IDENTIFYING OUTLYING AND INFLUENTIAL CLUSTERS IN THE ANALYSIS OF MULTIVARIATE SURVIVAL DATA

PhD (BIOSTATISTICS) THESIS

TSIRIZANI MWALIMU KAOMBE

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

AUGUST, 2020

IDENTIFYING OUTLYING AND INFLUENTIAL CLUSTERS IN THE ANALYSIS OF MULTIVARIATE SURVIVAL DATA

PhD (Biostatistics) Thesis

By

TSIRIZANI MWALIMU KAOMBE

MSc. Biostatistics (2012), University of Malawi

Submitted to the Department of Mathematical Sciences, Faculty of Science, in fulfillment of the requirements for the degree of Doctor of Philosophy (Biostatistics)

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

AUGUST, 2020

Declaration

I, the undersigned, hereby declare that this thesis is my own original work, which has not been submitted to any other institution for similar purposes. Where other people's work has been used, acknowledgements have been made.

Tsirizani Mwalimu Kaombe

Signature

Date

Certification

The undersigned certifies that this thesis represents the student's own work and effort, and has been submitted with my approval.

Signature:——————————— Date:———————

Samuel Manda, PhD (Professor)

Supervisor

Dedication

To my late brother, Christopher Malamula.

Acknowledgements

I firstly owe many thanks to my supervisor, Prof. Samuel Manda for his guidance, reassurance and assistance throughout the time of this PhD work. Despite his busy schedule, he was there every time I needed technical support on my inquiries and questions.

I also gratefully acknowledge the funding received for my PhD from the World Bank's Skills Development Project (SDP) at Chancellor College, University of Malawi. Likewise, I am thankful to the funding received through the Developing Excellence in Leadership Training and Science (DELTAS)-Sub-Saharan Africa Consortium for Advanced Biostatistical Training (SSACAB), that enabled me to attend and present at the South Africa Statistical Association (SASA) conferences in 2016 and 2018 held in South Africa, respectively.

To the Chancellor College management, I am thankful for allowing me space to travel for conferences, where I learnt a lot from the international statistical community. I am also grateful to the coordinator of the PhD-Statistics programme, Dr Jupiter Simbeye for his services during my studies. Similarly, I am indebted to the language editor, Dr Margaret Wazakili for proofreading this PhD thesis.

Finally, I appreciate the moral support received from my wife, Patricia. To my children: Pythagoras, Vanessa, Leticia, Mayankho, and Sifa, many thanks for your patience during my studies. You deserved my full parental attention, which was not the case due to my study commitments, thank you. Thanks also to all my relatives and friends for the encouragement throughout my studies.

Abstract

Outlier and influence statistics play an important role in assessing individual or grouped observations that may have undue impact on the parameter estimates of a statistical model. The methods are well-developed for linear and linear mixed-effects models, and are easily implemented in most statistical packages. Though similar statistics exist for univariate survival models, not much has been done for models of multivariate survival data. The objective of this PhD work was to derive outlier and influence statistics for multivariate survival data models, by extending limited research work on such statistics for linear mixed-effects and univariate survival models. The derived statistics were evaluated using simulation studies and illustrated with an analysis of child survival data in Malawi, which had 56 sub-districts (clusters), from both rural and urban areas. The proposed statistics had a high performance of well over 90% in identifying correctly the outlying or influential clusters, and the performance improved with increasing cluster size. In the application to clustered survival data, mostly off-city clusters were identified as having a different child survival pattern and impactful on regression coefficients and variance estimates. This study recommends incorporating outlier and influence assessments when analysing clustered survival data, otherwise the estimates of both regression slopes and variance components could be biased.

Contents

List of Tables

List of Figures

List of Abbreviations

AFT	Accelerated Failure-Time Model
BLUP	Best Linear Unbiased Predictor
CD	Cook's Distance
CP	Coverage Probability
DFBetas	Difference in Beta Standardised
DFFITS	Difference in Fit Standardised
EM	Expectation Maximisation
GLMM	Generalised Linear Mixed-Effects Model
LMM	Linear Mixed-Effects Model
LRT	Likelihood Ratio Test
LSE	Least Squares Estimation
MAD	Median of Absolute Deviations
MBP	Maximization By Parts
MCD	Minimum Covariance Determinant
MCMC	Monte Carlo Markov Chain
MDFFITS	Multivariate Difference in Fits Standardised
MDHS	Malawi Demographic and Health Survey
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MR	Reduced Model
MS	Saturated Model
MSE	Mean Squared Error
PH	Proportional Hazards

PHMM	Proportional Hazards Mixed-Effects Model
Q-Q	Quantile Quantile
REML	Restricted Maximum Likelihood
RMCD	Re-Weighted Minimum Covariance Determinant
SEA	Standard Enumeration Area

List of Notations and Symbols

M	the number of clusters in the study
Subscript 'j'	the j-th cluster
n_j	the number of observations in cluster j
$n = \sum_{j=1}^{M} n_j$	the number of observations in the study
Subscript 'i'	the i-th observation
p	the number of fixed effects
q	the number of random effects
T	event time random variable
$\mathbf{t}$	$n \times 1$ vector of observed values of the event-time variable T
$\mathbf{t}_j$	$n_j \times 1$ vector of event-times from the j-th cluster
t_{ij}	the observed event-time for i-th subject in j-th cluster
$\mathbf{y}$	$n \times 1$ vector of observed values of the response variable Y
$\mathbf{y}_j$	$n_j \times 1$ vector of responses from the j-th cluster
$\mathbf{y}_{(j)}$	$(n - n_j) \times 1$ vector $\mathbf{y}$ without observations from the j-th cluster
Y_i	the i-th subject observed response
$\mathbf{y}_{(i)}$	$(n-1) \times p$ vector $\mathbf{y}$ without the observation from i-th subject
$\hat{\mathbf{y}}$	$n \times 1$ vector of predicted or fitted values of the response variable Y
$\mathbf{X}$	$n \times p$ observed design matrix associated with the p fixed effects
$\mathbf{X}_j$	$n_j \times p$ matrix $\mathbf{X}$ from the j-th cluster only
$\mathbf{X}_{(j)}$	$(n - n_j) \times p$ matrix $\mathbf{X}$ without observations from the j-th cluster
$\mathbf{X}_{(i)j}$	$(n_j - 1) \times p$ matrix $\mathbf{X}$ without the i-th subject
$X_i^T = (X_{i1} X_{i2} ... X_{ip})$	the i-th observation fixed effect value
$X_{ij}^T = (X_{ij1} X_{ij2} ... X_{ijp})$	the i-th subject within the j-th cluster fixed effect value

β	$p \times 1$ vector of unknown fixed effects with design matrix $\mathbf{X}$
$\hat{\beta}$	$p \times 1$ vector of estimated value of β from n observations
$\hat{\beta}_{(i)}$	$p \times 1$ vector of estimated value of β without the i-th subject
$\hat{\beta}_{(j)}$	$p \times 1$ vector of estimated value of β without the j-th cluster
$\mathbf{Z}$	$n \times q$ observed design matrix of covariates that have random effects
$\mathbf{Z}_j$	$n_j \times q$ matrix of cluster-level covariates in the j-th cluster
$\mathbf{Z}_{(j)}$	$(n - n_j) \times q$ matrix $\mathbf{Z}$ without observations from the j-th cluster
$\mathbf{Z}_{(i)j}$	$(n_j - 1) \times q$ matrix $\mathbf{Z}_j$ without the i-th subject
$Z_i^T = (Z_{i1} Z_{i2} ... Z_{iq})$	the random effect covariate value for i-th subject
$Z_{ij}^T = (Z_{ij1} Z_{ij2} ... Z_{ijq})$	random effect covariate value for i-th subject in j-th cluster
$\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_M)$	$n \times q$ matrix of random effects with design matrix $\mathbf{Z}$
b_j	the $q \times 1$ j-th cluster unknown random effect
$\hat{b}_j$	the $q \times 1$ j-th cluster estimated random effect
$\hat{\mathbf{b}}$	$n \times q$ vector of predicted values of $\mathbf{b}$ from n observations
$\hat{\mathbf{b}}_{(i)}$	$q \times 1$ vector of predicted value of $\mathbf{b}$ without the i-th subject
$\hat{\mathbf{b}}_{(j)}$	$q \times 1$ vector of predicted value of $\mathbf{b}$ without the j-th cluster
ϵ	$n \times 1$ vector of unknown random errors
ϵ_i	i-th subject unknown random error
$\hat{\mathbf{e}}$	$n \times 1$ vector of estimated values of the error term from n subjects
$\hat{e}_i$	the $i-$ subject estimated error
$\hat{\epsilon}_j$	$n_j \times 1$ vector of estimated values of the error term from j-th cluster
$\hat{\epsilon}_{(i)}$	$(n-1) \times 1$ vector of estimated errors without i-th subject
$\hat{\epsilon}_{(j)}$	$(n - n_j) \times 1$ vector of estimated errors without j-th cluster
$\mathbf{D}$	$q \times q$ diagonal matrix with identical entries for each level of $\mathbf{b}$
$\mathbf{I}_n$	$n \times n$ identity matrix
$I(\beta)$	the Fisher information matrix for β
σ_ϵ^2	the unknown variance for the error term
σ_ϵ	the unknown standard deviation for the error term
$\hat{\sigma}_\epsilon^2$	the estimated variance of the error term from n subjects

$\hat{\sigma}^2_{\epsilon(i)}$	the estimated variance of the error without i-th subject
$\hat{\sigma}^2_{\epsilon(j)}$	the estimated variance of the error without j-th cluster
θ	diagonal elements of $\mathbf{D}$, given by $(\sigma^2_j/\sigma^2_\epsilon)I_{qj}$ for cluster j
$\mathbf{G} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma^2_\epsilon \mathbf{I}$	$n \times n$ diagonal matrix for covariance of Y in mixed model
$\mathbf{G}_{(i)}$	diagonal matrix $\mathbf{G}$ for variance components without subject i
$\mathbf{G}_{(j)}$	diagonal matrix $\mathbf{G}$ for variance components without cluster j
$\mathbf{g}_i$	the i-th row of matrix $\mathbf{G}$
g_{ii}	the i-th diagonal element of matrix $\mathbf{G}$
$\mathbf{C} = \mathbf{G}^{-1}$	the inverse of matrix $\mathbf{G}$
$\mathbf{C}_i$	the i-th column of matrix $\mathbf{G}^{-1}$
$\mathbf{C}_{(j)}$	the i-th column of matrix $\mathbf{G}^{-1}$ without j-th cluster
c_{ii}	the i-th diagonal element of matrix $\mathbf{G}^{-1}$
$h_{ij}(t)$	hazard of failure for i-th subject in j-th cluster
$H_{ij}(t)$	cumulative hazard function of subjects at observed event time t
$\hat{H}_{ij}(t)$	estimated cumulative hazard function at observed event time t_{ij}
$h_0(t)$	baseline hazard function of subjects
$\hat{H}_0(t)$	estimated cumulative baseline hazard function of subjects
$S_{ij}(t)$	survival function or probability of subject not failing before time t
$\hat{S}_{ij}(t)$	estimated survival function from the model
$S_0(t)$	baseline survival function
$\hat{S}_0(t)$	estimated baseline survival function
$\mathbf{v}$	$m \times 1$ vector of estimated group score residuals
v_j	the j-th element of vector v corresponding to cluster j
$\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$	$n \times n$ hat (projection/leverage) matrix in linear model
$\mathbf{W}_{(i)}$	$(n-1)\times(n-1)$ hat (projection) matrix without i-th observation
$\mathbf{W}_{(j)}$	$(n-n_j)\times(n-n_j)$ hat (projection) matrix without j-th cluster
w_{ii}	the i-th subject's leverage value or i-th diagonal element of $\mathbf{W}$
Δ	the unobserved censoring status of a subject
δ_{ij}	the observed censoring status for i-th subject in j-th cluster
$\mathbf{d}$	$n_j \times 1$ vector of deviance residuals n subjects

d_i	the deviance residual for i-th subject
d_{ij}	the deviance residual for i-th subject in j-th cluster
$d(a,b)$	the distance function between any two points, a and b, in a plane
U_β	the score function for β
U_b	the score function for random effects b
$\mathbf{m}$	the $n_j \times 1$ vector of martingale residual
$m(t_i)$	the martingale residual for i-th observation
$m(t_{ij})$	the martingale residual for i-th subject in cluster j
$L(.)$	the likelihood function
$l(.)$	the logarithm of likelihood function
$\mathbf{k}$	$m \times 1$ vector of within-cluster variance of deviance residual
k_j	the j-th element of vector $\mathbf{k}$ corresponding to j-th cluster
$\mathbf{k}^*$	$m \times 1$ vector of group outlier statistics
k_j^*	the j-th element of vector $\mathbf{k}^*$ corresponding to cluster j
a_i	the estimated Pena's statistic for i-th subject
λ	$n \times 1$ vector of studentised residual
λ_i	i-th subject studentised residual
$\varphi(X_i)$	$n \times 1$ vector of success probability for subjects with values X_i
ϖ_i	the weight of i-th subject in a statistic being computed

Chapter 1

Introduction

This chapter presents the research problem for this study as well as the study objectives. It further provides the outline of this thesis.

