Detecting influential data in multivariate survival models

Tsirizani M. Kaombe[11] ***Samuel O.M. Manda***[1,2,3]

[1]Department of Mathematical Sciences, Chancellor College, University of Malawi, Zomba, Malawi,
[2]Biostatistics Research Unit, South African Medical Research Council, Pretoria 0001, South Africa,
[3]Department of Statistics, University of Pretoria, Pretoria 0002, South Africa.

Abstract

Statistical techniques for detecting influential data are well developed and commonly used in linear regression, and to some extent in linear mixed-effects models. However, even though the application of multivariate survival models is widely done, the development of diagnostic tools for the models has been scarce. In this paper, we extend the martingale-based residuals and leverage commonly used in univariate survival regression to derive influence statistics for the multivariate survival model. The performance of the proposed influence statistics is illustrated with simulations, and the tools are applied to an analysis of child clustered survival data to identify influential clusters of observations and their effects on the estimate of fixed-effect coefficients.

Keywords: Clustered data; Survival model; Regression coefficients; Group influence.

[1]Corresponding author: e-mail: tkaombe@cc.ac.mw, Phone: +265-999-603-074

1 Introduction

Statistical inference for multivariate Cox proportional hazard model involves use of mixed regression effects, in order to account for dependence between event-times that are clustered into groups, such as families, communities, study sites, or recurrent events (Guo et al., 1994). In such clustered contexts, the independence assumptions for observed survival times that are commonly used for inference in the Cox regression are violated. The observational units from the same cluster share the unobserved covariates, hence the need to use multivariate methods for such survival data. A common model for correlated data is the so called shared frailty model, which introduces dependence in the model by having observations in the same cluster sharing the unobserved random (or frailty) effects (Ripatti & Palmgren, 2000; Manda, 2011).

The estimation methods for the multivariate survival data models are well-established. For example, this is done using penalised fixed effects partial conditional likelihood (Ripatti & Palmgren, 2000; Parner, 2001), marginal likelihood (Manda, 2001; Aalen et al., 2004), L_1 penalised (lasso) method (Goeman, 2010), and profile likelihood (Xu et al., 2009). These are now implemented in several statistical computer packages that either use standard Cox regression software or numerical technique packages, such as Newton-Raphson iterations, EM algorithm, Monte Carlo EM algorithm, and Bayesian MCMC (Ripatti et al., 2002; Manda, 2011). The available software programmes for fitting multivariate models include `the R pakages`: `coxph` (Fox, 2002), `phmm` (Donohue & Xu, 2010), and `lme4` (Bates, 2010). As is the case with other statistical models, for instance linear and linear mixed models, it is critical to undertake residual analysis and diagnosis assessment to identify observations that might exert excessive influence on parameter estimates from the fitted multivariate survival model, resulting into biased estimates (Xiang et al., 2002; Zewotir, 2008).

However, development and application of diagnostic statistics for multivariate survival data models have been lagging behind the application of the models. With

survival data, influence statistics are only available for univariate survival model (Therneau et al., 1990), or for single observations even when a multivariate survival model is fitted to clustered data, and usually with methods that are not analytic but numerically implemented (Tang et al., 2017). This paper concerns the derivation of a detection statistic for influential clusters of observations that could have impact on the estimation of parameters in multivariate survival data models. The performance of the proposed influence statistic is evaluated using simulation studies, and with further application on real child survival data from Malawi. The next section presents the multivariate survival model and reviews influence measures developed for univariate survival data. The third section develops the influence statistics for multivariate survival data model. This is followed by a simulation study and the application. The paper ends with conclusion of findings.

2 The model and review of influence statistics

2.1 Multivariate survival model

Suppose there are M distinct clusters, each with n_i units, $i = 1, 2, ..., M$. Let T denotes a random survival time, with t_{ij} its observed value for j-th unit in i-th cluster, $j = 1, 2, ..., n_i$, and t_i a set of all survival times in cluster i, whereas t be total of all the observed survival times. Further, let X_{ij} denotes a $p \times 1$ covariate vector for unit ij and β the corresponding vector of fixed effects. Also, we have Z_i as a $q \times 1$ vector of cluster-level covariates, that are associated with b_i random coefficient vector consisting of both random intercepts and random slopes. The random effect vectors are usually normally distributed, that is, $b_i \sim MVN(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is $q \times q$ diagonal covariance matrix. The q random effects are a subset of the p fixed effects. Furthermore, let δ_{ij} have value of 1 or 0 depending on whether or not the ij-th unit experienced the event. Conditional on random effects b_i, the hazard of failure for unit j in cluster i, denoted $\lambda_{ij}(t|\beta, b_i)$ (Abrahantes & Burzykowski, 2005; Xu et al.,

2009), is given by:

$$\lambda_{ij}(t|\beta, b_i) = \lambda_0(t) exp(X_{ij}^T \beta + Z_i^T b_i), \tag{1}$$

where $\lambda_0(t)$ is baseline hazard function.

The primary goal of model (1) is to estimate the effects of covariates on the risks for failure times (Guo et al., 1994). This can be done through constructing marginal likelihood, where the focus is on fixed effects and the dependence among failure times is treated as a nuisance, normally through integrating out the random effects (Ripatti & Palmgren, 2000; Manda, 2001). However, in large datasets, the marginal likelihood approach becomes computationally demanding, as the integrals are usually not in closed forms, such that numerical integration is engaged and also each group's random effect has to be separately integrated out from the likelihood (Guo et al., 1994; Manda, 2001). An alternative method is to use joint partial likelihood estimation, which considers a product of conditional likelihood function and the likelihood for random effects, and solve for the maximum likelihood estimators for the fixed and random effects simulateously (Ripatti & Palmgren, 2000). In this paper, we used the joint partial likelihood approach.

