5$, $k = 8$ and $k = 10$, fixing the parameters as $\beta_1 = 1$ and $\beta_2 = 0.5$, with the covariate x_{1i} generated from a uniform distribution in the range [0, 1] and the covariate x_{2i} generated from a binomial distribution with success probability 0.70, respectively. The censoring times, denoted by $C_1, \ldots, C_n$, were generated from a binomial distribution with failure probability 50%. Thus, a sequence of Bernoulli trials, corresponding to intervals visited by individuals, was generated with the probability of failure given by Eqs. (3)–(6). First we generated the sequence of Bernoulli trials by means of the logit link, and then we estimated the parameters β_1 and β_2 considering the four link functions proposed in this article. Analogously, we generated the Bernoulli trials by means of the complementary log–log link, and so on with the other link functions, allowing us to study the behavior of the four link functions in the different numbers of intervals. For each configuration of k and link function, all results were obtained from 1000 Monte Carlo replications, and the simulations were carried out using the Ox matrix programming language. For each fit, the mean square error (MSE) was calculated.

From the simulation results (Table 1), we can extract the following interpretations.

- The MSEs using the link logit and complementary log–log are smaller.
- By increasing the number of intervals, only the MSE of the log–log link function increases.
- The other link function does not change significantly when the number of intervals is increased.
- Future research should obtain bias corrections for these estimators, thus reducing their systematic errors in finite samples.

5. Application

In this paper, we consider a data set provided by the Instituto de Saúde Coletiva – Universidade Federal da Bahia. This data set was designed to evaluate the effect of vitamin-A supplementation on recurrent diarrheal episodes in small children (see, for example, Barreto et al., 1994). The data are from a randomized community trial that was designed to evaluate the effect

Table 1
MSE of $\hat{\beta}_1$ and $\hat{\beta}_2$, the maximum likelihood estimators of the regression models for grouped survival data.

k	Generation	Parameters	Logit	Complementary	Probit	Log–log
			MSE	MSE	MSE	MSE
5	Logit	β_1	0.333	0.366	0.460	0.496
		β_2	0.082	0.086	0.100	0.106
	Complementary	β_1	0.336	0.359	0.453	0.488
		β_2	0.081	0.083	0.102	0.109
	Probit	β_1	0.348	0.379	0.473	0.510
		β_2	0.081	0.087	0.106	0.110
	Log–log	β_1	0.347	0.382	0.471	0.501
		β_2	0.082	0.083	0.102	0.108
8	Logit	β_1	0.337	0.359	0.471	0.532
		β_2	0.082	0.084	0.106	0.117
	Complementary	β_1	0.350	0.374	0.487	0.535
		β_2	0.084	0.085	0.106	0.116
	Probit	β_1	0.335	0.338	0.464	0.521
		β_2	0.082	0.082	0.105	0.117
	Log–log	β_1	0.331	0.345	0.468	0.535
		β_2	0.084	0.084	0.101	0.113
10	Logit	β_1	0.335	0.362	0.506	0.587
		β_2	0.082	0.082	0.113	0.131
	Complementary	β_1	0.340	0.359	0.494	0.563
		β_2	0.082	0.084	0.107	0.122
	Probit	β_1	0.332	0.344	0.481	0.558
		β_2	0.088	0.089	0.113	0.129
	Log–log	β_1	0.344	0.358	0.508	0.589
		β_2	0.081	0.082	0.110	0.129

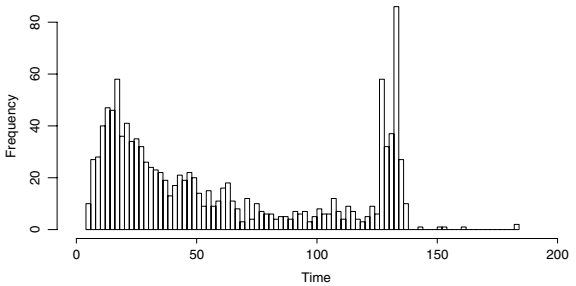

Fig. 1. Frequency graph for vitamin-A data.

of vitamin-A supplementation on diarrheal episodes in 1207 pre-school children, aged 6–48 months at the baseline, who were assigned to receive either a placebo or vitamin A. The study was conducted is a small town in the Northeast of Brazil from December 1990 to December 1991.

The vitamin-A dosage was 100,000 IU for children younger than 12 months and 200,000 IU for older children. The latter is the highest dosage guideline established by the World Health Organization (WHO) for the prevention of vitamin-A deficiency.

The total time was defined as the time from the first dose of vitamin A until the occurrence of an episode of diarrhea. An episode of diarrhea was defined as a sequence of days with diarrhea and a day with diarrhea was defined when 3 or more liquid or semi-liquid bowel movements were reported in a 24-h period. The information on the occurrence of diarrhea collected at each visit corresponds to a recall period of 48–72 h. The number of liquid and semi-liquid movements per 24 h was recorded. Of the 1207 children under study, 925 showed a diarrheal episode, which means that 282 children showed censored times. The explanatory variables considered in the models are x_{i1}: age at baseline (in months); x_{i2}: treatment (0 = placebo, 1 = vitamin A) and x_{i3}: gender (0 = girl, 1 = boy).