1.1 Background

Mixed-effects models have long been applied in statistics to describe heterogeneity in outcome variable data, when making inferences (McGilchrist & Aisbett, 1991; McGilchrist, 1993; Donner & Klar, 1994; Yau, 2001; Bienias et al., 2002; Glidden & Vittinghoff, 2004; Xu et al., 2009; Turkan & Toktamis, 2012). The commonly applied regression models, such as generalised linear model (glm) assume that a large portion of variation in the outcome variable is explained by the use of some fixed covariates in the model. In addition, the glm assumes that the realisations of the response variable are statistically independent. Both of these assumptions may not be entirely correct at certain times.

There are situations where the data have apparent clustering or grouping that can induce dependences of outcome observations from the same cluster. This may cause both the independence and variance-accounting assumptions of the model to be unrealistic. The generalised linear mixed-effects model (glmm) solves these shortcomings. This model maximises utilisation of complex data by making model inferences that account for variation in the responses attributable to subjects' clustering (Ziegler et al., 1998; Galbraith et al., 2010).

The mixed-effects model, also referred to as multi-level model involves specifying a probability distribution for the observation errors at first stage and another distribution for parameters called random effects in the model at subsequent higher stage (Laird & Ware, 1982; Langford & Lewis, 1998). A stage or level is defined as a unit of analysis, this can be a subject or a cluster of subjects. The random parameters belong to higher level and are assumed to vary across clusters or groups, as the observed clusters are a random sample from all clusters in the population.

Thus, the model assumes there are interactions of fixed covariates with subjects' group effects called random covariates, which are also supposed to be estimated by the model, hence the term mixed-effects model (Langford & Lewis, 1998). The random effects can be predicted for each group of subjects in the model. But the focus of the mixed-effects model is usually on measuring variation in outcome variable in the model that is contributed by the data clustering, known as variance components (Laird & Ware, 1982). Therefore, an explanatory variable enters the mixed-effects model as fixed or random effects variable.

In the context of a mixed-effects model, a fixed-effect predictor is a variable that the analyst expects to have effect on the response variable. By 'fixed' it means the variable is not random in the population, it is measured without errors. For example, 'source of drinking water' for a household can be a fixed-effect variable in a model that predicts the 'diarrhoea' outcome in children aged below five years. The odds of suffering from diarrhoea for a child whose household drinks piped water, for instance, are regarded as fixed in the population and the model tries to estimate these odds.

While random-effects variables are often the grouping or classification factors of observations for which the study tries to control their impact on the estimation of fixed effects. An example for the same diarrhoea model is the variable ‘village where a child lives’. A village is a discrete variable that may classify a location for a group of children. In this regard, the analyst may not be interested in the impact of a ‘child’s village’ on the ‘diarrhoea’ outcome, but how much variation in ‘diarrhoea’ outcome in the model is attributable to the ‘child’s village’ factor, when predicting effect of ‘source of drinking water’ on the diarrhoea outcome.

There are counterpart forms of the generalised linear mixed-effects model for survival data. Survival analysis deals with modelling of an outcome variable that reflects duration of time from some defined baseline, such as admission of a patient into hospital, until occurrence of some defined event, such as discharge from the hospital. When the duration of time is directly modelled on some covariates, the survival model is called accelerated failure time (AFT) model (Chiou et al., 2014). Alternatively, the rate of occurrence of the event, referred to as hazard rate, can be modelled as a function of the covariates, which is the case with Cox proportional hazard (PH) model (D. R. Cox, 1972).

The Cox PH model assumes that the hazard ratio for two subjects with different measurements on some covariate is a fixed proportionality term that is free of time. Thus, the covariates have proportional effects on the hazard function over time. This model is handy to implement in most statistical packages and easy to interpret, but like a generalised linear model, it assumes that event-times of subjects are independent (Xue & Schifano, 2017). So, it does not account for clustering of subjects, which can lead to biased estimates of the fixed effects due to possible under-estimation of their variances and standard errors (Liang & Zeger, 1993; Manda, 2011). For this reason, mixed-effect survival models are used where the data have some clustering structures (Guo et al., 1994; Liang et al., 1995;

Vaida & Xu, 2000).

The mixed-effect survival model can estimate fixed effects, predicted values of random effects for each cluster, and amount of variation in survival times attributable to clustering of data. A simple case is the shared frailty model, which incorporates a frailty term in the model that estimates single cluster-specific random effects shared by subjects in the same cluster (Ripatti & Palmgren, 2000; Ha et al., 2011). The shared cluster-effect represents unique features of the cluster that can affect its baseline risk to the event of interest. The use of cluster-specific frailties in the model comes from the fact that different clusters have different dispositions to failure that can cause subjects in some clusters to be more vulnerable to failure or be more frail compared to other clusters (Vaupel et al., 1979).

In recent years, the mixed-effects survival model has become an ideal choice for analysts to account for clustering of data, when applying a survival model to various designs of clustered survival data. These designs include multi-centre clinical trials (Ha et al., 2011), complex surveys (Manda, 2011), and longitudinal studies (Król et al., 2017). However, there is paucity of literature on the critical examination of the impact of unusual clusters on the inferences that can be drawn from the survival mixed model. Due to the uniqueness of subjects in different clusters, some clusters may be outliers to the mixed-effects model or may have large influence on parameter estimates in the model compared to others (Zewotir, 2008).

Moreover, the mixed-effects model is reportedly sensitive to outliers (Zewotir & Galpin, 2005; Turkan & Toktamis, 2012). This implies that ignoring outliers and influential observations assessments would cost the conclusions hugely, when applying the mixed-effects model on data that have some unusual subjects. This may apply to a survival mixed model. There have been advancements in parameter estimation for the survival mixed-effects model, for example using penalised partial likelihood method (Ripatti & Palmgren, 2000) or marginal partial likelihood technique (Manda, 2001) or the L_1 penalised (lasso) method (Goeman, 2010) among others. Nonetheless, little effort has been made to devise diagnostic assessment methods for the survival mixed model, especially the analysis of cluster outliers and influence.

The term 'outlier', in the context of this study, means a response value that is exceedingly large or small compared to others, when viewed from the fitted line or curve (Sarkar et al., 2011; Aguinis et al., 2013; Z. Zhang, 2016). Often times, outliers are indicative of some unusual process in the data. For example, a community with very tall inhabitants due to genetic factors would report very unfamiliar heights of subjects from the rest communities in a study that is recording height of subjects in the population. The model outliers are sometimes a result of data transcription errors. Whatever the cause for outlierness of a data point is, outliers are important data in the modelling process as they have a bearing on the appropriateness of assumptions made on the model's error variables. A more general term for model diagnostic statistic is 'residual', which simply means the difference between the observed and fitted outcomes. The smaller this is, the better the model's fit for the observation of interest (Aguinis et al., 2013; Z. Zhang, 2016).

Associated with the concept of *residual*, is a measure called 'leverage', which reports usefulness of a subject to the model-fit. The leverage of an observation is the distance of the subject's covariate value from an average of the values for that covariate (Sarkar et al., 2011; Z. Zhang, 2016). In consequence, subjects with very large or small covariate values have more leverage than those with intermediate values (Z. Zhang, 2016). Since a regression line or surface is a linear combination of covariates' values mapping to the mean of the response variable, a large leverage subject will pull the fitted line to pass closer compared to a small leverage subject. However, inference-specific importance of the subject to the model is analysed through a quantity called ‘influence’. This measures the effect of dropping a data point on the model’s inferences, such as fitted values, regression coefficients or likelihood (Das & Gogoi, 2015). It is a function of the outlier and leverage.

The measures described above are the focus of this study, especially in the context of clustered survival data. It is important to note that the implementation units for most national health policies in African states are provinces, districts, and communities. Thus, it is necessary to study methods of flagging outlying and influential communities with regard to various health outcomes of subjects, as these would help stakeholders in public health to plan easily for targeted implementation of the health policies.

As discussed in previous paragraphs, the biomedical field may involve studying recurrent events (Król et al., 2017), hence flagging outlying or influential groups of patients may guide researchers on future treatment options for unusual groups. In multi-centre clinical trials involving grouped health outcomes, for example, knowing outlying or influential communities may help in formulating targeted actions for the most vulnerable communities (Ha et al., 2011). When outlying groups are due to measurement errors, as can be observed during interim analyses in randomised controlled trials, the diagnostic assessment for groups can help in giving timely advice to the data management team to be cautious during the data collection phase of the clinical trials.

1.2 Preliminaries of multivariate survival data analysis

Multivariate survival data arise in different ways. For example, through clustered survival data, where failure-times of subjects from the same cluster are observed (Guo et al., 1994). These can be found in multi-centre randomised controlled trials, where each centre involves a number of participants (Glidden & Vittinghoff, 2004; Legrand et al., 2006; Ha et al., 2011), and family genetic studies, where members of the same family form a group (Xu, 2004; Maia et al., 2014). The other way is through recurrent events data, in which an individual may experience the event of interest and of the same type more than once. This could be re-hospitalisation data for patients of some chronic disease, such as diabetes (Król et al., 2017). In such scenarios, the interest of an analyst may be to study the variability of subjects' survival times across clusters (Xu, 2004). This study concerns with estimating associations between certain covariates and survival times, while taking into account the existing dependences among the survival times.

Suppose there are M distinct clusters, each with n_j subjects, $(j = 1, 2, ..., M; i = 1, 2, ..., n_j)$. Let T denotes a survival time random variable with t_{ij} its observed value for i-th subject in j-th cluster. Further, let X_{ij} denotes the $p \times 1$ covariate vector for fixed effect and β the corresponding $p \times 1$ vector of fixed effect coefficients, thus $X_{ij}^T = (X_{ij1} X_{ij2} ... X_{ijp})$ is a transpose of the vector of covariate values for X_{ij} for i-th subject within the j-th cluster. Furthermore, let δ_{ij} take the value of 1 or 0 depending on whether or not the subject experienced the event. Also, assume that each j-th cluster has specific $q \times 1$ random effects (or frailty) $b_j^T = (b_{j1} b_{j2} ... b_{jq})$ with Z_{ij}, $q \times 1$ vector of the covariates with the random effects, so that $Z_{ij}^T = (Z_{ij1} Z_{ij2} ... Zijq)$ becomes cluster covariate value for $i-$th subject in $j-$th cluster. The observed survival times t_{ij} for subjects i in cluster j are assumed to be conditionally independent, given the covariates X_j and random effect b_j (Skrondal & Rabe-Hesketh, 2009). Conditional on vector of cluster-specific random effect b_j, the hazard of failure for subject i in cluster j at time t, denoted $h_{ij}(t|\beta, b_j)$ (Abrahantes & Burzykowski, 2005; Xu et al., 2009) is given by:

$$h_{ij}(t|\beta, b_j) = h_0(t) exp(X_{ij}^T \beta + Z_{ij}^T b_j) \tag{1.1}$$

where $h_0(t)$ is unspecified baseline hazard function. The assumption with random effects is that b_j's are identically and independently distributed random variables from a distribution known up to a finite number of parameters. For example, random effects could be assumed to have multivariate normal distribution, i.e. $b_j \sim N(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is $q \times q$ diagonal covariance matrix with identical entries for each level of b_j.