Suppose that in cluster i there are r_i failure event-times, each denoted by d_{ij}, so that $r_i = \sum_{j=1}^{n_i} d_{ij}$. Using ideas developed in the original Cox model (Cox, 1992), let $t_{i1} < t_{i2} < ... < t_{ir_i}$ be the r_i ordered observed event-times in cluster i, with corresponding covariates $X_{i1}, X_{i2}, ..., X_{ir_i}$. Also, let $R(t_{il})$, $l = 1, 2, ..., r_i$ be the set of individuals who are at risk of failure at time t_{ij}, called the risk set. For simplicity, in this paper we restrict the multivariate survival model (1) to its special case, called shared frailty model with fixed covariates, and thus we let $Z_i = \{1\}$, in which case $b_i \sim N(0, D)$ (Ripatti & Palmgren, 2000). Then, the contribution to the full partial likelihood by observations in cluster i is given by:

$$L_i(\beta, b_i|t_i, X_{ij}) = \prod_{l=1}^{r_i} [\frac{exp(X_{il}^T \beta + b_i)}{\sum_{s \in R(t_{il})} exp(X_{is}^T \beta + b_i)}]. \tag{2}$$

The full joint partial likelihood of the data and random effects is the product of the conditional likelihood (2) over all clusters and with the likelihood function of the random effects, $f(b_i)$. This is given by:

$$L(\beta, \mathbf{b}|\mathbf{t}, \mathbf{X}) = \prod_{i=1}^{M} \prod_{j=1}^{n_i} [\frac{exp(X_{il}^T\beta + b_i)}{\sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)}] \times (2\pi)^{\frac{-n}{2}} |\mathbf{D}|^{\frac{-n}{2}} exp(-\frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}). \tag{3}$$

The full joint partial log-likelihood function is:

$$l(\beta, \mathbf{b}|\mathbf{t}, \mathbf{X}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}[(X_{ij}^T\beta + b_i) - ln \sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)] + ln[(2\pi)^{\frac{-n}{2}} |\mathbf{D}|^{\frac{-n}{2}}] - \frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}. \tag{4}$$

The score functions for β and $\mathbf{b}$ follow from the log-likelihood (4) and are, respectively, given by:

$$U_\beta = \frac{\partial l(\beta, \mathbf{b})}{\partial \beta} = \sum_{j=1}^{n_i}\sum_{i=1}^{M}[X_{ij} - \frac{\sum_{s\in R(t_{il})} X_{is} exp(X_{is}^T\beta + b_i)}{\sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)}], \tag{5}$$

and

$$U_{\mathbf{b}} = \frac{\partial l(\beta, \mathbf{b})}{\partial \mathbf{b}} = \sum_{j=1}^{n_i}\sum_{i=1}^{M}[1 - \frac{\sum_{s\in R(t_{il})} Z_i exp(X_{is}^T\beta + b_i)}{\sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)}] - \mathbf{b}^T\mathbf{D}^{-1}. \tag{6}$$

The estimates for β and b_i are found by solving the score functions (5) and (6) simultaneously, when they are equated to zero. The values of estimates are computed through numerical algorithms, such as Newton-Raphson method, since the equations (5) and (6) are not in closed form (Ripatti & Palmgren, 2000).

2.2 Influence statistics for univariate survival model

Suppose $\hat{\theta}$ is a set of maximum likelihood estimators of model parameters θ, with θ consisting of β, b_i, D, and other parameters, and let $\hat{\theta}_{(ij)}$ denotes the estimator of θ obtained from the data without j-th observation from i-th cluster. Then, the influence of j-th data record from i-th cluster on the estimator $\hat{\theta}$ is defined as the difference in estimators, $\Delta\hat{\theta}_{ij} = \hat{\theta} - \hat{\theta}_{(ij)}$ (Das & Gogoi, 2015; Cain & Lange, 1984). This can be obtained for each observation by manually deleting the observation from data and obtain the difference in parameter estimates upon refitting the model to the

reduced dataset. Also, for nonlinear models that use iteratave estimation techniques, $\Delta\hat{\theta}_{ij}$ can be manually obtained using one-step iterative approximation, upon removing a data record. But, these approaches are computationally demanding, since the model has to be refitted several times. In that regard, efficient model post-estimation influence statistics that result from fitting the model to data once are developed and made available in literature.

With generalised linear and linear mixed-effects models, where parameter estimators $\hat{\theta}$ are obtained analytically, influence measure $\Delta\hat{\theta}_{ij}$ is a function of model's basic building blocks, i.e. Studentized residuals, error contrast matrix, and inverse of covariance matrix of response variable (Zewotir & Galpin, 2005). In such models, $\Delta\hat{\theta}_{ij}$ is either computed analytically using methods like Cook's distance (Cook, 1977) or it is approximated for one-step ML estimation using updating formulae techniques (Zewotir, 2008; Nobre & Singer, 2011). Others use first-order Taylor series expansion on score function around $\hat{\theta}_{(ij)}$ (Xiang et al., 2002). For Cox proportional hazard (PH) model, the analytic influence techniques such as Cook's distance do not apply, since subjects enter the likelihood as members of various risk sets, such that deleting a data point affects a number of these risk sets other than one (Cox, 1992).

Therefore, various approximations for influence statistics have been developed for univariate survival data. One technique is through first-order Taylor series expansion about a unity weight ϖ_{ij} of an observation in score function, where $\varpi_{ij} = 0$ for a subject that has been removed from data and $\varpi_{ij} = 1$ otherwise (Cain & Lange, 1984). The weights ϖ_{ij} of observations result into a weighted partial likelihood $L(\beta(\varpi_{ij}))$, as well as weighted score function $U_{\beta(\varpi_{ij})}$ for the model. Subsequently, the weighted ML estimators $\beta(\hat{\varpi}_{ij})$ become $\hat{\beta}(1) = \hat{\beta}$ or $\hat{\beta}(0) = \hat{\beta}_{(ij)}$, where $\hat{\beta}_{(ij)}$ is the estimator obtained upon dropping ij-th case in the dataset, and $\hat{\beta}$ the one obtained from full data. Then, using first-order Taylor series expansion about $\varpi_{ij} = 1$, an estimate of influence is given by $\Delta\hat{\beta}_{ij} = \hat{\beta} - \hat{\beta}_{(ij)} = \partial\hat{\beta}/\partial\varpi_{ij}$, which is obtained by solving for

$\partial\hat{\beta}/\partial\varpi_{ij}$ when the score function is equated to zero (Cain & Lange, 1984), as follows:

$$\begin{aligned}(\partial U/\partial\hat{\beta})(\partial\hat{\beta}/\partial\varpi_{ij}) + \partial U/\partial\varpi_{ij} = 0 \\ \therefore \partial\hat{\beta}/\partial\varpi_{ij} = (-\partial U/\partial\hat{\beta})^{-1}\partial U/\partial\varpi_{ij}.\end{aligned} \tag{7}$$

where the likelihood for univariate model is: $L(\beta|\mathbf{t},\mathbf{X}) = \prod_r [\frac{exp(X_{ij}^T\beta)}{\sum_{s\in R(t_{il})}\varpi_{ij}exp(X_{is}^T\beta)}]^{\varpi_{ij}}$, and the weighted score function is first derivative of logarithm of $L(\beta|\mathbf{t},\mathbf{X})$ with respect to β. The approach in equation (7) is also referred to as infinitesimal jackknife measure of influence of a data record on $\hat{\beta}$ (Therneau et al., 1990).