To have an idea about the behavior of the vitamin-A data, we constructed a frequency histogram for the data set (Fig. 1).

Fig. 1 shows a high frequency of observations at the initial times and near the end of the study. This indicates that a significant number of ties exist, and for this reason, the grouped data methodology is an alternative to model the vitamin-A data set.

We categorized the lifetimes into eight intervals $\{(4, 21], (21, 38], \ldots, (126, 185]\}$. No criterion exists to establish the number of intervals. The only assumption is the presence of at least one failure in each interval. Hence, for the sake of convenience, we choose $k = 8$ intervals. Next, we tabulated a life table (Table 2) by considering eight intervals.

Table 2
Life table for vitamin-A data.

I_j Interval	Number of failures	Number of censorings	Number at risk	$\hat{S}(t)$
[4; 21)	292	0	1207	1.000
[21; 38)	243	4	915	0.758
[38; 55)	138	6	668	0.556
[55; 73)	101	2	524	0.441
[73; 90)	46	3	421	0.356
[90; 108)	49	6	372	0.317
[108; 126)	46	11	317	0.275
[126; 185)	10	250	260	0.234

Table 3
AIC, BIC and CAIC for the regression model parameters for vitamin-A grouped data and the values for model selection criteria using the four link functions.

Measures	Logit link	Complementary log–log link	Log–log link	Probit link
AIC	4378.86	4378.87	4379.93	4382.26
BIC	4434.91	4434.93	4435.99	4438.31
CAIC	4445.91	4379.88	4446.99	4449.31

Table 4
Maximum likelihood and jackknife estimates for the parameters of the regression model for grouped survival data fitted to vitamin-A data.

$\boldsymbol{\theta}$	Maximum likelihood estimates				Jackknife estimates		
	Estimate	S.E.	p-value	95% C.I.	Estimate	S.E.	95% C.I.
γ_1	−0.079	0.119	0.506	(−0.312; 0.154)	−0.081	0.123	(−0.323; 0.161)
γ_2	0.104	0.127	0.415	(−0.145; 0.353)	0.104	0.133	(−0.156; 0.364)
γ_3	−0.174	0.143	0.222	(−0.454; 0.106)	−0.177	0.148	(−0.467; 0.114)
γ_4	−0.212	0.156	0.175	(−0.518; 0.094)	−0.218	0.162	(−0.536 ; 0.100)
γ_5	−0.859	0.192	<0.001	(−1.235; −0.483)	−0.858	0.199	(−1.248; −0.467)
γ_6	−0.627	0.191	0.001	(−1.001; −0.253)	−0.627	0.197	(−1.013; −0.241)
γ_7	−0.465	0.198	0.019	(−0.853; −0.077)	−0.471	0.202	(0.868; −0.074)
γ_8	−1.212	0.350	0.001	(−1.898; −0.526)	−1.164	0.368	(−1.885; −0.442)
β_1	−0.037	0.003	0.000	(−0.043; −0.031)	−0.038	0.003	(−0.044; −0.031)
β_2	−0.161	0.076	0.033	(−0.310; −0.012)	−0.162	0.076	(−0.311; −0.012)
β_3	−0.029	0.076	0.699	(−0.178; 0.120)	−0.029	0.076	(−0.179; 0.120)

Now, by considering the regression model for grouped survival data (see Section 2), the link functions can be described as follows.

- Logit link: $p_j(\mathbf{x}_i) = \exp(\gamma_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})/\big[1 + \exp(\gamma_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\big]$.
- Complementary log–log link: $p_j(\mathbf{x}_i) = 1 - \exp[-\exp(\gamma_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]$.
- Log–log link: $p_j(\mathbf{x}_i) = \exp\{-\exp[-(\gamma_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]\}$.
- Probit link: $p_j(\mathbf{x}_i) = \Phi(\gamma_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$.

Here $i = 1, \ldots, 1207$ and $j = 1, \ldots, 8$.

To select a link function among the logit, complementary log–log, log–log and probit so as to proceed with the analysis, we used the AIC, BIC and CAIC. The results in Table 3 show that the AIC, BIC and CAIC measurements are similar. However, the lowest value for the selection criteria was observed for the logit link function, except when the CAIC was considered. Hence, we decided to continue analyzing the vitamin-A data by using the logit link function. Note that these results are similar to those obtained in Section 4.

5.1. Maximum likelihood and jackknife results

To obtain the MLEs for the parameters in the regression model for fitted grouped survival data, we used the subroutine MaxBFGS in Ox, whose results are given in Table 4. Additionally, Table 4 gives the jackknife estimates for the parameters of the regression model for fitted grouped survival data.

The explanatory variables x_1 (age) and x_2 (treatment) are significant in the model at a 5% significance level. However, the explanatory variable gender is not significant at this level.

Although the estimates from the two methods seem to be very similar, the MLEs appear more conservative. Therefore, since for this sample size ($n = 1207$) normality is expected for the jackknife estimator, one may also expect some symmetrical distribution for the MLEs with heavy tails. We therefore continue the analysis by using the MLEs.