As a consequence of (1.1), the corresponding integrated hazard function is:

$$\begin{aligned} H_{ij}(t|\beta, b_j) &= \int_0^t h_{ij}(t|\beta, b_j)dt \\ &= \int_0^t h_0(t) exp(X_{ij}^T\beta + Z_{ij}^T b_j)dt \\ &= H_0(t) exp(X_{ij}^T\beta + Z_{ij}^T b_j). \end{aligned} \tag{1.2}$$

where $H_0(t) = \int_0^t h_0(t)dt$ is unspecified cumulative baseline hazard function. While the survival function is:

$$\begin{aligned} S_{ij}(t|\beta, b_j) &= exp\{-\int_0^t h_{ij}(t|\beta, b_j)dt\} \\ &= exp\{-\int_0^t h_0(t) exp(X_{ij}^T\beta + Z_{ij}^T b_j)dt\} \\ &= exp\{-\int_0^t h_0(t)dt\}^{exp(X_{ij}^T\beta + Z_{ij}^T b_j)} \\ &= [S_0(t)]^{exp(X_{ij}^T\beta + Z_{ij}^T b_j)}. \end{aligned} \tag{1.3}$$

where $S_0(t) = exp\{-\int_0^t h_0(t)dt\}$ is unspecified baseline survival function.

The application of model (1.1) aims to estimate the parameters β and $\mathbf{D}$ from the observed survival times data $(t_{ij}, X_{ij}, Z_{ij}, \delta_{ij})$ (Abrahantes & Burzykowski, 2005; Xu et al., 2009). Let $\mathbf{Z}$ represents $n \times q$ covariates with random effects while $\mathbf{X}$ represents $n \times p$ covariates with fixed effects (Xu, 2004; Xu et al., 2009; M. Crowther, 2017) and often covariates in $\mathbf{Z}$ are included in fixed effects covariates $\mathbf{X}$. If $q = 1, Z_{ij} = 1$ in model (1.1), then the model becomes a usual univariate frailty (or random-intercept) model. Further, the log-hazard scale of model (1.1) is analogous to a linear mixed-effects model, that belongs to the class of models called generalised linear mixed-effects model (GLMM) (Ibrahim et al., 2001; Xiang et al., 2002; Skrondal & Rabe-Hesketh, 2009). For this reason, model (1.1) is also referred to as Cox proportional hazards mixed-effects model (PHMM) (Palmgren & Ripatti, 2002; Abrahantes & Burzykowski, 2005; Xu et al., 2009).

The model (1.1) is multivariate because a cluster of subjects (as opposed to individual subjects) is observed in the random-effects design (Skrondal & Rabe-Hesketh, 2009). The random-effects b_j, which represent various sources of variations for child survival times that are unique to j-th cluster, relate to logarithm of hazards of failure linearly and are additive with fixed-effects terms in the model. In this study, the continuous event-time data will be used to develop group diagnostic methods, as opposed to discrete survival-times data. A discussion of fitting a discrete survival-time model to data is presented in Manda & Meyer (2005). Furthermore, time-independent covariates is assumed. The inference under the random effect Cox model (1.1) considers two sets of data: fixed and random effects.

Suppose t_{ij} is the time subject i in cluster j leaves the study, either by experiencing the event ($\delta_{ij} = 1$) or by surviving to the end of study ($\delta_{ij} = 0$). If the subject experiences the event, then its contribution to the likelihood is $f(t_{ij}|\beta, b_j)$ but if the subject survives, its contribution to the likelihood is $S(t_{ij}|\beta, b_j)$. Thus, assuming independence of subjects within a cluster given random effects, the contribution of subject i in cluster j to the likelihood is given by:

$$\begin{aligned}L_{ij}(t|\beta, b_j) &= [f(t_{ij}|\beta, b_j)]^{\delta_{ij}} \times [S(t_{ij}|\beta, b_j)]^{1-\delta_{ij}} \\ &= [h(t_{ij}|\beta, b_j)S(t_{ij}|\beta, b_j)]^{\delta_{ij}} \times [S(t_{ij}|\beta, b_j)]^{1-\delta_{ij}} \\ &= \left[h_0(t)exp(X_{ij}^T\beta + Z_{ij}^T b_j)\right]^{\delta_{ij}} \times [S_0(t|\beta, b_j)]^{exp(X_{ij}^T\beta + Z_{ij}^T b_j)}.\end{aligned} \tag{1.4}$$

The whole likelihood, conditional on the random effect b_j, is

$$\begin{aligned} L(t|\beta,\mathbf{b}) &= \prod_{j=1}^{M}\prod_{i=1}^{n_j} L_{ij}(t|\beta,b_j) \\ &= \prod_{j=1}^{M}\prod_{i=1}^{n_j} \left[h_0(t)exp(X_{ij}^T\beta + Z_{ij}^T b_j)\right]^{\delta_{ij}} \times [S_0(t)]^{exp(X_{ij}^T\beta+Z_{ij}^T b_j)}. \end{aligned} \tag{1.5}$$

Now, at second stage of the model (1.1) the random effects are considered in the likelihood. Thus, the complete joint likelihood for β and $\mathbf{b}$ is a product of the whole conditional likelihood in (1.5) and the likelihood of the random effects b_j and it is given by:

$$\begin{aligned} L(\beta,\mathbf{b}|\mathbf{t},\mathbf{X},\mathbf{Z}) &= L(t|\beta,b_j) \times \prod_{j=1}^{M} f(b_j) \\ &= \left\{\prod_{j=1}^{M}\prod_{i=1}^{n_j}[h_0(t)exp(X_{ij}^T\beta + Z_{ij}^T b_j)]^{\delta_{ij}} \times [S_0(t)]^{exp(X_{ij}^T\beta+Z_{ij}^T b_j)}\right\} \times \prod_{j=1}^{M} f(b_j). \end{aligned} \tag{1.6}$$

where $\mathbf{t} = \mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_M$ with each component a $n_j \times 1$ vector of survival times, $\mathbf{b} = \mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_M$ with each $\mathbf{b}_j$ a $q_j \times 1$ vector of random variables, $\mathbf{X} = \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M$ where each element is $n_j \times p$ matrix of covariates with fixed effects, and $\mathbf{Z} = \mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_M$ with each element a $n_j \times q_j$ matrix of covariates with random effects.

The maximisation of the likelihood equation (1.6) requires specification of the distributions of the baseline hazard function, $h_0(t)$, baseline survival function $S_0(t)$, and the random effects $f(b_j)$ (Ripatti & Palmgren, 2000; Manda, 2001). In this study, the multivariate normal distribution was assumed for the random effects. One approach that is used to obtain the maximum likelihood estimators for the observed data is through engaging marginal likelihood for β and $\mathbf{D}$ (Manda, 2001). This is done by integrating out the random-effects $\mathbf{b}_j$ from the complete joint likelihood function (1.6) in all clusters, that is,

$$\begin{aligned} L(\beta,\mathbf{D}) &= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} [L(\beta,\mathbf{b}|\mathbf{t},\mathbf{X},\mathbf{Z})] db_1...db_M \\ &= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{M} \prod_{i=1}^{n_j} [h_0(t) exp(X_{ij}^T\beta + Z_{ij}^T b_j)]^{\delta_{ij}} \times [S_0(t)]^{exp(X_{ij}^T\beta + Z_{ij}^T b_j)} \right\} \times \prod_{j=1}^{M} f(b_j) db_1...db_M. \end{aligned} \tag{1.7}$$

The challenge with the marginal likelihood (1.7) is that the integrals are not of closed forms. Iterative algorithms such as EM algorithm (Manda, 2001) are therefore used to get the estimate for β and $\mathbf{D}$. Alternatively, joint likelihood function (1.8) can be used to obtain maximum likelihood estimators for both fixed and random effects simulatenously. This is what is done in the penalized partial likelihood estimation method (Ripatti & Palmgren, 2000; Palmgren & Ripatti, 2002). With this method, the partial joint likelihood function for β and b_j is constructed from a product of conditional density of T given random effect b_j and the density of random-effects $f(b_j)$, which is very similar to methods that treat random effects density as a penalty function (Ripatti & Palmgren, 2000; Palmgren & Ripatti, 2002; Therneau, 2015). The penalised partial joint likelihood function is given by:

$$L_p(\beta,\mathbf{b}|\mathbf{t},\mathbf{X},\mathbf{Z}) = \left\{ \prod_{j=1}^{M} \prod_{i=1}^{n_j} \left[\frac{exp(X_{ij}^T\beta + Z_{ij}^T b_j)}{\sum_{k=1}^{n} R_k(t_{ij}) exp(X_{ijk}^T\beta + Z_{ijk}^T b_j)} \right]^{\delta_{ij}} \right\} \times \prod_{j=1}^{M} f(b_j). \tag{1.8}$$

where $R_k(t_{ij})$ is an indicator showing whether k-th subject is still at risk, that is, not yet experienced the event, at event-time t_{ij}, and δ_{ij} is the censoring indicator.

The estimators $\hat{\beta}$ and $\hat{b}_j$ are obtained by using numerical techniques, such as Newton-Raphson method, because the penalised partial log-likelihood from the likelihood (1.8) is not analytic for one to solve for the parameters. This is done by alternating between iteratively solving the score functions U_β and U_{bj} obtained from equation (1.8) for β and b_j when U_β and U_{bj} are equated to zero. Then, the Laplace approximation is engaged to complete the estimation of covariance parameters θ in $\mathbf{D}$ by using the estimators $\hat{\beta}$ and $\hat{b}_j$ to update covariance elements in $\mathbf{D}$ through maximizing the approximate profile likelihood (Palmgren & Ripatti, 2002; Abrahantes & Burzykowski, 2005) given by:

$$l_{pl}(\hat{\beta}(\theta), \hat{\mathbf{b}}(\theta), \theta) \approx -\frac{n}{2}|logD(\theta)| - \frac{1}{2}log|\frac{\partial^2}{\partial\mathbf{b}\partial\mathbf{b}^T} l_p(\hat{\beta}, \hat{\mathbf{b}})| - \frac{1}{2}\hat{\mathbf{b}}^T D^{-1}(\theta)\hat{\mathbf{b}}. \quad (1.9)$$

where $l_p(\hat{\beta}, \hat{\mathbf{b}})$ is estimated penalised partial log-likelihood.

However, estimators for standard errors for $\hat{\beta}$ obtained from Laplace approximations are said to be slightly biased as they ignore variation that is brought by the estimated covariance $\hat{\mathbf{D}}$ (Palmgren & Ripatti, 2002). Instead, the inverse of observed information matrix of Louis (1982) is used to obtain standard errors for $\hat{\beta}$ and $\hat{\mathbf{D}}$ (Louis, 1982; Vaida & Xu, 2000; Palmgren & Ripatti, 2002; Abrahantes & Burzykowski, 2005), which is given by:

$$I^{-1}(\hat{\beta}, \hat{\theta}) = E\left(\left[-\frac{\partial^2 l_p(t_{ij}, \hat{b}_j|\hat{\beta}, \hat{\theta})}{\partial(\hat{\beta}, \hat{\theta})^2}|t_{ij}, \hat{\beta}, \hat{b}_j\right] - var\left[\frac{\partial l_p(t_{ij}, \hat{b}_j|\hat{\beta}, \hat{\theta})}{\partial(\hat{\beta}, \hat{\theta})}|t_{ij}, \hat{\beta}, \hat{b}_j\right]\right), \quad (1.10)$$

where off diagonal elements are zeroes.

Other numerical estimation techniques that are used for model (1.1) parameter estimation include the EM algorithm (Manda, 2001), the Monte Carlo EM algorithm and the Bayesian MCMC (Ripatti et al., 2002; Hadfield, 2010; Manda, 2011). These are implemented in `R` software through packages like `coxph` (Fox, 2002), `phmm` (Donohue & Xu, 2010), and `lme4` (Bates, 2010).