A related method is the score residual, which is a product of a subject's residual and its extremity in covariate value (Therneau et al., 1990). It is given by:

$$v_{ij}(\hat{\beta}) = \int_0^\infty [X_{ijp}(t) - \bar{X}_p(\hat{\beta},t)]dm(t_{ij}), \tag{8}$$

where $m(t_{ij}) = N(t_{ij}) - \int_0^{t_{ij}} Y_{ij}(t)exp(X_{ij}^T(t)\hat{\beta})d\hat{H}_0(t)$ is residual of ij-th unit at time t_{ij}, also called martingale residual, which measures excess number of events; and p denotes number of covariates; while $\bar{X}_p = \frac{\sum X_{ijp}exp(X_{ij}^T\hat{\beta})}{\sum_{s\in R(t_{il})}exp(X_{is}^T\hat{\beta})}$ is the weighted average of covariate X_{ijp} over $R(t_{il})$ risk sets. The measure (8) is used to estimate sensitivity of log-likelihood to infinitesimal displacements of $\hat{\beta}$. Using a weighted partial likelihood, Therneau et al. (1990) showed that the residual (8) is similar to the jackknife measure (7) and that $\partial U/\partial\varpi_{ij} = (v_{ij1}, v_{ij2}, ..., v_{ijp})^T$.

The third method is the augmented or perturbed regression model (Storer & Crowley, 1985; Therneau et al., 1990), which is a one-step update in $\hat{\theta}$ when a single indicator covariate is added to the model. The added covariate has value 1 for ij-th data point and 0 for all other observations (Therneau et al., 1990). The augmented model influence statistic for univariate survival model (Storer & Crowley, 1985) is

given by:

$$\begin{aligned} \hat{\beta}_1 &= \hat{\beta}_0 + I^{-1}(\hat{\beta}_0)\dot{l}(\hat{\beta}_0) \\ \Rightarrow \hat{\beta}_1 - \hat{\beta}_0 &= I^{-1}(\hat{\beta}_0)\dot{l}(\hat{\beta}_0) \\ &= \frac{-I^{-1}(\hat{\beta}_0)\xi_{ij}}{\pi_{ij} - \xi_{ij}^T I^{-1}(\hat{\beta}_0)\xi_{ij}} m(t_{ij}) \end{aligned} \tag{9}$$

where $m(t_{ij})$ is the martingale residual defined along with equation (8), $\xi_{ijp} = \hat{H}_0(X_{ijp} - \bar{X}_p(\hat{\beta}))exp(X_{ij}^T\hat{\beta})$ represents a column vector from matrix $\mathbf{X}$ corresponding to $1's$, $\pi_{ij} = \hat{H}_0(t)(1 - \bar{c}_{ij}(\hat{\beta}))exp(\hat{\beta}^T X_{ij}^T)$ is the diagonal identity matrix with entries 1 throughout, except for the subject that has been removed, which has 0 entry, and c_{ij} is the indicator covariate that has been added to the dataset (Storer & Crowley, 1985).

These methods are related, since they are a function of subject's leverage and residual measures. Moreover, Therneau et al. (1990) showed that the three methods yield similar estimates of influence, but the score residual (8) has a number of advantages, including simplicity of interpretation. For this reason, we applied the score residual approach to derive the influence statistic for the multivariate survival model (1).

3 Proposed influence statistic for multivariate survival model

We notice from Section 2.1 that the estimation of β for model (1) is completed using numerical methods. This means that effect of dropping a cluster on $\hat{\beta}$ can be approximated manually by one-step Newton-Raphson process, through refitting the model to the data for each removal of a cluster. However, this is time-consuming, because it requires refitting the model for each removal of a cluster. We thus propose an extension of the score residual (8) (Therneau et al., 1990), that results from fitting the model to data once, to study influence of clusters on fixed effects estimators from model (1). As for estimates of random effects $\mathbf{b}$, model (1) assumes that b_i are

mutually independent between clusters, hence deleting a cluster will not affect the estimator $\hat{\mathbf{b}}$ for the other clusters. This has been shown for linear mixed-effects models using first-order Taylor-series expansion on score function (Xiang et al., 2002). This study therefore focuses on deriving group influence statistic for fixed effect estimators $\hat{\beta}$, that depend on observations from all clusters.

To study influence for grouped observations, we first define a leverage and a residual for a single unit ij at a given time t_{ij}. The score process (5) derived for the model (1) is essentially a row vector of differences between the individual ij covariate value and the average for the covariates of all individuals at risk at time t_{ij}. In essence, this is analogous to leverage in linear models (Sarkar et al., 2011; Zhang, 2016). For individual ij, we let $r_{ij} = exp(X_{ij}^T\hat{\beta} + \hat{b}_i)$ be a risk score of ij-th unit. Then, at the il-th event time t_{il}, the Schoenfeld residual (or leverage) (Schoenfeld, 1982), denoted by w_{il}, is given by:

$$\begin{aligned} w_{il} &= X_{il} - \frac{\sum_{s\in R(t_{il})} r_{is}X_{is}}{\sum_{s\in R(t_{il})} r_{is}}, \\ &= X_{il} - \bar{X}(\hat{\beta}, \hat{b}_i, t_{il}), \end{aligned} \tag{10}$$

where $r_{is} = exp(X_{is}^T\hat{\beta} + \hat{b}_i)$ is the risk score for unit ij in the risk set $R(t_{il})$, and X_{il} is the covariate vector of the individual experiencing the event at time t_{il}. Further, $\hat{\beta}$ and $\hat{b}_i$ are, respectively, fixed and random effects terms estimated from the log-likelihood (4). In addition, $\bar{X}(.)$ is a vector whose elements are the conditional weighted means of the covariates values for the individuals at risk of event at time t_{ij}. Hence, the dimension of (10) is $1 \times p$ vector corresponding to each ij-th unit in the risk set.

The quantity (10) is also a residual proposed by Schoenfeld (1982) that sums the score processes (5) of units with failure time at each unique event, assuming no ties. Denote $\mathbf{W}_{il}$ as leverages w_{il} for all n_l data points in the risk set and p covariates, then $\mathbf{W}_{il}$ will be $n_l \times p$ matrix. Furthermore, $w_{il} \in [-\infty, +\infty]$, with mean

$E(w_{il}) = E(X_{il}) - E[\bar{X}(\hat{\beta}, \hat{b}_i, t_{il})] = E(X_{il}) - E(X_{il}) = 0$. The value 0 of w_{il} corresponds to observations with intermediate covariates values and are thus close to the weighted average for covariate X_{il}, and hence their leverage on the fitted survival curve is negligible. While large negative and positive values of w_{il} correspond to observations that have unusual covariates values, that are far from the weighted average of X_{il}, and hence they have high leverage on the fitted survival curve (Zhang, 2016).