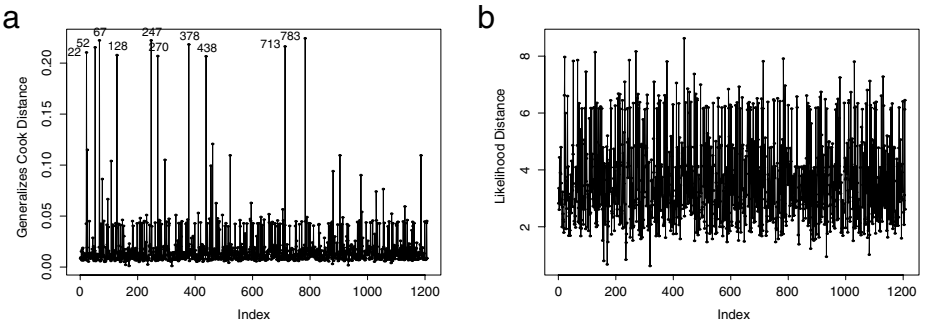

Fig. 2. (a) Index plot of $GD_i(\boldsymbol{\theta})$ and (b) index plot of $LD_i(\boldsymbol{\theta})$ from the fit of the regression model for vitamin-A grouped survival data.

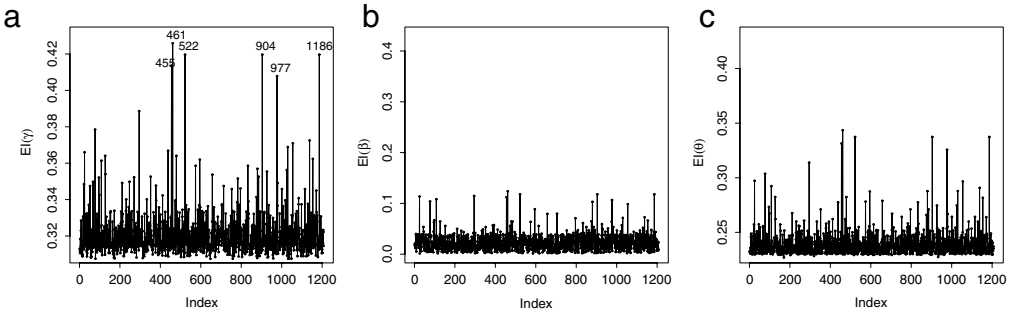

Fig. 3. (a) Index plot of $IE(\boldsymbol{\theta})_i$, (b) index plot of $IE(\boldsymbol{\beta})_i$ and (c) index plot of $IE(\boldsymbol{\gamma})_i$ from the fit of the regression model for vitamin-A grouped survival data.

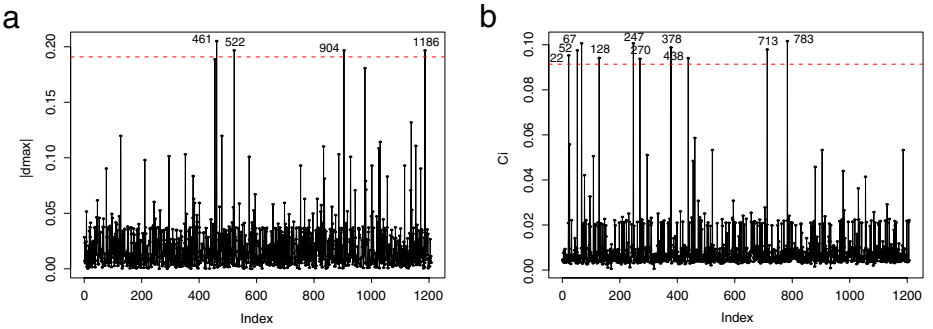

Fig. 4. Index plot for $\boldsymbol{\theta}$ (case-weight perturbation): (a) $|\mathbf{d}_{max}|$ and (b) total local influence C_i. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2. Sensitivity analysis

Global influence analysis

In this subsection, we use Ox to compute the case-deletion measures $GD(\boldsymbol{\theta})_i$, $LD(\boldsymbol{\theta})_i$, $IE(\boldsymbol{\theta})_i$, $IE(\boldsymbol{\gamma})_i$ and $IE(\boldsymbol{\beta})_i$ presented in Section 3.1. From Fig. 2, we can see that cases ♯22, ♯52, ♯67, ♯128, ♯247, ♯270, ♯378, ♯438, ♯713 and ♯783 are possible influential observations. Fig. 3 presents plots for the influential estimate (IE_i), where observations ♯455, ♯461, ♯522, ♯904, ♯977 and ♯1186 again stand out.

Local and total influence analysis

In this section, we analyze the local influence for the vitamin-A grouped survival data using a regression model for grouped survival data.