1.3 Outlier and influence statistics in multivariate survival data

The subject of outlier and influence analysis has received considerable attention in the last four decades. This spans various types of statistical models, such as linear model (D. Cook, 1977; Andrews & Pregibon, 1978; D. Cook, 1979; Belsley et al., 2005; D. Cook & Weisberg, 1982), generalised linear model (Pregibon, 1981; Andersen, 1992; Sarkar et al., 2011), and linear mixed model (Langford & Lewis, 1998; Fung et al., 2002; Z. Pan & Lin, 2005; Zewotir & Galpin, 2005; Cerioli, 2010; Nieuwenhuis et al., 2012; Turkan & Toktamis, 2012). In linear mixed models, group outlier assessment is accomplished through some computation of distance of observations from location measures (Cerioli, 2010). Such techniques have not been studied for clustered survival data. Moreover Langford & Lewis (1998), upon studying outliers in multilevel linear models, proposed further research in non-linear multilevel models.

In linear and linear mixed models, the influence examination involves perturbing some metric such as log-likelihood by allowing different weights to its components. Case deletion is a special example where all cases are given the weight of 1, except the case of interest which is given 0 weight (Zhu et al., 2001). These approaches are directly applicable to other exponential family models in which the observations are independent and identically distributed (Tang et al., 2000; Lee & Xu, 2004). Because of the independence of the components of the metrics, the impact of individual subjects can be precisely quantified by merely removing a term from a metric corresponding to the case(s) of interest (Zewotir & Galpin, 2005; Zewotir, 2008).

As for multinomial models, the terms in the likelihood function corresponding to the cells are not independent. Thus, it does not make sense to merely perturb a term in the likelihood function. Simultaneous perturbations of cell probabilities that take into account dependences have been developed and successfully used to detect influential multinomial observations (Nyangoma et al., 2006). Although Song et al. (2007) demonstrated efficiency of maximization by parts (MBP) algorithm proposed in Song et al. (2005) over expectation-maximisation (EM) algorithm when determining influence of outliers on model fit in linear mixed-effects model using multivariate t distribution, they acknowledge that use of the approach to other model setups such as clustered survival model remains to be investigated.

With the Cox proportional hazard (PH) model (D. R. Cox, 1972), a subject contributes to the partial likelihood that is summed over several risk sets. Thus, dropping one observation affects the likelihood function over many risk sets, making assessment of case influence a bit complex (Cain & Lange, 1984). While the delete-one approximation can be obtained analytically from a one-step Newton-Raphson iteration on the maximum likelihood solution in problems involving likelihood from exponential families (Pregibon, 1981), it is not easily done with the partial likelihood techniques. In the partial likelihood, the one-step approximations are obtained by re-doing a Newton-Raphson step (A. Cook, 2008), thus re-computing and re-inventing the information matrix for each observation, which is computationally expensive (Wei & Su, 1999).

To develop influence examination methods in linear mixed-effects model, the approach by Zewotir & Galpin (2005) is to use basic building blocks of case deletion, through techniques proposed in Christensen et al. (1992), which necessitate re-calculation of updated model parameter estimators resulting from dropping a data record. Then, various residuals for this model are developed by simply substituting the updated estimators in the existing diagnostic methods from linear models, such as Cook's distance (D. Cook, 1977). Zewotir (2008) extended the approach of updating formulae to assessing joint influence of two or more cases to the linear mixed-effects model. Such an approach has not been exploited for influence assessment of the clustered survival model (1.1). There have been however advances in software development for parameter estimation in the clustered survival model, as demonstrated by Fox (2002); Leucuţa & Cadariu (2008); Munda et al. (2012); Loy & Hofmann (2014). But the model still lacks structured diagnostic methods. The various diagnostic measures developed for linear mixed-effects model as discussed in this section may not directly apply to the clustered survival model (1.1), which invites further research for this model.

1.4 Purpose of the study

This study aimed to derive, validate and apply group outlier and influence statistics for the clustered failure-time data analysis. This involved extending similar statistics derived for the linear, liner mixed-effects, and univariate failure-time models to develop appropriate diagnostic statistics for the clustered survival models. In particular, the martingale-based residuals for univariate Cox model (Therneau et al., 1990) and concepts of visual inspection of standardised residuals for group outlier detection in linear mixed model (Langford & Lewis, 1998) were extended to develop group outlier residuals for the clustered survival models. Influence approximations based on one-step Newton-Raphson method for maximum likelihood estimators (Therneau et al., 1990; Cain & Lange, 1984; Storer & Crowley, 1985) were extended to derive the influence statistic for the clustered survival models. The performance of the proposed methods was evaluated using extensive simulation studies and the proposed statistics were implemented through an analysis of mortality of children under the age of five years in Malawi.

The next chapter reviews the residuals and influence measures in various models, with existing application in clustered survival data. Then the derivation of the proposed method for group outlier analysis in multivariate survival models is presented in Chapter 3 together with its numerical examples. This is followed by the derivation of the proposed method for cluster influence analysis in Chapter 4. The simulation study and application for influence method are presented in Chapter 5. Then Chapter 6 is the last one, it discusses the findings and provides the conclusions of this study.

Chapter 2

A Review of Diagnostic Statistics

This chapter discusses various residuals and their use in different statistical models. The current application of some of the diagnostic statistics in clustered survival data is reviewed.

2.1 General assumptions of statistical models

For single-valued response linear models, the structure falls into the form:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{2.1}$$

where $\mathbf{y}$ is $n \times 1$ vector of responses, $\mathbf{X}$ a $n \times p$ design matrix of covariates, β a $p \times 1$ vector of regression parameters, and ϵ is $n \times 1$ vector of unobserved random errors from $N(0, \sigma_\epsilon^2 I)$.

The general assumptions for these linearised models can be summarised into the following: a) the observed covariates X_i on subject i jointly affect the measured response Y_i linearly and additively; b) the errors ϵ for any two subjects are independent; c) the errors have constant variance, and d) the errors have normal distribution with mean zero (Yang, 2012). The linearity and additivity assumptions also apply to the non-linear clustered survival model (1.1), where the covariates are related with the logarithm of the hazard function. The extra assumptions for model (1.1) are that e) the observed covariates X_{ij} for subjects are independent of measured event-times t_{ij} or the hazards of failure for any two subjects are proportional to one another, referred to as proportional hazards (PH) assumption, and that f) the random effect values b_j are *iid* Gaussian with mean zero and some positive covariance.

Similarly, all assumptions of linear model (2.1) and some for clustered survival model (1.1) apply to the linear mixed-effects model (Laird & Ware, 1982; Zewotir & Galpin, 2005; Gharibvand & Liu, 2009; Turkan & Toktamis, 2012; J. Pan et al., 2014; D. Zhang et al., 2016) given by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon, \tag{2.2}$$

where $\{\mathbf{y}, \mathbf{X}, \epsilon\}$ are as defined in model (2.1), only that the vectors and matrices are stacked over time or location (cluster), $\{\mathbf{Z}, \mathbf{b}\}$ are as defined in Section 1.2, $\epsilon \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$, while $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D})$. Then, $var(\mathbf{y}) = var(\mathbf{X}\beta) + var(\mathbf{Z}\mathbf{b}) + var(\epsilon) = \mathbf{0} + \mathbf{Z}var(\mathbf{b})\mathbf{Z}^T + \sigma_e^2\mathbf{I} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma_e^2\mathbf{I} = \mathbf{G}$ is the overall covariance matrix, which is the sum of covariances from individual errors and random effects. Also, the cluster random effects and individual subjects random errors are assumed to be independent, that is $\mathbf{b} \perp \epsilon$. The diagonal elements of $\mathbf{G}$ are referred to as variance components. These constitute the model parameters to be estimated along with the vectors of fixed effects β and random effects $\mathbf{b}$ (D. Zhang et al., 2016). The model assumes that the fixed-effects parameters are not static but vary across clusters and the modelling tries to capture the varying correlation within cluster.

When $\mathbf{Z} = I$, model (2.2) takes a special case called multi-level or random-intercept model (Skrondal & Rabe-Hesketh, 2009) given by:

$$Y_{ij} = X_{ij}^T\beta + b_j + \epsilon_{ij}, \tag{2.3}$$

where Y_{ij} is the observed response value for subject i in cluster j, b_j is the cluster-specific random effect which is the deviation from mean intercept β_0, while the rest of the terms are as defined in models (1.1), (2.1) or (2.2). The model (2.3) has fixed-effects covariates only whose intercept varies across j clusters (Langford & Lewis, 1998; Zewotir & Galpin, 2005; Skrondal & Rabe-Hesketh, 2009). Therefore, the modelling estimates the constant correlation within cluster.

The modelling of data in these various models aims to make statistical estimates and predictions about the response variable Y or time variable T for survival models in the context of the assumptions holding true. Diagnostic statistics therefore, serve to examine fulfillment of the model assumptions so as to generate evidence on accuracy and adequacy of the fitted model. The assessments are done through visual inspection or numerical tests (Aguinis et al., 2013). In the next few sections, the diagnostic statistics for verifying these assumptions in different models are reviewed.

2.2 Diagnostics for linearity and additivity assumptions

For the generalised linear model (2.1), $E(\mathbf{y}) = E(\mathbf{X}\beta) + E(\epsilon) = \mathbf{X}\beta$ and $var(\mathbf{y}) = var(\mathbf{X}\beta) + var(\epsilon) = \sigma_e^2\mathbf{I}$. Because of the normal probability distribution assumption for ϵ in the model (2.1), $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma_e^2\mathbf{I})$. Therefore, the likelihood function for β, denoted $L(\beta|\mathbf{y},\mathbf{X})$ is given by:

$$L(\beta|\mathbf{y},\mathbf{X}) = (2\pi\sigma_e^2)^{-n/2} exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\right), \tag{2.4}$$

hence, the log-likelihood function, denoted $l(\beta|\mathbf{y},\mathbf{X})$ is found by taking the logarithm of the likelihood function (2.4) which gives:

$$l(\beta|\mathbf{y},\mathbf{X}) = -n/2log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta). \tag{2.5}$$

Differentiating the log-likelihood function (2.5) with respect to β gives the expression:

$$\begin{aligned} \frac{d}{d\beta}l(\beta|\mathbf{y},\mathbf{X}) &= \frac{2}{2\sigma_e^2}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\beta) \\ &= \frac{1}{\sigma_e^2}(\mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{X}\beta). \end{aligned} \tag{2.6}$$

The expression (2.6) is called the score vector or score function, which is a function of regression parameters and it shows how the likehood function changes with small changes in each β. Therefore, the Maximum Likelihood (ML) estimator for β is found by solving for β, when the score function (2.6) is equated to zero. Thus the ML estimator for β is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.7}$$

The Maximum Likelihood estimator $\hat{\beta}$ in (2.7) is the same as can be obtained using Least Square estimation (LSE) procedure. The LSE method finds the estimator $\hat{\beta}$ that minimises the sum of squared errors in the model.