A residual, on the other hand, means the difference between the observed and fitted outcome. The smaller this is, the better the model's fit for the observation of interest (Aguinis et al., 2013; Zhang, 2016). For survival data, one of the residuals is the martingale, defined along equation (8), which is an estimate of difference in counts of observed and estimated events at each observation time (Therneau et al., 1990). Extending the univariate martingale residual to multivariate survival data model (1), we obtain an $n_l \times 1$ stacked vector of residuals for units in the risk set $R(t_{il})$ given by:

$$m(t_{il}) = N(t_{il}) - \hat{\Lambda}_0(t) exp(X_{il}^T \hat{\beta} + \hat{b}_i)$$

$$\Rightarrow \begin{bmatrix} m(t_{11}) \\ \vdots \\ m(t_{1n_1}) \\ m(t_{21}) \\ \vdots \\ m(t_{2n_2}) \\ \vdots \\ m(t_{M1}) \\ \vdots \\ m(t_{Mn_M}) \end{bmatrix} = \begin{bmatrix} N(t_{11}) - \hat{\Lambda}_0(t) exp(X_{11}^T \hat{\beta} + \hat{b}_1) \\ \vdots \\ N(t_{1n_1}) - \hat{\Lambda}_0(t) exp(X_{1n_1}^T \hat{\beta} + \hat{b}_1) \\ N(t_{21}) - \hat{\Lambda}_0(t) exp(X_{21}^T \hat{\beta} + \hat{b}_2) \\ \vdots \\ N(t_{2n_2}) - \hat{\Lambda}_0(t) exp(X_{2n_2}^T \hat{\beta} + \hat{b}_2) \\ \vdots \\ N(t_{M1}) - \hat{\Lambda}_0(t) exp(X_{M1}^T \hat{\beta} + \hat{b}_M) \\ \vdots \\ N(t_{Mn_M}) - \hat{\Lambda}_0(t) exp(X_{Mn_M}^T \hat{\beta} + \hat{b}_M) \end{bmatrix}. \tag{11}$$

where $\hat{\Lambda}_0(t) = \int_{-\infty}^{t} \lambda_0(h) dh$ is the estimated cumulative baseline hazard. The residual (11) has values in the range $(-\infty, 1]$, because $N(t_{il})$ is either 0 or 1 and $\hat{\Lambda}_0(t) exp(X_{il}^T \hat{\beta} +$

$\hat{b}_i$) has values in the interval $[0, \infty)$. In addition, $E(m(t_{il})) = E(N(t_{il})) - E(\hat{\Lambda}_0(t) exp(X_{il}^T \hat{\beta} + \hat{b}_i)) = E(N(t_{il})) - E(N(t_{il})) = 0$, since the off-minus quantity in (11) is the average number of events.

Both leverage quantity (10) and residual (11) have correlated values for subjects that are in the same cluster due to shared random effect, but independent values between clusters. Due to this property, we utilise the independence of clusters to derive an influence statistic for detecting impact of dropping a cluster on the estimate of β. Influence of an observation on regression parameter estimates is a product of its outlier and leverage values. Many studies, for example (Cook, 1977) for linear models, (Zewotir & Galpin, 2005) for linear mixed-effects models, (Therneau et al., 1990) for univariate survival models, have shown this. Thus, in deriving influence statistics, appropriate case-deletion residual and leverage measures need to be defined first. Using the residual defined in (11) and leverage in (10) for model (1), we propose an analogue of the score residual (8) (Therneau et al., 1990) to measure influence of a cluster on $\hat{\beta}$ for the model (1) as a vector product of values of vector (11) and those of columns of matrix (10) for subjects under risk set $R(t_{il})$ in the same cluster i, given by:

$$v_i(\hat{\beta}) = [m(t_{il})]^T \times \mathbf{W}_{il}. \tag{12}$$

The extended score residual (12) is an $((1 \times n_1) \times (n_1 \times p) ... (1 \times n_M) \times (n_M \times p)) = M \times p$ matrix, as the value $v_1(\hat{\beta})$ for first cluster will be a $(1 \times n_1) \times (n_1 \times p) = 1 \times p$ vector reflecting influence of first cluster on each $\hat{\beta}$ for p covariates, while $v_2(\hat{\beta})$ for second cluster will be a $(1 \times n_2) \times (n_2 \times p) = 1 \times p$ vector, and so forth. The measure (12) will quantify joint influence of observations in a cluster on the estimate $\hat{\beta}$, since each of its components is a measure of joint extremity of cluster observations in terms of survival outcomes, as well as in covariates' values off the fitted survival curve. Since $\mathbf{W}_{il}$ in (12) has elements $w_{il} \in [-\infty, +\infty]$ and $m(t_{il}) \in (-\infty, 1]$, both with mean 0, then the proposed influence statistic (12) is expected to have mean 0.

Large positive value of the proposed statistic (12) means a cluster has majority of subjects that have high positive values in w_{il} that coincide with high positive values in $m(t_{il})$, or large negative values in w_{il} coinciding with large negative values in $m(t_{il})$. Technically, this means the cluster has majority of large positive leverage subjects that experienced more events (i.e. failed too early) than predicted by the model or has most subjects with large negative leverage that survived longer than predicted by the model. Hence, such a cluster requires further investigation.

On the other hand, large negative value of (12) implies that a cluster has majority of subjects that have large positive leverage w_{il} that coincide with large negative values of the residual $m(t_{il})$ or viceversa. In other words, this implies that the cluster has majority of large positive leverage observations that experienced fewer events (i.e. survived longer) than predicted by the model or has majority of large negative leverage subjects that failed too early than predicted by the model. Again, such a cluster will need further investigation. The values of (12) that are close to zero imply most subjects of the corresponding clusters have either leverage close to zero or residual close to zero, hence such clusters have no issues for follow up investigation. To decide on influential groups, some studies in linear mixed-effects models have used a cutoff of $\pm 2/\sqrt{M}$ for the values of the influence statistic (Belsley et al., 2005; Nieuwenhuis et al., 2012). However, graphical methods or relative comparisons of influence values for groups are commonly used (Zewotir & Galpin, 2007). We applied graphical techniques in the next two sections to examine influential clusters to the fitted survival mixed models using the proposed influence statistic (12).