Case-weight perturbation

By applying the local influence theory developed in Section 3.2, where case-weight perturbation is used, $C_{\mathbf{d}_{max}} = 1.14$ was obtained as the maximum curvature. In Fig. 4a, the plot of the eigenvector corresponding to $C_{\mathbf{d}_{max}}$ is presented, and the total influence C_i is shown in Fig. 4b. Observations ♯22, ♯52, ♯67, ♯128, ♯247, ♯270, ♯378, ♯438, ♯461, ♯522, ♯713, ♯783, ♯904 and ♯1186 stand out most from the others. Note that the red line in the figures (local influence) helps to visualize those individuals that are furthest away.

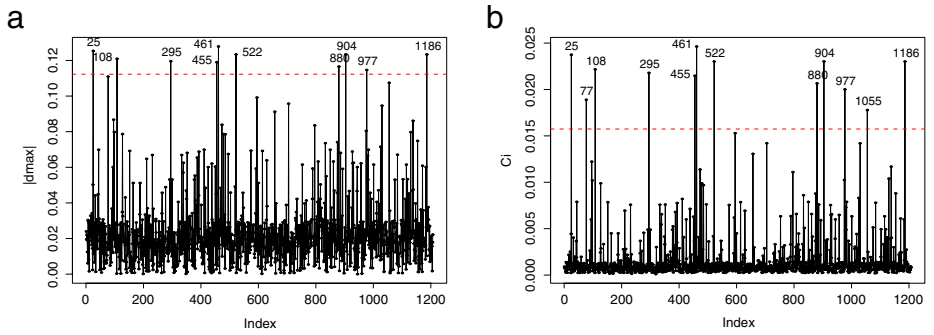

Fig. 5. Index plot for $\boldsymbol{\theta}$ (age explanatory variable perturbation): (a) $|\mathbf{d}_{max}|$ and (b) total local influence C_i.

Explanatory variable perturbation

The perturbation of the vector for the covariate age ($\mathbf{x}_1$) is investigated here. For perturbation of this covariate, $C_{\mathbf{d}_{max}} = 1.41$ was obtained as the maximum curvature. The respective plots of $|\mathbf{d}_{max}|$ as well as total local influence C_i against the observation index are shown in Fig. 5a and 5b, respectively. Observations ♯25, ♯77, ♯108, ♯295, ♯455, ♯461, ♯522, ♯880, ♯904, ♯977, ♯1055 and ♯1186 stand out most from the others.

5.3. Impact of the detected influential observations

From the previous sections, we can consider observations ♯455, ♯461, ♯522, ♯904, ♯977 and ♯1186 as possible influential or outlier observations. Note that these observations appear often in the figures of influence (global, local and total). These observations represent individuals grouped in the eighth interval. They are girls who received the placebo treatment.

Table 5 shows the relative changes (in percentage) of each parameter estimate, defined by $RC_{\beta_m} = [(\hat{\beta}_m - \hat{\beta}_m(I))/\hat{\beta}_m] \times 100$, and the corresponding p-values, where $\hat{\beta}_m(I)$ denotes the MLE of β_m after set I of observations has been removed, for all $m = 1, 2, 3$. Note that, in Table 5,

$$
\begin{aligned}
&\text{set } I_1 = \{\sharp 455\}, \quad \text{set } I_2 = \{\sharp 461\}, \quad \text{set } I_3 = \{\sharp 522\}, \quad \text{set } I_4 = \{\sharp 904\}, \quad \text{set } I_5 = \{\sharp 977\},\\
&\text{set } I_6 = \{\sharp 1186\}, \quad \text{set } I_7 = \{\sharp 455, \ \sharp 461\}, \quad \text{set } I_8 = \{\sharp 455, \ \sharp 522\}, \quad \text{set } I_9 = \{\sharp 455, \ \sharp 904\},\\
&\text{set } I_{10} = \{\sharp 455, \ \sharp 977\}, \quad \text{set } I_{11} = \{\sharp 455, \ \sharp 1186\}, \quad \text{set } I_{12} = \{\sharp 461, \ \sharp 522\},\\
&\text{set } I_{13} = \{\sharp 461, \ \sharp 904\}, \quad \text{set } I_{14} = \{\sharp 461, \ \sharp 977\}, \quad \text{set } I_{15} = \{\sharp 461, \ \sharp 1186\},\\
&\text{set } I_{16} = \{\sharp 522, \ \sharp 904\}, \quad \text{set } I_{17} = \{\sharp 522, \ \sharp 977\}, \quad \text{set } I_{18} = \{\sharp 522, \ \sharp 1186\},\\
&\text{set } I_{19} = \{\sharp 904, \ \sharp 977\}, \quad \text{set } I_{20} = \{\sharp 904, \ \sharp 1186\}, \quad \text{set } I_{21} = \{\sharp 977, \ \sharp 1186\},\\
&\text{set } I_{22} = \{\sharp 455, \ \sharp 461, \ \sharp 522\}, \quad \text{set } I_{23} = \{\sharp 455, \ \sharp 461, \ \sharp 904\},\\
&\text{set } I_{24} = \{\sharp 455, \ \sharp 461, \ \sharp 977\}, \quad \text{set } I_{25} = \{\sharp 455, \ \sharp 461, \ \sharp 1186\},\\
&\text{set } I_{26} = \{\sharp 461, \ \sharp 522, \ \sharp 904\}, \quad \text{set } I_{27} = \{\sharp 461, \ \sharp 522, \ \sharp 977\},\\
&\text{set } I_{28} = \{\sharp 461, \ \sharp 522, \ \sharp 1186\}, \quad \text{set } I_{29} = \{\sharp 522, \ \sharp 904, \ \sharp 977\},\\
&\text{set } I_{30} = \{\sharp 522, \ \sharp 904, \ \sharp 1186\}, \quad \text{set } I_{31} = \{\sharp 904, \ \sharp 977, \ \sharp 1186\},\\
&\text{set } I_{32} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 904\}, \quad \text{set } I_{33} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 977\},\\
&\text{set } I_{34} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 1186\}, \quad \text{set } I_{35} = \{\sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 977\},\\
&\text{set } I_{36} = \{\sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 1186\}, \quad \text{set } I_{37} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 977\},\\
&\text{set } I_{38} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 1186\},\\
&\text{set } I_{39} = \{\sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 977, \ \sharp 1186\} \quad \text{and}\\
&\text{set } I_{40} = \{\sharp 455, \ \sharp 461, \ \sharp 522, \ \sharp 904, \ \sharp 977, \ \sharp 1186\}.
\end{aligned}
$$