One useful diagnostic for model (2.1) is the residual, defined in Chapter 1 as the difference between the observed response vector $\mathbf{y}$ and the estimated response vector $\hat{\mathbf{y}}$, given by

$$\begin{aligned} \hat{\mathbf{e}} &= \mathbf{y}-\hat{\mathbf{y}} \\ &= \mathbf{y}-\mathbf{X}\hat{\beta} \\ &= \mathbf{y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{I}_n-\mathbf{W})\mathbf{y}, \end{aligned} \tag{2.8}$$

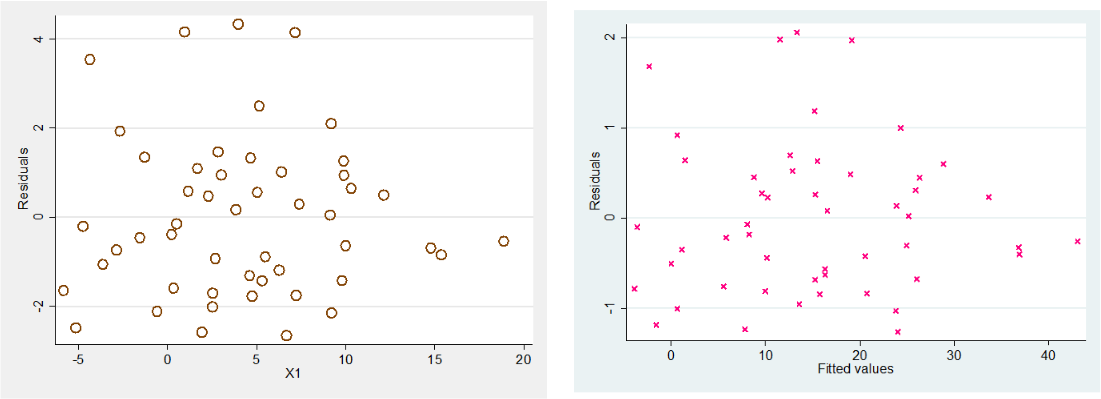
where $W = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the $n \times n$ hat matrix, which is a matrix responsible for changing elements of $\mathbf{y}$ into $\hat{\mathbf{y}}$. When the residual (2.8) is plotted against values of each independent variable X, the graph is used to examine fulfillment of the linearity assumption of model (2.1) for the variable X. Where the plots show no pattern but pure random points, it means the model is linear in the covariate X. If some pattern is noticed, it implies the response variable Y is related with some transformation of the explanatory variable X (Yang, 2012). The additivity assumption is examined using plots of the same residual (2.8) against the fitted values $\hat{Y}_i = X_i\hat{\beta}$. The plots are expected to consistently wag around 0 to show that there is no nonlinear term in the covariates to be added to the model. The same plots can also be used to verify the constant variance or homoscedasticity assumption (Yang, 2012). In this case, the shape of the graph is supposed to be the same along the horizontal axis. If the graph widens up or narrows down, it will imply violation of constant variance assumption of the error term in the model.

As an example, linearity and additivity assumptions are examined on a covariate in a linear model that was fitted to some simulated data. The data had 50 observations and were simulated from a linear model with two covariates using Stata software version 12. The values of the covariates were sampled from normal distributions, i.e. $X_1 \sim N(3.2,6)$ and $X_2 \sim N(10,3.5)$ and the model's error term was generated from $N(0,2.1)$. The model used is given by:

$$Y_i = \beta_0 + X_{i1}^T\beta_1 + X_{i2}^T\beta_2 + \epsilon_i \tag{2.9}$$

where Y_i is the outcome value for subject i, ϵ_i the random error for observation i, the parameters $\beta_0 = 1.5$, $\beta_1 = 2$, and $\beta_2 = 0.5$.

Upon fitting the model to the data simulated by model (2.9), the residual and fitted values were computed and the results in Figures 2.1 (a) and (b) are typical examples of a model that fulfills both linearity and additivity, plus constant variance assumptions, respectively.



(a) Scatter plots of residual versus X_1 for data generated from model (2.9), showing fulfillment of linearity assumption.

(b) Scatter plots of residual versus $\hat{Y}_i$ from model (2.9), showing fulfillment of additivity and constant variance assumptions.

Figure 2.1: Examples of graphs for testing linearity, additivity and constant variance assumptions in linear regression models. Source: Researcher.

To balance off the differences in leverages of different subjects, a scaled residual, which is also referred to as studentised or standardized residual, is used to serve the same purposes of a residual highlighted above (Sarkar et al., 2011). The studentised residual is given by:

$$\begin{aligned}\lambda &= [var(\hat{\mathbf{e}})]^{-1/2}\hat{\mathbf{e}} \\ &= [var[(\mathbf{1}-\mathbf{W})^T\mathbf{y}]]^{-1/2}\hat{\mathbf{e}} \\ &= \left[(\mathbf{1}-\mathbf{W})^T var(\mathbf{y})(\mathbf{1}-\mathbf{W})\right]^{-1/2}\hat{\mathbf{e}} \\ &= \hat{\sigma}^{-1}(\mathbf{1}-\mathbf{W})^{-1/2}\hat{\mathbf{e}},\end{aligned} \tag{2.10}$$

where $\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a leverage matrix as defined before and computation of variance involves its diagonal elements $\mathbf{w}_{ii}$ (Loy & Hofmann, 2014).

With the Cox univariate model (D. R. Cox, 1972), the linearity and additivity relationships of covariates are with the logarithm of the hazard function $h_i(t|\beta)$, given by:

$$h_i(t|\beta) = h_0(t)exp(X_i^T\beta). \tag{2.11}$$

A counterpart residual in survival analysis analogous to residual (2.8) for linear models is the martingale residual given by:

$$m(t_i) = \delta_i - \hat{H}(t_i) = \delta_i - \hat{H}_0(t) exp(X_i^T \hat{\beta}), \tag{2.12}$$

where δ_i is censoring status of i-th subject and $\hat{H}_0(t)$ the fitted cumulative baseline hazard function. This measures excess events at each observation time t_i by computing the difference between the observed and expected number of events over the interval $[0, t_i]$ given the model (Therneau et al., 1990; Fitrianto & Jiin, 2013). The values of the martingale residual (2.12) are expected to be uncorrelated with mean zero when the model is correct (Therneau et al., 1990). Individuals who fail earlier than expected have positive martingale residuals and those who survive longer have negative martingale residuals.

The assessment of linearity assumption for Cox model (2.11) is also done through graphical inspection of the values of martingale residual (2.12) plotted against each covariate X. The plots are expected to consistently average around zero when the variable X has correct linear form with the logarithm of hazard function $logh(t_i|\beta)$ (Therneau et al., 1990; Fox, 2002; Nguyen & Rocke, 2002; Wilson, 2013). This is assessed with the help of smoothing functions such as 'Lowess' (Fox, 2002). The 'lowess' smoother is supposed to be a horizontal straight line passing through zero (Fox, 2002; Wilson, 2013). An example is given in Figure 2.2, where the martingale residual has been computed from the Cox PH model that was fitted on recidivism data discussed in Fox (2002). The data are from an experimental study of 432 male prisoners, who were observed for re-arrest during the first year after release from jail, the data are available on ulr, http://socserv.mcmaster.ca/jfox/Books/Companion/data/Rossi.txt (Fox, 2002). The fitted model is:

$$\hat{h}_i(arrest) = \hat{h}_0(arrest)exp(0.698 \times fin + 0.944 \times age + 0.71 \times race$$
$$+ 0.89 \times wexp + 1.53 \times mar + 0.91 \times paro + 1.09 \times prio + 0.83 \times edu) \quad (2.13)$$

where arrest' = duration of time of release from jail to re-arrest,

$\hat{h}_0(.)$ = baseline hazard,

'age' = age at time of release,

'fin' = whether a person received financial aid or not after release,

'race' = race of a person,

'wexp' = whether a person had full-time job or not prior to arrest,

'mar' = marital status at time of release,

'paro' = whether a person was released on parole or not,

'prior' = number of prior convictions,

'edu' = highest education level (Fox, 2002).

Two of the covariates in the fitted model (2.13), that is, *age* and *prio* were significant, and hence a reduced model with these two covariates was used for post-estimation analysis examples. Figure 2.2 is an example of martingale residual for *age*, which shows that the linearity assumption was slightly violated by the model.

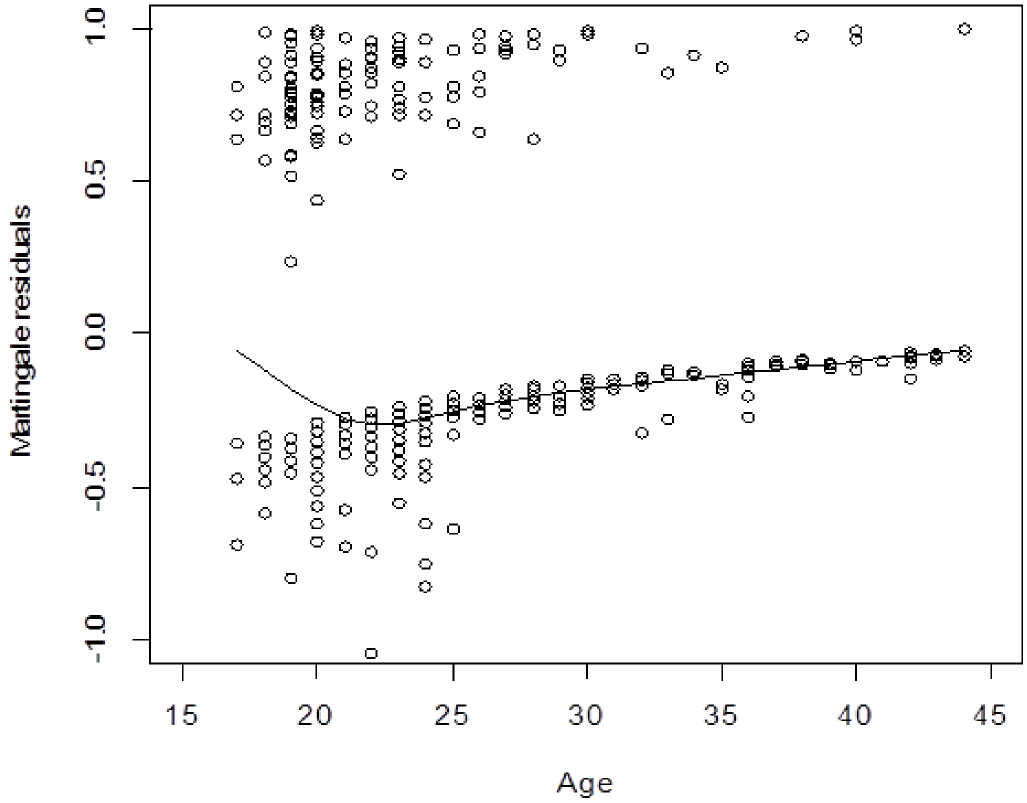


Figure 2.2: Martingale residual with Lowess smoother for recidivism Cox model (2.13). Source: (Fox, 2002)

The current use of martingale residual is limited to individual level and not clustered data. Hence, proper extensions have been defined in this current work for the clustered survival data.

2.2.1 Linearity and additivity assessments in mixed-effects models

From model (2.2), $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{ZDZ}^T + \sigma_\epsilon^2\mathbf{I})$ and in this work, $\mathbf{G} = \mathbf{ZDZ}^T + \sigma_\epsilon^2\mathbf{I}$. Based on the normality assumption for $\mathbf{y}$, the conditional likelihood function for β is given by:

$$L(\beta|\mathbf{y}, \mathbf{X}, \mathbf{b}) = (2\pi)^{-n/2}|\mathbf{G}|^{-1/2}exp\Big\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{G}^{-1}(\mathbf{y}-\mathbf{X}\beta)\Big\}. \qquad (2.14)$$

This gives the conditional log-likelihood function as:

$$l(\beta|\mathbf{y},\mathbf{X},\mathbf{b}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log|\mathbf{G}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{G}^{-1}(\mathbf{y}-\mathbf{X}\beta). \tag{2.15}$$

The maximum likelihood estimator $\hat{\beta}$ of the parameter β is found by taking the partial first derivative of the conditional log-likelihood (2.15) with respect to β and solve for β when the result is equated to zero. The partial first derivative of the conditional log-likelihood function (2.15) is:

$$\begin{aligned}\frac{\partial l(\beta|\mathbf{y},\mathbf{X},\mathbf{b})}{\partial\beta} &= \mathbf{X}^T\mathbf{G}^{-1}(\mathbf{y}-\mathbf{X}\beta)\\ &= \mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}-\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}\beta.\end{aligned} \tag{2.16}$$

Then, equating the equation (2.16) to zero and solving for β gives the ML estimator $\hat{\beta}$ as:

$$\begin{aligned}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}-\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}\beta &= 0\\ \mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}\beta &= \mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}\\ (\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})\beta &= (\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}\\ \therefore\hat{\beta} &= (\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}.\end{aligned} \tag{2.17}$$