4 Simulation study

In order to evaluate performance of the proposed influence statistic in equation (12), a simulation study was carried out. A shared frailty model given below was assumed. The covariates $X_1 \sim Bernoulli(0.7)$ and $X_2 \sim N(0,1)$, and random effects $b_i \sim N(0, 0.4^2)$ were used. The event-time $T \sim Exp(1)$ was generated from the model

below using cumulative hazard inversion method (Brilleman et al., 2018):

$$\lambda_{ij}(t|b_j, X_{ij}) = \lambda_0(t)exp(\beta_1 X_{ij1} + \beta_2 X_{ij2} + b_i), \tag{13}$$

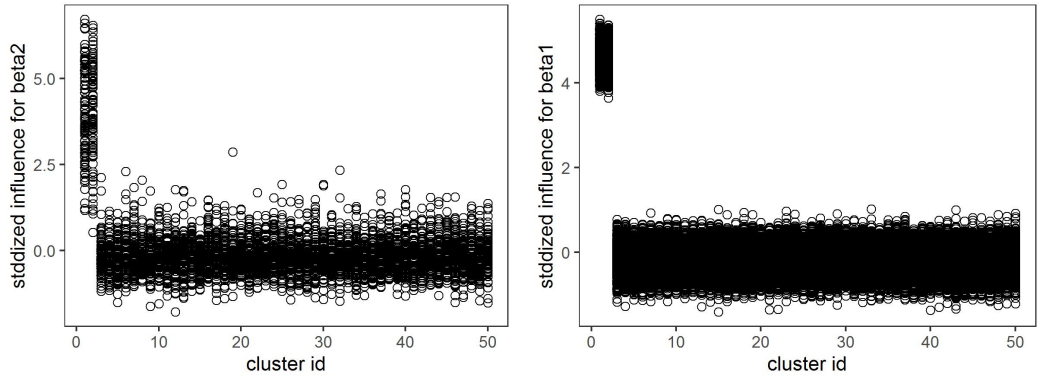
where $\lambda_0(t) \sim Weibull(0.1, 1)$, i.e. $\lambda_0(t) = cdt^{d-1}$, with $c = 0.1$ and $d = 1$ making $\lambda_0(t) = 0.1$; $\beta_1 = 0.5$ and $\beta_2 = 1$. The inversion method derived t_{ij} from $t^*_{ij} = \Lambda^{-1}_{ij}(-log(S(t_{ij})))$, where $S(t_{ij}) \sim Unif(0, 1)$ and hence $\Lambda_{ij}(t) = -log(Unif(0, 1)) \sim Exp(1)$. The random censoring data were generated from $Bernoulli(0.4)$, based on literature (Manda & Meyer, 2005). The `R package simsurv` (Moriña & Navarro, 2014; Brilleman et al., 2018) was used to set up and draw data from the exponential distribution. Samples of size 10, 20, and 50 clusters each were generated. The cluster sizes were set at 80 and 500. This helped to assess effect of both sample size per cluster and number of clusters per dataset on performance of the statistic (12). The sampling in each case was replicated 100 and 1000 times.

The examination process involved simulating regular data set as per model (13), and then perturbing data in first two clusters in each case (Zewotir & Galpin, 2006). The first perturbations involved $\beta_1 = 1.8, 2.7$ followed by $\beta_2 = 2.0, 2.5$, leaving the rest parameters unchanged each time. Then, perturbing jointly $\beta_1 = 1.8, 2.7$ and $\beta_2 = 2.0, 2.5$, leaving b_i intact. The success of the proposed influence statistic (12), in picking cluster 1 or 2 as influential to $\hat{\beta}_1$ or $\hat{\beta}_2$ was evaluated using proportion of correct identification of the two clusters among all 100 or 1000 simulations given a cutoff (Xiang et al., 2002). Upon fitting model (13) to the data, the proposed statistic (12) was computed.

4.1 Results of simulation study

We inspected performance of the derived influence statistic prior to its detailed evaluation. The findings in Figure 1 indicate that the proposed statistic had detected influence of the first two clusters on $\hat{\beta}_1$ and $\hat{\beta}_2$, at varying cutoffs. The values of the statistic were outstandingly higher in the first two clusters, where the coefficients were

perturbed, than in the other clusters. This study therefore assessed success rates of the proposed influence statistic under each simulation scenario using different cutoffs that were informed by prior graphical inspections.



(a) Scatter plots of influence statistic vs cluster id for a case of data with perturbed $\beta_2 = 2.0$ in 2 of 50-clusters sample, each with 80 subjects and with 100 replications

(b) Scatter plots of influence statistic vs cluster id for a case of data with perturbed $\beta_1 = 2.7$ in 2 of 50-clusters sample, each with 500 subjects and with 1000 replications

Figure 1: Plots of cluster influence on $\hat{\beta}_1$ or $\hat{\beta}_2$ under different simulations. Source: Researcher

Table 1 provides success rates of the proposed influence statistic in detecting impact of cluster 1 or 2 on $\hat{\beta}_1$ over 100 and 1000 simulations. The results show that the statistic correctly identified the two influential clusters with high percentage, when the perturbations involved β_1 or β_1 and β_2 jointly. The rates for influence of cluster 1 or 2 on $\hat{\beta}_1$ were relatively low, when it was β_2 that was twirked. The results also show that the sensitivity of the proposed residual improved with cluster sample size, such that the success rates were as high as 100% where cluster size was 500 and lower with varying degrees when cluster size was 80 subjects. In addition, performance of the statistic improved with size of perturbed parameter value, and this was noticeable where cluster sample sizes were low.

It is also shown that performance of the influence statistic was not different between 100 and 1000 simulation sizes, when cluster sample size was 500 subjects. But the success rates generally slumped in 1000 replications, when cluster size was 80. Finally, the results show that the influence statistic was equally effective across different number of clusters per dataset.