Table 5 shows that the maximum likelihood estimates from the regression model for grouped survival data are not highly sensitive under deletion of the outlying observations. In general, the significance of the parameter estimates does not change (at 5%) after removing set I. Therefore, we do not have inferential changes after removing the observations handed out in the diagnostic plots.

6. Concluding remarks

In this paper, a regression model for grouped survival data is proposed as an alternative to model lifetime in the presence of many tied data. We used the quasi-Newton algorithm to obtain the maximum likelihood estimates and performed

Table 5
Relative changes [RC- in %], parameter estimates and their p-values in parentheses for the indicated set.

Dropping	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
All observations	[−]	[−]	[−]
	−0.037	−0.161	−0.029
	(0.000)	(0.033)	(0.699)
Set I_1	[−1]	[−4]	[−25]
	−0.038	−0.168	−0.037
	(0.000)	(0.026)	(0.629)
Set I_2	[−1]	[−5]	[−26]
	−0.038	−0.169	−0.037
	(0.000)	(0.026)	(0.625)
Set I_3	[−1]	[−5]	[−26]
	−0.038	−0.169	−0.039
	(0.000)	(0.026)	(0.627)
Set I_4	[−1]	[−5]	[−26]
	−0.038	−0.169	−0.037
	(0.000)	(0.026)	(0.627)
Set I_5	[−1]	[−4]	[−24]
	−0.038	−0.168	−0.036
	(0.000)	(0.027)	(0.630)
Set I_6	[−1]	[−5]	[−26]
	−0.038	−0.169	−0.037
	(0.000)	(0.026)	(0.627)
Set I_7	[−3]	[−9]	[−52]
	−0.038	−0.176	−0.044
	(0.000)	(0.020)	(0.557)
Set I_8	[−3]	[−9]	[−51]
	−0.038	−0.176	−0.044
	(0.000)	(0.020)	(0.559)
Set I_9	[−3]	[−9]	[−51]
	−0.038	−0.176	−0.044
	(0.000)	(0.020)	(0.559)
Set I_{10}	[−2]	[−9]	[−50]
	−0.038	−0.176	−0.044
	(0.000)	(0.021)	(0.562)
Set I_{11}	[−3]	[−9]	[−51]
	−0.038	−0.176	−0.044
	(0.000)	(0.020)	(0.559)
Set I_{12}	[−3]	[−9]	[−53]
	−0.038	−0.177	−0.045
	(0.000)	(0.020)	(0.555)
Set I_{13}	[−3]	[−9]	[−53]
	−0.038	−0.177	−0.045
	(0.000)	(0.020)	(0.555)
Set I_{14}	[−3]	[−9]	[−51]
	−0.038	−0.176	−0.044
	(0.000)	(0.020)	(0.559)
Set I_{15}	[−3]	[−9]	[−53]
	−0.038	−0.177	−0.045
	(0.000)	(0.020)	(0.555)
Set I_{16}	[−3]	[−9]	[−52]
	−0.039	−0.176	−0.045
	(0.000)	(0.020)	(0.557)
Set I_{17}	[−3]	[−9]	[−50]
	−0.038	−0.176	−0.044
	(0.000)	(0.021)	(0.561)
Set I_{18}	[−3]	[−9]	[−52]
	−0.038	−0.176	−0.045
	(0.000)	(0.020)	(0.557)
Set I_{19}	[−3]	[−9]	[−50]
	−0.038	−0.176	−0.044
	(0.000)	(0.021)	(0.561)
Set I_{20}	[−3]	[−9]	[−52]
	−0.038	−0.176	−0.045
	(0.000)	(0.020)	(0.557)
Set I_{21}	[−3]	[−9]	[−50]
	−0.038	−0.176	−0.044
	(0.000)	(0.021)	(0.561)
Set I_{22}	[−4]	[−14]	[−79]