The estimator for **b** is found by maximising the complete joint likelihood function for **b** and **y** (Xiang et al., 2002; Turkan & Toktamis, 2012; D. Zhang et al., 2016). Both variables have the normal distribution. Hence, the complete joint likelihood function for **b** and **y** is just the product of the likelihood functions for **b** and conditional likelihood for **y** in (2.14). Once again, from model (2.2) $\epsilon \sim N(\mathbf{0},\sigma_\epsilon^2\mathbf{I})$, $\mathbf{y} \sim N(\mathbf{X}\beta,\mathbf{ZDZ}^T+\sigma_\epsilon^2\mathbf{I})$, and $\mathbf{b} \sim N(\mathbf{0},\mathbf{D})$. Therefore, the complete joint likelihood function will be:

$$L(\mathbf{b},\beta|\mathbf{y},\mathbf{X},\mathbf{Z}) = \frac{exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zb})^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zb})-\frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}\right\}}{(2\pi)^{(n)/2}|\mathbf{D}|^{n/2}|\sigma_\epsilon^2\mathbf{I}|^{n/2}} \tag{2.18}$$

Therefore the complete joint log-likelihood function is:

$$l(\mathbf{b},\beta|\mathbf{y},\mathbf{X},\mathbf{Z}) = -\frac{1}{2}[(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{b})^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{b})+\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}] - log\{(2\pi)^{n/2}|\mathbf{D}|^{n/2}|\sigma_\epsilon^2\mathbf{I}|^{n/2}\}. \tag{2.19}$$

Similar to fixed effects, the estimator of random effects **b** is found by taking first partial derivative of the joint log-likelihood (2.19) with respect to **b** and solve for **b** when the result is equated to zero (Xiang et al., 2002; Turkan & Toktamis, 2012; D. Zhang et al., 2016). The partial first derivative of (2.19) with respect to **b** is given by:

$$\begin{aligned}\frac{\partial l(\mathbf{b},\beta|\mathbf{y},\mathbf{X},\mathbf{Z})}{\partial\mathbf{b}} &= \frac{2}{2}\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{b}) - \frac{2}{2}\mathbf{D}^{-1}\mathbf{b}\\ &= \mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})-1\mathbf{y} - \mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{X}\beta - \mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}\mathbf{b} - \mathbf{D}^{-1}\mathbf{b}\\ &= \mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta) - (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})\mathbf{b}.\end{aligned} \tag{2.20}$$

It follows that equating the result (2.20) to zero and solve for the random effects **b** yields the ML estimator or predictor for **b** given by:

$$\begin{aligned}\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta) - (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})\mathbf{b} &= 0\\ (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})\mathbf{b} &= \mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta)\\ (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}(\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})\mathbf{b} &= (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\beta)\\ \therefore \hat{\mathbf{b}} &= (\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}^T(\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})\\ &= (\mathbf{Z}^T\mathbf{I}\mathbf{Z}+\mathbf{D}^{-1}\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{I}(\mathbf{y}-\mathbf{X}\hat{\beta})\\ &= \mathbf{D}\mathbf{Z}^T(\mathbf{Z}\mathbf{D}\mathbf{Z}^T+\sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})\\ &= \mathbf{D}\mathbf{Z}^T\mathbf{G}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}).\end{aligned} \tag{2.21}$$

The result in equation (2.21) can also be obtained by applying the formula for computing conditional mean of a joint multivariate normal distribution of **y** and **b**, given by: $\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\end{bmatrix} \sim MVN\{\begin{bmatrix}\mathbf{X}\beta\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G} & \mathbf{Z}\mathbf{D}\\ \mathbf{D}\mathbf{Z}^T & \mathbf{D}\end{bmatrix}\}$. Thus, $\hat{\mathbf{b}}$ is found by calculating conditional expectation of $\mathbf{b}$ given $\mathbf{y}$, that is: $E(\hat{\mathbf{b}}|\mathbf{y}) = E(\hat{\mathbf{b}}) + cov(\hat{\mathbf{b}}, \mathbf{y}^T)var^{-1}(\mathbf{y})(\mathbf{y} - E(\mathbf{y}))$.

Then, the fitted value $\hat{\mathbf{y}}$, residual $\hat{\mathbf{e}}$, and studentized residual λ for linear mixed-effects model (2.2) must be a linear combination of estimated fixed- and random-effects (Zewotir & Galpin, 2005; Nobre & Singer, 2011; Zare & Rasekh, 2011; Turkan & Toktamis, 2012). The fitted value is thus given by:

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{b} \\
&= \mathbf{X}\hat{\beta} + \mathbf{Z}D\mathbf{Z}^T G^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \\
&= \mathbf{X}\hat{\beta} + (\mathbf{G} - \mathbf{I}_n)\mathbf{G}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \\
&= \mathbf{X}\hat{\beta} + (\mathbf{I}_n - \mathbf{G}^{-1})(\mathbf{y} - \mathbf{X}\hat{\beta}) \\
&= (\mathbf{I}_n - \mathbf{G}^{-1})\mathbf{y} + (\mathbf{I}_n - (\mathbf{I}_n - \mathbf{G}^{-1}))\mathbf{X}\hat{\beta} \qquad (2.22) \\
&= (\mathbf{I}_n - \mathbf{G}^{-1})\mathbf{y} + \mathbf{G}^{-1}\mathbf{X}\hat{\beta} \\
&= \mathbf{y} - \mathbf{G}^{-1}\mathbf{y} + \mathbf{G}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y} \\
&= \left[\mathbf{I}_n - (\mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1})\right]\mathbf{y} \\
&= (\mathbf{I}_n - \mathbf{R})\mathbf{y},
\end{aligned}$$

where $\mathbf{R} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}$ is shorthand for the symmetric matrix in equation (2.22) that transforms observations into residual (Zewotir & Galpin, 2005; Turkan & Toktamis, 2012). As in univariate linear model, the fitted value is linear in $\mathbf{y}$ for the linear mixed-effects model.

The residual vector follows from the fitted value as:

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{R}\mathbf{y}, \qquad (2.23)$$

and its studentised form is:

$$\lambda = \hat{\sigma}_{\epsilon}^{-1}\mathbf{R}^{-1/2}\hat{\mathbf{e}}, \tag{2.24}$$

Once the residual statistic (2.23) or (2.24) is obtained, the linearity assumption is checked in a similar manner as in univariate linear model, that is, through plots of the residual against each covariate. To discretely study the fit of linear mixed model (2.2), Schabenberger (2005) and Loy & Hofmann (2014) use a segmented residual with three levels. Level-1, also called conditional residual is just the residual form defined in equation (2.23). Then, plotting the residual (2.23) against the fitted values (2.22) will assess the model misspecification, that is, linearity and additivity (Schabenberger, 2005; Loy & Hofmann, 2014).

The level-2 residual, called random-effects residual is just the estimates of random-effects in equation (2.18) obtained through restricted maximum likelihood (Zewotir & Galpin, 2005; Schabenberger, 2005; Turkan & Toktamis, 2012) or through empirical Bayes prediction or other similar methods (Skrondal & Rabe-Hesketh, 2009; Loy & Hofmann, 2014). The estimators (2.21) of random effects are said to be the best linear unbiased predictors (BLUPs) of random effects **b** in model (2.2), due to the fact that they are linear in y, unconditionally unbiased, and 'best' because they minimise marginal sampling variance of prediction error (Zewotir & Galpin, 2005; Skrondal & Rabe-Hesketh, 2009). Hence, they are used to assess the normality assumption of the random-effects through normal quantile-quantile (Q-Q) plots (Loy & Hofmann, 2014). They also serve to examine linearity assumption of the random effects in the model.

In addition, the random effects predictors help in investigating additional explanatory variables in data that contribute significantly to the model. This is done using scatter diagrams, for continuous covariate, and box plots, for categorical covariate, plotted against averages of these potential covariates (Loy & Hofmann, 2014). The level-3, marginal or composite residual consists of the fixed-effects residual only, as in equation (2.8). This diagnostic is used to analyse the marginal covariance structure of the model, among others (Schabenberger, 2005; Loy & Hofmann, 2014).

The definition of a residual in linear mixed-effects model reviewed in this section, i.e., with both fixed- and random-effects parts of the model, has been adapted to develop diagnostics for the clustered survival model in the present study.

2.3 Assessing distributional assumptions of the fitted model

With the generalised linear model (2.1), the normality assumption for the error term is usually checked graphically using, for example, histograms, quantile quantile $Q-Q$ or stem-and-leaf plots of the residual vector $\hat{\mathbf{e}}$ versus subjects' indexes (Yang, 2012). The idea is that if the model is correct, these plots should follow the normal distribution. Alternatively, the residual is plotted against the fitted values $\hat{\mathbf{y}}$ and the graph is inspected if it matches the normal distribution. A substantial criticism for use of graphical approach in assessing the normal distribution assumption is that it pools all covariates together in making conclusions for the model fit, yet the model assumes the linear relationship between the variable X and response Y is conditional on each covariate's values mapping to the mean of the response variable (Yang, 2012).

As for Cox univariate survival model (2.11), the formulation does not provide for the error term, but assumes all sources of individual noise in the event-time variable T are captured by the observable independent variable X. The estimated cumulative hazard function $\hat{H}(t_i)$ also called Cox-Snell or generalised residual (D. Cox & Snell, 1968; Nguyen & Rocke, 2002; Wilson, 2013) is used to assess the general model fit. Theoretically, the cumulative hazard function $H(t_i)$ from equation (1.2) should have a unit exponential distribution (Hosmer Jr et al., 2011). This is because the survivorship function $S(t_i)$ presented in equation (1.3), which is used to compute the cumulative hazard function, has the property $S(k) \leq S(w)$ for $k \geq w$ and hence $S(t_i)$ is a non-increasing function that is bounded below at 0. For this reason, the probability distribution of $S(t_i)$ can be specified as follows:

$$P(S(t_i) \leq x) = P(t_i > S^{-1}(x)) = S(S^{-1}(x)) = x, \tag{2.25}$$

where $S^{-1}(x)$ is the inverse of $S(t_i)$, x the maximum value in the range of $S(t_i)$, and the inequality sign is reversed due to the fact that $S(t_i)$ is a decreasing function.

The above result implies that the density of the variable $S(t_i)$ is $f(S(t_i)) = 1$ since its cdf $P(S(t_i) \leq x) = \int_0^x f(u)du = x$. Therefore, $S(t_i)$ has a Uniform$(0,1)$ distribution, with $T\epsilon[0, \infty)$. Then, through transformation of random variables, it can be shown that the cumulative hazard function $H(t_i)$ will indeed have exponential distribution with parameter 1. Further, $H(t_i)$ is an increasing function with no bound as time t_i gets large, i.e. as $t_i \to \infty$, $H(t_i) \to -log[1 - F(\infty)] = -log[0] = \infty$. The estimated cumulative hazard function or generalised residual from the fitted Cox model is given by:

$$r_{CS,i} = \hat{H}(t_i) = -log[\hat{S}(t_i)] = \hat{H}_0(t_i)exp(X_i^T\hat{\beta}). \tag{2.26}$$

The assessment of model fit is done by plotting the values of the generalised residual (2.26) against its raw values. When the graph is a straight line through the origin with gradient 1, it means estimates of the survivor-times from the model $\hat{S}(t_i)$ match the true survivor-times $S(t_i)$ in the population and hence the Cox model is correctly specified (D. Cox & Snell, 1968). The points above the plotted line imply the model over-predicts failure and those below it suggest under-prediction of failure (D. Cox & Snell, 1968; Nguyen & Rocke, 2002; Wilson, 2013). Upon computing the Cox-Snell residual for the recidivism model (2.13), the results in Figure 2.3 indicate that the model generally fitted the data well, with few cases that were over-predicted.