Table 1: Percentage of simulations[1] that identified cluster 1 or 2 as influential to $\hat{\beta}_1$

				100 replicates		1000 replicates	
M	n_j	β_1	β_2	%Cluster1	%Cluster2	%Cluster1	%Cluster2
10	80	1.8	1	84	87	60.9	59.4
	80	2.7	1	100	100	99.3	99.2
10	500	1.8	1	100	100	100	100
	500	2.7	1	100	100	100	100
20	80	1.8	1	74	75	46.4	44.9
	80	2.7	1	99	99	95.3	95.1
20	500	1.8	1	100	100	100	100
	500	2.7	1	100	100	100	100
50	80	1.8	1	34	31	10.7	11.8
	80	2.7	1	75	75	52.7	55.3
50	500	1.8	1	100	100	100	100
	500	2.7	1	100	100	100	100
10	80	0.5	2.0	19	22	42	49
	80	0.5	2.5	36	38	32.3	36.3
10	500	0.5	2.0	27	29	13.1	15.1
	500	0.5	2.5	47	39	41.1	38.5
20	80	0.5	2.0	27	25	10.6	13.1
	80	0.5	2.5	27	31	36.9	40.4
20	500	0.5	2.0	29	30	18.9	20.4
	500	0.5	2.5	60	51	43.9	45
50	80	0.5	2.0	30	29	13.2	12.9
	80	0.5	2.5	60	54	43.7	42.8
50	500	0.5	2.0	30	28	23.5	22.1
	500	0.5	2.5	63	62	47.6	48.6
10	80	1.8	2.0	69	77	57.5	59
	80	2.7	2.5	99	96	84.6	83
10	500	1.8	2.0	100	100	100	100
	500	2.7	2.5	100	100	100	100
20	80	1.8	2.0	70	69	43.8	46.2
	80	2.7	2.5	92	92	76.6	74.7
20	500	1.8	2.0	100	100	100	100
	500	2.7	2.5	100	100	100	100
50	80	1.8	2.0	67	51	45.8	44.9
	80	2.7	2.5	86	87	71.9	70.6
50	500	1.8	2.0	100	100	100	100
	500	2.7	2.5	100	100	100	100

[1] No perturbations were done to data in other clusters than 1 and 2, in those other clusters model (18) had $\beta_1 = 0.5$, $\beta_2 = 1$.

The results in Table 2 are for success rates over 100 and 1000 replications for the proposed influence statistic in identifying cluster 1 or 2 as having influence on $\hat{\beta}_2$.

The findings show that the proposed influence statistic highly detected impact of first two clusters on $\hat{\beta}_2$, when it was β_2 or jointly β_2 and β_1 that was perturbed during data generation. The success rates of the statistic in detecting influence of cluster 1 or 2 on $\hat{\beta}_2$ were low when it was β_1 that was perturbed.

As was the case with $\hat{\beta}_1$, the success rates of the statistic in sensing impact of cluster 1 or 2 on $\hat{\beta}_2$ improved with cluster size, as the rates were consitently high for cluster sizes of 500 and low with cluster sizes of 80 subjects. Again, the performance of the statistic improved with size of perturbed parameter value, a situation that was also noticeable in low cluster sizes like before. Likewise, there was no difference in performance of the statistic between 100 and 1000 simulation sizes, this was much apparent in large cluster sample sizes. Lastly, it is also shown that the influence statistic performed equally across different number of clusters per sample.

Table 2: Percentage of simulations[1] that identified cluster 1 or 2 as influential to $\hat{\beta}_2$

				100 replicates		1000 replicates	
M	n_j	β_1	β_2	%Cluster1	%Cluster2	%Cluster1	%Cluster2
10	80	1.8	1	2	2	0.9	0.7
	80	2.7	1	4	4	1.2	1.3
10	500	1.8	1	0	0	0	0
	500	2.7	1	0	0	2.6	2.3
20	80	1.8	1	14	12	4.8	5.5
	80	2.7	1	0.8	0.6	4.6	4.6
20	500	1.8	1	0.9	1.2	1.3	1.4
	500	2.7	1	1	0.8	2	1.2
50	80	1.8	1	34	40	13.7	14.8
	80	2.7	1	34	33	19.6	20.0
50	500	1.8	1	26	18	13.4	11
	500	2.7	1	18	14	8.5	7.4
10	80	0.5	2.0	94	97	93.8	93.5
	80	0.5	2.5	98	100	98.5	97.9
10	500	0.5	2.0	100	100	100	100
	500	0.5	2.5	100	100	100	100
20	80	0.5	2.0	98	98	93.4	92.7
	80	0.5	2.5	100	100	97.7	97.4
20	500	0.5	2.0	100	100	100	100
	500	0.5	2.5	100	100	100	100
50	80	0.5	2.0	99	97	94.4	94.6
	80	0.5	2.5	100	100	97.3	97.6
50	500	0.5	2.0	100	100	100	100
	500	0.5	2.5	100	100	100	100
10	80	1.8	2.0	72	77	39.1	42.5
	80	2.7	2.5	99	92	81.6	81.3
10	500	1.8	2.0	100	100	100	100
	500	2.7	2.5	100	100	100	100
20	80	1.8	2.0	64	73	46.7	45
	80	2.7	2.5	88	81	65.2	63.7
20	500	1.8	2.0	100	100	100	100
	500	2.7	2.5	100	100	100	100
50	80	1.8	2.0	59	53	43.4	43.4
	80	2.7	2.5	78	74	63.6	61.5
50	500	1.8	2.0	100	100	99.9	99.8
	500	2.7	2.5	100	100	100	100

[1] No perturbations were done to data in other clusters than 1 and 2, in those other clusters model (18) had $\beta_1 = 0.5$, $\beta_2 = 1$.

5 Application to clustered survival data from Malawi

We applied the proposed influence statistic on child survival data that were collected as part of 2015-16 Malawi Demographic and Health Survey (MDHS) data. Malawi is a country in south-eastern Africa that borders Tanzania to the north, Zambia to the west, and Mozambique to its east, south, and west. The country is divided into 28 administrative districts, with a total population of just over 17 million people (Malawi National Statistical Office (NSO), 2019). The MDHS was conducted between 19 October 2015 and 18 February 2016 and it collected child survival data from women respondents and caregivers aged 15-49 years, who provided birth histories. The survey used two-stage stratified sampling design, with emuneration areas as primarily sampling units and households as secondary units (Malawi National Statistical Office (NSO) & ICF, 2017). We analyse mortality data for 17,286 children who were born within the last 5 years of the survey. The data are available at `www.DHSprogram.com`.

The analysis used district's rural and urban strata as clusters, totalling to 56 clusters. Child birth order and sex were used as predictors of child mortality, based on findings from previous studies (Manda, 1999). We used death of a child from any cause before 60 months of age as event of interest and age-at-death in months or censoring point as event-time. We transformed into random uniform (0,1) values all zero ages-at-death to reflect proportion of month-days for the event-time. There were 5% of children who experienced the event of death. We censored all children who were still alive during the survey or who had survived up to 60 months. We fitted the Cox frailty model to the data and computed values of the proposed influence statistic for each of the 56 clusters. The fitted frailty model is:

$$\lambda_{ij}(age) = \lambda_0(age)exp(-0.2127 \times Female - 0.0085 \times Birthorder + cluster). \quad (14)$$

The model results showed that female children had significantly lower risk of death than their male counterparts (p-value = 0.0024). Studies attribute this trend

to genetic and biological makeup as well as preconception environments that put male babies to higher risk of suffering from diseases than female children (Pongou, 2013). While children with higher birth order had slight reduced risk of death, but this was not statistically significant (p-value = 0.6100). Other studies have observed that the relationship between birth order and logarithm of child mortality is quadratic and not linear (Manda, 1999). Thus, the insignificant result for birth order could reflect the form in which the covariate was entered in the survival model. The variance of cluster random-effects was 0.0465 and it was significantly different from zero (p-value = 0.003). We proceeded to analyse influence of each cluster on effect of sex on child mortality and ignored effect of birth order due to its insignificance in the model. We used the national under-five mortality rate of 63 deaths per 1000 live births (Malawi National Statistical Office (NSO) & ICF, 2017) as baseline hazard when computing influence values. Upon identifying the influential clusters to the model, we analysed their impact on regression parameter estimates by re-fitting the model to data without the detected clusters and observe the changes in parameter estimates.