Table 5 (*continued*)

Dropping	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.490)
Set I_{23}	[−4]	[−14]	[−79]
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.490)
Set I_{24}	[−4]	[−14]	[−77]
	−0.039	−0.184	−0.052
	(0.000)	(0.016)	(0.493)
Set I_{25}	[−4]	[−14]	[−79]
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.490)
Set I_{26}	[−4]	[−14]	[−80]
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.488)
Set I_{27}	[−4]	[−14]	[−78]
	−0.039	−0.184	−0.052
	(0.000)	(0.015)	(0.491)
Set I_{28}	[−4]	[−14]	[−80]
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.488)
Set I_{29}	[−4]	[−14]	[−77]
	−0.039	−0.184	−0.052
	(0.000)	(0.016)	(0.493)
Set I_{30}	[−4]	[−14]	[−79]
	−0.039	−0.184	−0.053
	(0.000)	(0.015)	(0.490)
Set I_{31}	[−2]	[−10]	[−43]
	−0.038	−0.178	−0.042
	(0.000)	(0.019)	(0.581)
Set I_{32}	[−6]	[−19]	[−106]
	−0.040	−0.192	−0.061
	(0.000)	(0.011)	(0.425)
Set I_{33}	[−5]	[−19]	[−105]
	−0.039	−0.192	−0.060
	(0.000)	(0.012)	(0.429)
Set I_{34}	[−6]	[−19]	[−106]
	−0.040	−0.192	−0.061
	(0.000)	(0.011)	(0.425)
Set I_{35}	[−6]	[−19]	[−106]
	−0.039	−0.192	−0.060
	(0.000)	(0.012)	(0.427)
Set I_{36}	[−6]	[−19]	[−107]
	−0.040	−0.192	−0.061
	(0.000)	(0.011)	(0.423)
Set I_{37}	[−7]	[−24]	[−133]
	−0.040	−0.200	−0.069
	(0.000)	(0.009)	(0.377)
Set I_{38}	[0]	[0]	[0]
	−0.037	−0.161	−0.029
	(0.000)	(0.033)	(0.699)
Set I_{39}	[0]	[0]	[0]
	−0.037	−0.161	−0.029
	(0.000)	(0.033)	(0.699)
Set I_{40}	[0]	[0]	[0]
	−0.037	−0.161	−0.029
	(0.000)	(0.033)	(0.699)

asymptotic tests for the parameters based on the asymptotic distribution of the maximum likelihood estimators. On the other hand, as an alternative analysis, the paper discusses the use of the jackknife estimator for the regression model for grouped survival data. In the applications within real data, we observed that all estimation methods presented similar results. Furthermore, this article compared the performance of the proposed model considering four link functions (logit, complementary log–log, log–log and probit). We also applied global and local influence methodologies in a regression model for grouped survival data. The necessary matrices for application of the techniques were obtained by taking into account some usual perturbations in the model/data. By applying the procedures on a data set from the medical area, we could assess the sensitivity aspects of the maximum likelihood estimates under some perturbation schemes as well as check the goodness-of-fit of the postulated model. Although the diagnostic plots detected some possible influential observations, their

deletion did not cause inferential changes in the results. Furthermore, this article compared the performance of the proposed model under four link functions and different numbers of intervals based on mean squared error through a simulation study. These simulations suggest that regression models for grouped survival data can be used for modeling data considering logit and complementary log–log link functions. The approach was applied to real data sets, which indicates the usefulness of the approach.

Appendix. Hessian matrix $\ddot{\mathbf{L}}(\boldsymbol{\alpha})$

Here, we derive the necessary formulas to obtain the second-order partial derivatives of the log-likelihood function. After some algebraic manipulations, we obtain

$$\begin{aligned}
\mathbf{L}_{\beta_m\beta_y} &= \sum_{j=1}^{k}\left\{\sum_{i\in F_j}\left[\frac{\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y}}{p_j(\mathbf{x}_i)}-\left(\frac{\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}}{\left[p_j(\mathbf{x}_i)\right]^2}\right)\right]-\sum_{i\in R_j}\left[\frac{\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y}}{1-p_j(\mathbf{x}_i)}+\right.\right.\\
&\left.\left.+\left(\frac{\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}}{\left[1-p_j(\mathbf{x}_i)\right]^2}\right)\right]-\frac{1}{2}\sum_{i\in C_j}\left[\frac{\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y}}{1-p_j(\mathbf{x}_i)}+\left(\frac{\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}}{\left[1-p_j(\mathbf{x}_i)\right]^2}\right)\right]\right\},\\
\mathbf{L}_{\beta_m\gamma_h} &= \sum_{i\in F_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h}}{p_h(\mathbf{x}_i)}-\left(\frac{[\dot{p}_h(\mathbf{x}_i)]_{\beta_m}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}}{[p_h(\mathbf{x}_i)]^2}\right)\right]-\sum_{i\in R_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h}}{1-p_h(\mathbf{x}_i)}+\right.\\
&\left.+\left(\frac{[\dot{p}_h(\mathbf{x}_i)]_{\beta_m}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}}{[1-p_h(\mathbf{x}_i)]^2}\right)\right]-\frac{1}{2}\sum_{i\in C_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h}}{1-p_h(\mathbf{x}_i)}+\left(\frac{[\dot{p}_h(\mathbf{x}_i)]_{\beta_m}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}}{[1-p_h(\mathbf{x}_i)]^2}\right)\right],\\
\mathbf{L}_{\gamma_h\gamma_s} &= \sum_{i\in F_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s}}{p_h(\mathbf{x}_i)}-\frac{[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}}{[p_h(\mathbf{x}_i)]^2}\right]-\sum_{i\in R_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s}}{1-p_h(\mathbf{x}_i)}+\right.\\
&\left.+\frac{[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}}{[1-p_h(\mathbf{x}_i)]^2}\right]-\frac{1}{2}\sum_{i\in C_j}\left[\frac{[\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s}}{1-p_h(\mathbf{x}_i)}+\frac{[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}}{[1-p_h(\mathbf{x}_i)]^2}\right],
\end{aligned}$$