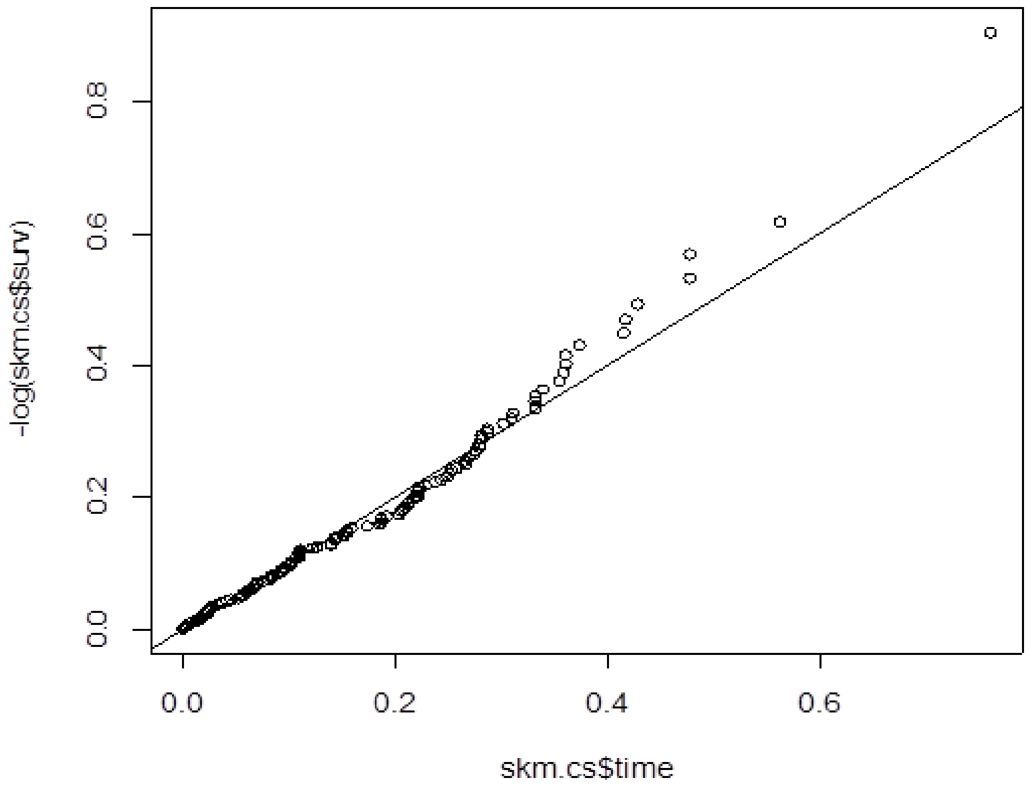


Figure 2.3: Estimated survival curve for recidivism Cox model (2.13). Source: (Fox, 2002)

Few criticisms for the Cox-Snell residual (2.26) relate to difficulties in its interpretation (Zhao et al., 2011; Wilson, 2013) and over-reliance on sample size (Nguyen & Rocke, 2002). Unlike the $Q-Q$ plots and histograms that are used for examination of normal distributional assumptions in linear models, the Cox-Snell residual plots are based on exponential distribution assumption, which is hard to interpret by non-technical audience (Wilson, 2013). Further, closeness of the Cox-Snell residual distribution to unit exponential depends on sample size (Nguyen &

Rocke, 2002). In addition, Zhao et al. (2011) observed that the plots of Cox-Snell residual may not give exact points of departure when the survival model is incorrectly specified.

The main assumption for the Cox PH model (2.11) is the PH assumption stated in Section 2.1. Each covariate is assessed against this assumption using Schoenfeld residual (Schoenfeld, 1982; Fitrianto & Jiin, 2013). This residual determines whether the difference between observed and expected value of each covariate X at each time point t_i is independent of time t_i. The computation makes use of the elements in the score function for β, i.e. U_β (D. R. Cox, 1972; Grambsch & Therneau, 1994). The partial likelihood function for univariate Cox model (2.11), which takes contribution of subjects in the risk sets (D. R. Cox, 1972; Grambsch & Therneau, 1994), is given by:

$$L(\beta|\mathbf{t},\mathbf{X}) = \prod_{i=1}^{n} \left[\frac{exp(X_i^T\beta)}{\sum_{k=1}^{n} R_k(t_i) exp(X_{ik}^T\beta)} \right]^{\delta_i}, \tag{2.27}$$

where $R_k(t_i)$ is an indicator variable showing whether k-th subject is still at risk, that is, not yet experienced the event, at time t_i, and δ_i is the censoring indicator. The log-likelihood function is:

$$l(\beta|\mathbf{t},\mathbf{X}) = \sum_{i=1}^{n} \delta_i \left[X_i^T\beta - log \sum_{k=1}^{n} R_k(t_i) exp(X_{ik}^T\beta) \right]. \tag{2.28}$$

From the log-likelihood function (2.28), the score function for β is found by differentiating the quantity (2.28) with respect to β as:

$$\begin{aligned} U_\beta &= \frac{dl(\beta|\mathbf{t},\mathbf{X})}{d\beta} \\ &= \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{k=1}^{n} R_k(t_i) X_{ik} exp(X_{ik}^T\beta)}{\sum_{k=1}^{n} R_k(t_i) exp(X_{ik}^T\beta)} \right] \\ &= \sum_{i=1}^{n} \delta_i (X_i - \bar{X}(\beta)), \end{aligned} \tag{2.29}$$

where $\bar{X}(\beta) = \frac{\sum_{k=1}^{n} R_k(t_i) X_{ik} exp(X_{ik}^T \beta)}{\sum_{k=1}^{n} R_k(t_i) exp(X_{ik}^T \beta)}$ and it is a weighted average of each covariate's values for members of the risk set. Therefore, the Schoenfeld residual is the component $r_{S,i}$ in the score vector (2.29) given by:

$$r_{S,i} = \delta_i [X_i - \bar{X}(\hat{\beta})]. \tag{2.30}$$

The values of the residual (2.30) are uncorrelated with mean 0 when the model is correct (Schoenfeld, 1982; Nguyen & Rocke, 2002). To improve the diagnostic power of this residual, scaled Schoenfeld residuals are used (Grambsch & Therneau, 1994). The plots of $r_{S,i}$ or its scaled version against the observed survival times t_i show a random pattern around zero if the PH assumption holds true. If there is any systematic pattern, it suggests that there is evidence of dependence of the covariate on time t_i (Fox, 2002; Nguyen & Rocke, 2002). For example, upon computing the Schoenfeld residual for the recidivism model (2.13) for the variables *prio* and *age*, the results in Figure 2.4 show that the *prio* variable fulfilled the PH assumption as many points are departing the smoothing spline band. While, the variable *age* violated the PH assumption as the smoothing spline band appears to gain more points as time increases.

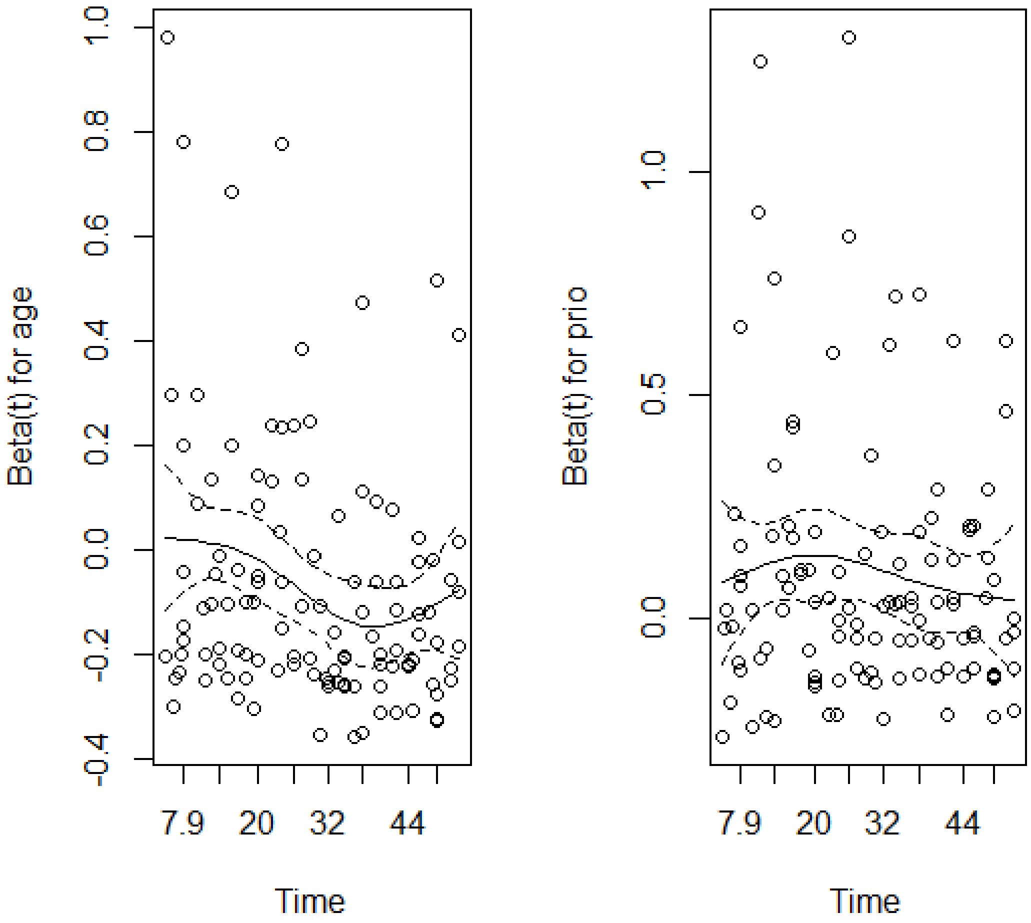


Figure 2.4: Plots of scaled Schoenfeld residual for age and prio against time in the Cox recidivism model, with smoothing lines and confidence bands. Source: (Fox, 2002)

.

The scaled Schoenfeld residual also gives a structure for formal test of the PH assumption (Grambsch & Therneau, 1994; Nguyen & Rocke, 2002). This is accomplished by the function:

$$\beta(t_i) = \beta + \rho g(t_i), \tag{2.31}$$

where $g(t_i)$ is a time-function, β the coefficient of a variable X being investigated, and ρ the slope of the relationship between β and time t_i. The assessment tests the hypothesis $H_0 : \rho = 0$. The result of the test provides a complementary decision to that of graphical examinations (Grambsch & Therneau, 1994; Fox, 2002). When H_0 is rejected, it means the data provide evidence that the variable X is independent of time. Otherwise, failing to reject $H_0 : \rho = 0$ implies the covariate in question is related to time. In such cases, proper transformation of the log cumulative hazard function, such as stratified regression, is recommended (Mehrotra et al., 2012). For the stratified Cox model, a covariate enters the model in strata forms of some specified intervals. In addition, time-dependent Cox regression may be opted for, if stratification approach is not the best solution (Thomas & Reyes, 2014).

The PH assumption extends to the clustered Cox survival model (1.1), although the current form of the Schoenfeld residual (2.30) does not.

2.3.1 Assessment of assumption for the random effect distribution in mixed-effects models

With linear mixed-effects model (2.2), a common method for assessing the normality assumption of random-effects is through using random-effect or level-2 residual (2.18) described by Claeskens & Hart (2009); Loy & Hofmann (2014). As already stated in Section 2.2.1, this is done using normal $Q-Q$ plots, histograms, or stem-and-leaf plots against subjects' indexes. The use of random-effect residual for this purpose is based on the fact that they are the Best Linear Unbiased Predictors (BLUPs) of the model's random effects (Zewotir & Galpin, 2005).

2.4 Statistics for identifying outlying data points

Reportedly, the probability distribution of the residual (2.8) for generalised linear model (2.1) is slightly skewed (Sarkar et al., 2011). Through some transformation of the residual (2.8), a normally distributed quantity is obtained. One of the common transformations is a 'deviance' residual and it serves to examine availability of potential outliers to the model (2.1). Taking a case of the logistic regression model, which is a member of the model (2.1) that analyses binary response data, expressed as:

$$Y_i = \varphi(X_i) + \epsilon_i, \tag{2.32}$$

where:

$$P(Y_i = 1|X) = \varphi(X_i) = \frac{exp(X_i^T\beta)}{1 + exp(X_i^T\beta)}, \tag{2.33}$$

with Y_i is the measured response for subject i with $Y_i = 1$ for a subject that possesses the feature of interest and $Y_i = 0$ otherwise; X_i is a covariate value observed for subject i; β the regression coefficient; ϵ_i the random error for subject i with unknown probability distribution; and $\varphi(X_i)$ is the conditional probability of achieving the feature of interest for subject i whose observed covariate is X_i.