5.1 Application results

The results in Figure 2 show that, at a cutoff of ± 2, the proposed influence statistic detected *Kasungu* rural cluster as having outright positive influence on effect of female gender on child mortality. This means that *Kasungu* rural cluster had majority of children with high positive leverage on estimated mortality that had also died too early than predicted by the model, such that dropping this cluster from the model would cause a significant change on estimated effect of female gender on child mortality. It may also imply this cluster had majority of children with high negative leverage on the mortality curve that had also survived longer than predicted by the model. While *Salima* urban cluster was identified as having negative borderline influence on effect of being female on child mortality. This implies that Salima urban cluster had majority of children with high positive leverage on estimated mortality, who survived longer than predicted by the model, such that removing the cluster would impact on

estimated effect of female gender on child mortality. It may also mean the cluster had majority of children with high negative leverage that died too early than predicted by the model. Thus, the two clusters required further investigations.

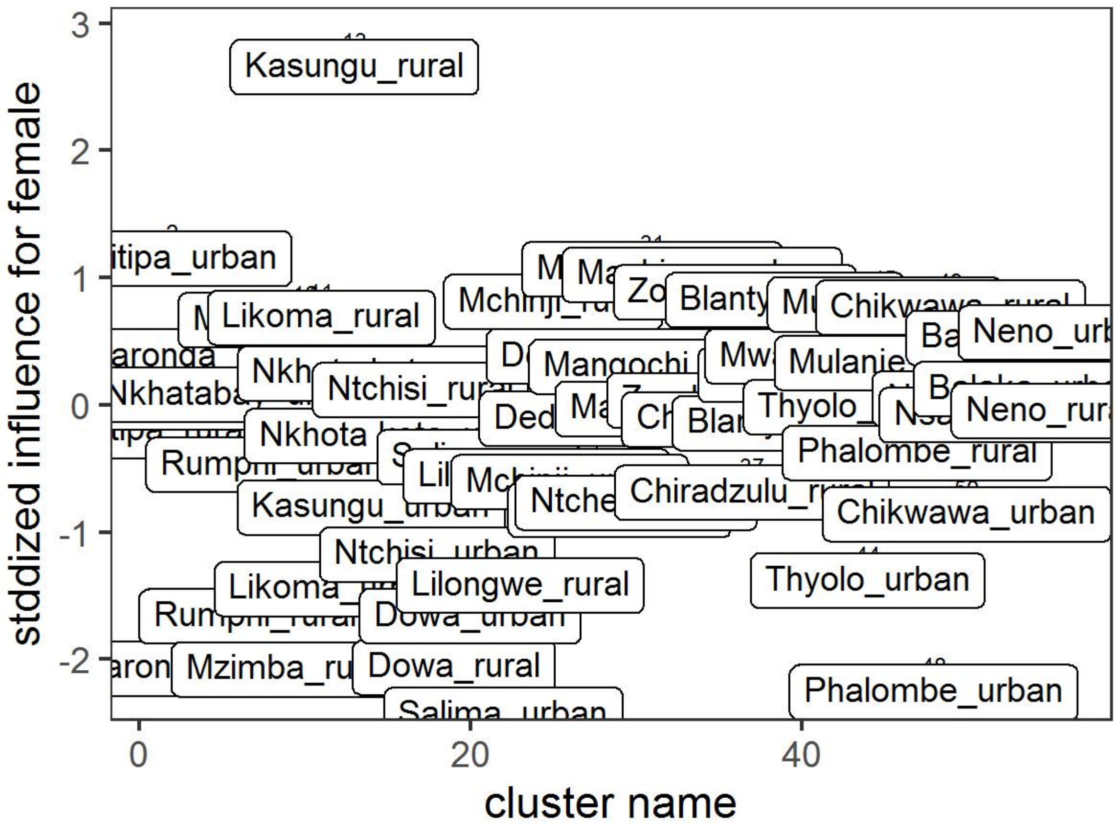


Figure 2: Sub-district level estimates of the proposed influence statistic for effect of female gender upon fitting a frailty Cox hazard regression model to Malawi child survival data, 2015-16 MDHS. Source: Researcher

We performed further analyses to study impact of the idetified influential clusters on regression coefficients. Table 3 shows results for model estimates using full Malawi child survival dataset and the data without each of the two identified influential clusters. The findings show that removal of *Kasungu* rural cluster from analysis resulted in further reduction in hazard of death for female children by 0.0145. We also noticed a reduction in p-value by 0.001. Thus, the survival model was better off without data from *Kasungu* rural cluster. While dropping Salima urban cluster increased the hazard of death in female children by 0.0014. The p-value also got higher upon removing this cluster. This means that the data from Salima urban cluster were required in the model. Removing both clusters from analysis resulted in

reduction of hazard of death, but not as higher as when *Kasungu* rural cluster was dropped alone. Thus, the effect of dropping the two clusters at the same time did not add value to the estimation compared to dropping each one of them separately. This was the case since Kasungu rural cluster had positive influence on effect of being female on mortality, while Salima urban had negative influence. The standard errors of the parameter estimates slightly increased in each case, implying that the original estimates from full data were biased. The variance of random effects also got lower in both cases. Further, the results vindicate the magnitude of influence of each of the two clusters as reported by our proposed statistic in the previous paragraph. It is shown in Table 3 that impact of *Kasungu* rural cluster on the estimate of effect of female gender on mortality was so huge compared to that of *Salima* urban cluster.