for the following functions

- *Logit link*

$$\begin{aligned}
&p_j(\mathbf{x}_i)=\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})},\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}=\frac{x_{im}\exp(\eta_{ij})}{\left[1-\exp(\eta_{ij})\right]^2},\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}=\frac{x_{iy}\exp(\eta_{ij})}{\left[1-\exp(\eta_{ij})\right]^2},\\
&[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}=[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}=\frac{\exp(\eta_{ij})}{\left[1+\exp(\eta_{ij})\right]^2},\qquad \left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_m}=(x_{im})^2\exp(\eta_{ij})\frac{1-\exp(\eta_{ij})}{\left[1+\exp(\eta_{ij})\right]^3},\\
&\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y}=x_{im}x_{iy}\exp(\eta_{ij})\frac{1-\exp(\eta_{ij})}{\left[1+\exp(\eta_{ij})\right]^3},\qquad [\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s}=\exp(\eta_{ij})\frac{1-\exp(\eta_{ij})}{\left[1+\exp(\eta_{ij})\right]^3}\\
&\text{and}\quad [\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h}=x_{im}\exp(\eta_{ij})\frac{1-\exp(\eta_{ij})}{\left[1+\exp(\eta_{ij})\right]^3}.
\end{aligned}$$

- *Complementary log–log link*

$$\begin{aligned}
&p_j(\mathbf{x}_i)=\exp\left[-\exp(\eta_{ij})\right],\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}=x_{im}\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right],\\
&\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}=x_{iy}\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right],\qquad [\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}=[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}=\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right],\\
&\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_m}=(x_{im})^2\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right]\left[1-\exp(\eta_{ij})\right],\\
&\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y}=x_{im}x_{iy}\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right]\left[1-\exp(\eta_{ij})\right],\\
&[\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s}=\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right]\left[1-\exp(\eta_{ij})\right]\\
&\text{and}\quad [\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h}=x_{im}\exp(\eta_{ij})\exp\left[-\exp(\eta_{ij})\right]\left[1-\exp(\eta_{ij})\right].
\end{aligned}$$

- *Log–log link*

$$\begin{aligned}
&p_j(\mathbf{x}_i)=\exp\left[-\exp(-\eta_{ij})\right],\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m}=x_{im}\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right],\\
&\left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y}=x_{iy}\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right],\qquad [\dot{p}_h(\mathbf{x}_i)]_{\gamma_h}=[\dot{p}_h(\mathbf{x}_i)]_{\gamma_s}=\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right],\\
&\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_m}=(x_{im})^2\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right]\left[\exp(-\eta_{ij})-1\right],
\end{aligned}$$

$$\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y} = x_{im}x_{iy}\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right]\left[\exp(-\eta_{ij})-1\right],$$
$$[\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s} = \exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right]\left[\exp(-\eta_{ij})-1\right]$$
$$\text{and}\quad [\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h} = x_{im}\exp(-\eta_{ij})\exp\left[-\exp(-\eta_{ij})\right]\left[\exp(-\eta_{ij})-1\right].$$

- *Probit link*

$$p_j(\mathbf{x}_i) = \Phi(\eta_{ij}),\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_m} = x_{im}\phi(\eta_{ij}),\qquad \left[\dot{p}_j(\mathbf{x}_i)\right]_{\beta_y} = x_{iy}\phi(\eta_{ij}),$$
$$[\dot{p}_h(\mathbf{x}_i)]_{\gamma_h} = [\dot{p}_h(\mathbf{x}_i)]_{\gamma_s} = \phi(\eta_{ij}),\qquad \left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_m} = -(x_{im})^2\eta_{ij}\phi(\eta_{ij}),$$
$$\left[\ddot{p}_j(\mathbf{x}_i)\right]_{\beta_m\beta_y} = -x_{im}x_{iy}\eta_{ij}\phi(\eta_{ij}),\qquad [\ddot{p}_h(\mathbf{x}_i)]_{\gamma_h\gamma_s} = -\eta_{ij}\phi(\eta_{ij})\quad\text{and}\quad [\ddot{p}_h(\mathbf{x}_i)]_{\beta_m\gamma_h} = -x_{im}\eta_{ij}\phi(\eta_{ij}),$$

where $\eta_{ij} = \gamma_j + \mathbf{x}_i^T\boldsymbol{\beta}$, $\Phi(.)$ is the standard normal cumulative distribution function, $\phi(.)$ is the standard normal probability density function, $i = 1,\ldots,n, j = 1,\ldots,k, m, y = 1,\ldots,p$ and $h, s = 1,\ldots,k$.