Table 3: Estimates of effect of being female on mortality with and without Kasungu rural or Salima urban clusters or both in the Malawi child survival dataset

Parameter	Full data	Without Kasungu rural (diff[1])	Without Salima urban (diff[1])	Without Both (diff[1])
$\hat{\beta}$	-0.2127	-0.2270 (0.0145)	-0.2113 (-0.0014)	-0.2256 (0.0130)
$se(\hat{\beta})$	0.0701	0.0710 (-0.0009)	0.0702 (-0.0001)	0.0711 (-0.0010)
p-value	0.0024	0.0014 (0.001)	0.0026 (-0.0002)	0.0015 (0.0009)
$var(re)$	0.0465	0.0443 (0.0021)	0.0462 (0.0003)	0.0440 (0.0025)

diff[1] = estimate under full data - estimate from reduced data, $se(\hat{\beta})$ is standard error of $\hat{\beta}$, $var(re)$ is variance of random effects.

6 Conclusion

This paper has developed an influence statistic for analysing impact of a cluster of observations on regression coefficient estimates from multivariate survival model. This was accomplished by extending the score residual that was available for univariate survival data. A simulation study has shown that the proposed statistic is very effective in identifying influential clusters to the model's fixed effect estimates. The proposed statistic detects both direction and magnitude of influence of a cluster on regression parameter estimates. Evaluation of the proposed statistic has shown that its application requires no definitive cutoff, but relative comparisons of its values suffice, with large positive or large negative values indicating clusters that require

further investigation (Zewotir & Galpin, 2006).

For the identified influential clusters, one could investigate contribution of individual units in making the clusters as such (Xiang et al., 2002; Zewotir & Galpin, 2006). We recommend such analyses for future research. Further, we recommend that an analysis of multivariate survival data should be accompanied by assessment of influential clusters to avoid having biased estimates and inaccurate conclusions.

Acknowledgements

This work was supported through DELTAS Africa Initiative, SSACAB, Grant:107754/Z/15/Z. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences' (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK government. The views expressed in this publication are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government. We are sincerely grateful to the World Bank's Skills Development Project (SDP) at Chancellor College, University of Malawi for initial funding towards this work.

Conflict of Interest

The authors have declared no conflict of interest.

References

Aalen, O. O., Fosen, J., Weedon-Fekjær, H., Borgan, Ø., & Husebye, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics*, *60*(3), 764–773.

Abrahantes, J., & Burzykowski, T. (2005). A version of the em algorithm for proportional hazard model with random effects. *Biometrical Journal*, *47*(6), 847–862.

Aguinis, H., Gottfredson, R., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with r.* New York: Springer.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571). New York: John Wiley & Sons.

Brilleman, S., Rory, W., Moreno-Betancur, M., & Crowther, M. (2018). simsurv: A package for simulating simple or complex survival data. In *UseR! Conference 2018, Brisbane, Australia.*

Cain, K., & Lange, N. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, *40*(2), 493–499.

Cook, D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cox, D. R. (1992). *Regression models and life-tables. Breakthroughs in statistics.* New York: Springer.

Das, M., & Gogoi, B. (2015). Influential observations and cutoffs of different influence measures in multiple linear regression. *International Journal of Computational and Theoretical Statistics*, *2*(2), 79–85.

Donohue, M., & Xu, R. (2010). phmm: Proportional hazards mixed-effects model (phmm). *R package version 0.6*, *3*.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS Companion to Applied Regression*, *2002*.

Goeman, J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, *52*(1), 70–84.

Guo, S., , & Lin, D. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, *50*(3), 632–639.

Langford, I., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*(2), 121–160.

Malawi National Statistical Office (NSO). (2019). *2018 malawi population and housing census: Main report.* Zomba: Author.

Malawi National Statistical Office (NSO), & ICF. (2017). 2015-16 malawi demographic and health survey: Key findings. *Zomba, Malawi, and Rockville, Maryland, USA: Author*.

Manda, S. (1999). Birth intervals, breastfeeding and determinants of childhood mortality in malawi. *Social Science & Medicine*, *48*(3), 301–312.

Manda, S. (2001). A comparison of methods for analysing a nested frailty model to child survival in malawi. *Australian & New Zealand Journal of Statistics*, *43*(1), 7–16.

Manda, S. (2011). A nonparametric frailty model for clustered survival data. *Communications in Statistics-Theory and Methods*, *40*(5), 863–875.

Manda, S., & Meyer, R. (2005). Age at first marriage in malawi: a bayesian multilevel analysis using a discrete time-to-event model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(2), 439–455.

Moriña, D., & Navarro, A. (2014). The r package survsim for the simulation of simple and complex survival data. *Journal of Statistical Software*, *59*(2), 1–20.

Nieuwenhuis, R., te Grotenhuis, H., & Pelzer, B. (2012). Influence. me: tools for detecting influential data in mixed effects models: Retrived from https://repository.ubn.ru.nl/handle/2066/103101.

Nobre, J., & Singer, J. (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics*, *38*(5), 1063–1072.

Parner, E. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, *28*(2), 295–302.

Pongou, R. (2013). Why is infant mortality higher in boys than in girls? a new hypothesis based on preconception environment and evidence from a large sample of twins. *Demography*, *50*(2), 421–444.

Ripatti, S., Larsen, K., & Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis*, *8*(4), 349–360.

Ripatti, S., & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, *56*(4), 1016–1022.

Sarkar, S., Midi, H., & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*, *11*(1), 26–35.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Storer, B. E., & Crowley, J. (1985). A diagnostic for cox regression and general conditional likelihoods. *Journal of the American Statistical Association*, *80*(389), 139–147.

Tang, A.-M., Tang, N.-S., & Zhu, H. (2017). Influence analysis for skew-normal semiparametric joint models of multivariate longitudinal and multivariate survival data. *Statistics in medicine*, *36*(9), 1476–1490.

Therneau, T., Grambsch, P., & Fleming, T. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160.

Xiang, L., Tse, S.-K., & Lee, A. H. (2002). Influence diagnostics for generalized linear mixed models: applications to clustered data. *Computational Statistics & Data Analysis*, *40*(4), 759–774.

Xu, R., Vaida, F., & Harrington, D. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, *19*, 819–842.

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in Statistics-Theory and Methods*, *37*(7), 1071–1084.

Zewotir, T., & Galpin, J. (2005). Influence diagnostics for linear mixed models. *Journal of data science*, *3*(2), 153–177.

Zewotir, T., & Galpin, J. S. (2006). Evaluation of linear mixed model case deletion diagnostic tools by monte carlo simulation. *Communications in Statistics-Simulation and Computation*, *35*(3), 645–682.

Zewotir, T., & Galpin, J. S. (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. *Test*, *16*(1), 58–75.

Zhang, Z. (2016). Residuals and regression diagnostics: focusing on logistic regression. *Annals of Translational Medicine*, *4*(10), 1–8.