References

Aranda-Ordaz, F.J., 1983. An extension of proportional-hazard model for grouped data. Biometrics 39, 109–117.

Barreto, M.L., Santos, L.M.P., Assis, A.M.O., Araújo, M.P.N., Farenzena, G.G., Santos, P.A.B., Fiaccone, R.L., 1994. Effect of vitamin A supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in Brazil. Lancet 344, 228–231.

Carrasco, J.M.F., Ortega, E.M.M., Paula, G.A., 2008. Log-Modified Weibull regression models with censored data: sensitivity and residual analysis. Computational Statistics and Data Analysis 52, 4021–4029.

Chalita, L.V.A.S., 1997. Modelos para dados agrupados e censurados. Doctor Thesis, Departamento de Ciências Exatas, University of São Paulo, Brasil (in Portuguese).

Colosimo, E.A., Giolo, S.R., 2006. Análise de Sobrevivência Aplicada. Edgard Blücher, São Paulo.

Colosimo, E.A., Chalita, L.V.A.S., Demétrio, C.G.B., 2000. Tests of proportional hazards and proportional odds models for grouped survival data. Biometrics 56, 1233–1240.

Cook, R.D., 1977. Detection of influential observations in linear regression. Technometrics 19, 15–18.

Cook, R.D., 1986. Assessment of local influence. Journal of the Royal Statistical Society 48, 133–169.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall, New York.

Cox, D.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 34, 187–220.

Doornik, J., 2007. Object-Oriented Matrix Programming Using, 5th ed. Timberlake Consultants Ltd., London.

Escobar, L.A., Meeker, W.Q., 1992. Assessing influence in regression analysis with censored data. Biometrics 48, 507–528.

Hashimoto, E.M., 2008. Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados. Master Thesis, Departamento de Ciências Exatas, University of São Paulo, Brasil (in Portuguese).

Hashimoto, E.M., Ortega, E.M.M., Cancho, V.G., Cordeiro, G.M., 2010. The log-exponentiated Weibull regression model for interval-censored data. Computational Statistics and Data Analysis 54, 1017–1035.

Heitjan, D.F., 1991. Regression with bivariate grouped data. Biometrics 47, 549–562.

Hertz-Piccioto, I., Rockhill, B., 1997. Validity and efficiency of approximation methods for tied survival times in Cox regression. Biometrics 53, 1151–1156.

Kalbfleisch, J.D., Prentice, R.L., 1973. Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267–278.

Lam, K.F., Ip, D., 2003. REML and ML estimation for clustered grouped survival data. Statistics in Medicine 22, 2025–2034.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, 2nd ed. John Wiley & Sons, New Jersey.

Lipsitz, S.R., Laird, N.M., Harrington, D.P., 1990. Using the jackknife to estimate the variance of regression estimators from repeated measures studies. Communications in Statistics: Theory Methods 19, 821–845.

Manly, B.F.J., 1997. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Ortega, E.M.M., Bolfarine, H., Paula, G.A., 2003. Influence diagnostic in generalized log-gamma regression model. Computational Statistics and Data Analysis 42, 165–186.

Ortega, E.M.M., Cancho, V.G., Bolfarine, H., 2006. Influence diagnostics in exponentiated-Weibull regression models with censored data. Statistics and Operations Research Transactions 30, 172–192.

Ortega, E.M.M., Cancho, V.G., Paula, G.A., 2009. Generalized log-gamma regression models with cure. Lifetime Data Analysis 15, 79–106.

Pettitt, A.N., Bin Daud, I., 1989. Case-weighted measures of influence for proportional hazards regression. Applied Statistics 38, 51–67.

Silva, G.O., Ortega, E.M.M., Cancho, V.G., Barreto, M.L., 2008. Log-Burr XII regression models with censored Data. Computational Statistics and Data Analysis 52, 3820–3842.

Thompson Jr., W.A., 1977. On the treatment of grouped observation in life studies. Biometrics 33, 463–470.

Wang, P., Puterman, M.L., Cockburn, I., Le, N., 1996. Mixed poisson regression models with covariate dependent rates. Biometrics 52, 381–400.

Xie, F.-Ch., Wei, B.-Ch., 2007. Diagnostic analysis for log-Birnbaum-Saunders regression models. Computational Statistics and Data Analysis 51, 4692–4706.

Yu, B., Tiwari, R.C., Cronin, K.A., Feuer, E.J., 2004. Cure fraction estimation from the mixture cure models for grouped survival data. Statistics in Medicine 23, 1733–1747.