The term $\varphi(X_i)$ in equation (2.33) is also called a logistic function because it resembles the logistic curve. Further, the link function $\eta = E(Y|X)$ for model (2.32) is the logarithm of odds of having $Y_i = 1$ given a covariate value X_i, which is:

$$\eta = E(Y|X) = \hat{\varphi}^* = log\left[\frac{\varphi(X_i}{1 - \varphi(X_i)}\right] = X_i^T\beta. \tag{2.34}$$

The i-th subject residual for the logistic regression model (2.32) is a binary term (Sarkar et al., 2011) given by:

$$\hat{e}_i = Y_i - \hat{\varphi}(X_i) = \begin{cases} 1 - \hat{\varphi}(X_i) & if \quad Y_i = 1 \\ -\hat{\varphi}(X_i) & if \quad Y_i = 0, \end{cases} \tag{2.35}$$

where $\hat{\varphi}(X_i) = \frac{exp(X_i^T\hat{\beta})}{1+exp(X_i^T\hat{\beta})}$ is the fitted conditional probability of success given covariate values X_i.

The residual (2.35) implies that the variance of the error term as well as the response variable in logistic regression is a function of the covariates, as $var(\hat{e}_i) = var(Y|X) = \varphi(X_i)(1-\varphi(X_i))$. This is a departure from the convention set in the general linear model (2.1), where covariates contribute zero variance to the response. Moreover, plotting the residual (2.35) against the fitted values Y_i will provide some hard-to-interpret information about the model due to the unknown distribution of the error term in the logistic regression model. Hence, a transformed measure called Pearson residual is used instead (Sarkar et al., 2011), and it is given by:

$$\psi_i = \frac{\hat{e}_i}{\sqrt{\hat{\varphi}(X_i)(1-\hat{\varphi}(X_i))}} = \frac{\hat{Y}_i - \hat{\varphi}(X_i)}{\sqrt{\hat{\varphi}(X_i)(1-\hat{\varphi}(X_i))}}. \tag{2.36}$$

The square of Pearson residual (2.36) measures contribution of each response Y_i to the Pearson chi-square test statistic. But the measure does not follow approximate chi-square distribution (Sarkar et al., 2011). To utilise this residual in assessing problematic observations, it is standardised so as to have an approximate normal distribution (Sarkar et al., 2011), given by:

$$\lambda_i = \frac{\hat{e}_i}{\sqrt{\hat{\varphi}(X_i)(1-\hat{\varphi}(X_i))(1-w_{ii})}} = \frac{\psi_i}{\sqrt{(1-w_{ii})}}, \tag{2.37}$$

where w_{ii} is the i-th diagonal element of the estimated hat matrix (or Pregibon leverage) $\mathbf{W} = \hat{\mathbf{G}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{G}}^{1/2}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{G}}^{1/2}$ and $\hat{\varphi}(X_i)(1-\hat{\varphi}(X_i))$ is conditional variance of the response variable given X (Sarkar et al., 2011). The residual (2.37) helps in identifying influential subjects to model (2.32), when the measure (2.37) is plotted against fitted values or subjects indexes (Sarkar et al., 2011).

The rest residuals for the logistic regression model (2.32) build on the Pearson statistic (2.36) or its studentised form (2.37). One such residual is the deviance residual. A model's deviance $Dv = 2(l(MS) - l(MR))$, where MS and MR stand for saturated and reduced model, respectively, measures the distance between a component of the log-likelihood of the fitted model and the corresponding component that would result if each point was fitted exactly. The models that use maximum likelihood estimation aim at minimising the sum of deviance residuals. So, these statistics are computed from the fitted model deviance in order to track potential outliers and mis-specified subjects. The deviance residual for the i-th subject is the signed square root of the contribution of that subject to the sum for the fitted model deviance (Sarkar et al., 2011) given by:

$$d_i = sgn(Y_i - \hat{\varphi}(X_i))\left[-2(Y_i log \hat{\varphi}(X_i)) + (1 - Y_i) log(1 - \hat{\varphi}(X_i))\right]^{\frac{1}{2}}, \tag{2.38}$$

where ‘sgn’ is the sign of the i-th subject residual $\hat{e}_i$, that is, plus or minus. The plots of the deviance residual (2.38) against the estimated probabilities $\hat{\varphi}(X_i)$ will show the model’s outliers at cutoff ±2 (Sarkar et al., 2011).

Schall & Dunne (1988) use a different approach to study outliers. They specify a separate model with raw residual $\hat{\mathbf{e}}$ as the response vector and engage some tests to identify outliers to the this model. The outliers to the residual model are also deemed unusual subjects to the main model with response variable Y. Others use the scaled residual (2.10) directly to analyse outliers to a model, as by Chebyshev theorem not more than 5% of values of the studentized measure (2.10) should be outside the bounds ±1.96, while not more than 1% will be beyond ±2.58 (Dobson & Barnett, 2008; Sarkar et al., 2011). The observations outside these limits are considered outliers.

With Cox survival model (2.11), the martingale residual (2.12) has the same weakness of skewed distribution as the raw residual (2.8) in generalised linear models. This is because the subject’s censoring condition δ_i in the measure (2.12) can only take values of 1 or 0, while the cumulative hazard $\hat{H}(t_i)$ has strictly positive values in the interval $[0,\infty)$. This makes the statistic (2.12) to have highly positively-skewed distribution, with values in the range $(-\infty,1]$. So, the quantity (2.12) may not detect outliers using values on both ends of its distribution. A counterpart deviance residual for examining outliers in survival models was studied by Therneau et al. (1990). Taking the baseline hazard of Cox model as nuisance parameter, Therneau et al. (1990) engaged the Lagrange multiplier maximization technique to derive the deviance structure from the fitted model's deviance. In so doing, the residual (2.12) was transformed into a statistic that is symmetrical about zero (Therneau et al., 1990; Fitrianto & Jiin, 2013).

Thus, the deviance residual for univariate Cox model was defined as follows:

$$d_i = sgn(m(t_i))\left[-2(m(t_i)+\delta_i log(\delta_i - m(t_i))\right]^{\frac{1}{2}}, \tag{2.39}$$

where $\hat{m}(t_i)$ is the martingale residual, δ_i censoring status, and 'sgn' is the sign of the measure (2.9), which is plus or minus. As in generalised linear model, plotting values of the deviance residual (2.39) against linear predictors $log\hat{h}_i$ will show potential outliers to the survival model. The values of (2.39) outside the range ±2.5 are usually considered outliers (Nguyen & Rocke, 2002). Upon computing the measure (2.39) for the Cox model (2.13), the results in Figure 2.5 show five people, who were re-arrested earlier than estimated by the model.

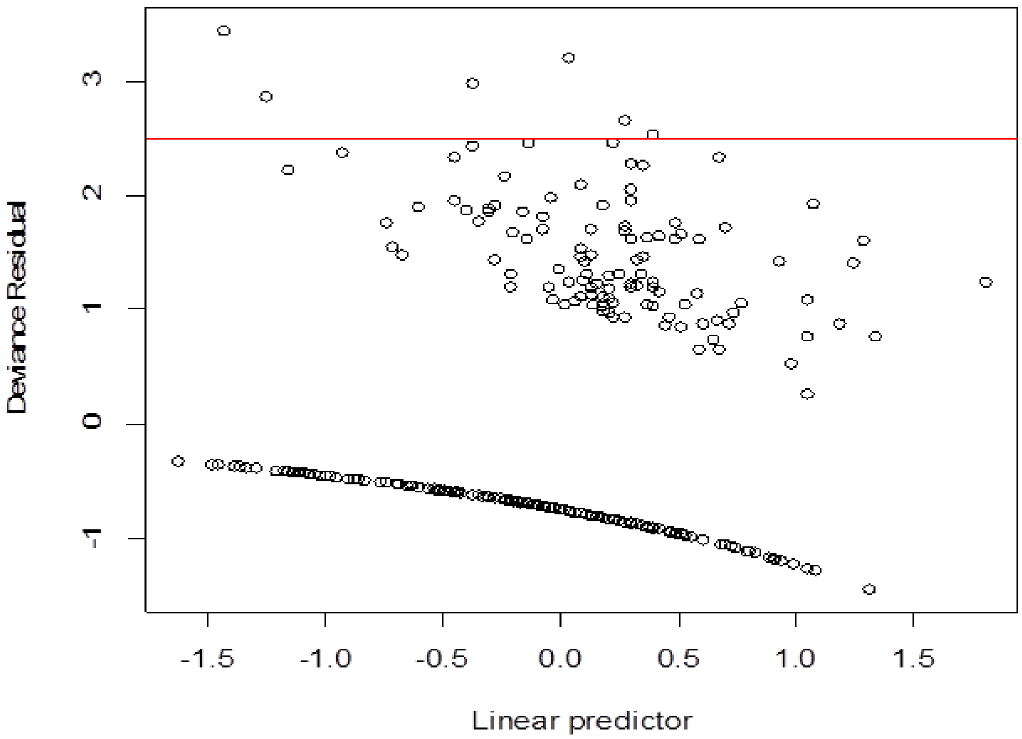


Figure 2.5: Scatter plot of deviance residual versus linear predictor that had covariates age and prior conviction in the recidivism Cox model. Source: (Fox, 2002)

This current work extended the definitions of martingale and deviance residuals reviewed in this section to the clustered survival model (1.1). This was done in order to explore methods for the group outlier examination in the clustered survival data.

2.4.1 Outlier identification in mixed-effects models

With the linear mixed-effects model (2.2), the outlier assessment methods are similar to those of the generalised linear model (2.1). The only difference is that the methods are segregated according to levels of data (Langford & Lewis, 1998; Bell & Malacova, 2004; Loy & Hofmann, 2014). For example, Bell & Malacova (2004) analysed outlying education outcomes to a multilevel logistic regression model applied on university applicants in the UK, with two stages: high school progress and university admissions. While Langford & Lewis (1998) studied outlying schools or pupils to the multilevel model using the UK's local education authority data from 66 schools with 2478 students in 136 school-years. In both cases, the measures that were used for group outlier examination in the multilevel model were extensions of those that are used for individual subjects in linear models. In particular, Langford & Lewis (1998) use the distribution of standardised residuals when plotted against clusters to assess the clusters that deviate from the rest in the model.

As highlighted already in Section 2.2.1, another approach for group outlier analysis in linear mixed models is to use the random-effects or level-2 residual (2.21) plotted against group identities (Loy & Hofmann, 2014). The shortfall of this approach is that it does not fully utilise the fixed-effects component of the model in computing the residual, which may lead to unrealistic estimates of true group outliers. This current work has explored group outlier methods for the survival mixed model that exhaust all the data structures in the mixed survival model.

A general method for trapping multivariate outliers from all levels of data in linear mixed-effects model was suggested by Cerioli (2010). The approach uses the re-weighted minimum covariance determinant (RMCD), similar to Mahalanobis distance (Cerioli, 2010). The MCD component in RMCD is part of the sample of h data points, $n/2 \leq h < n$, whose covariance has the smallest determinant (Cerioli, 2010). The method is given by:

$$d^2_{i(RMCD)} = (Y_i - \hat{\mu}_{(RMCD)})^T \hat{\Sigma}^{-1}_{(RMCD)} (Y_i - \hat{\mu}_{(RMCD)}), \tag{2.40}$$

where

$$\hat{\mu}_{(RMCD)} = \frac{1}{m} \sum_{i \in Y_{MCD}} \varpi_i Y_i \tag{2.41}$$

is re-weighted MCD estimate of location and estimate of scatter $\hat{\Sigma}$ proportional to dispersion matrix is:

$$\hat{\Sigma} = \frac{k_{(RMCD)(m,n,v)}}{m-1} \times \sum_{i \in Y_{MCD}} \varpi_i (Y_i - \hat{\mu}_{(RMCD)})^T (Y_i - \hat{\mu}_{(RMCD)}) \tag{2.42}$$

with $m = \sum_{i \in Y_{MCD}} \varpi_i$; v dimension of covariance matrix Σ; and $k_{(RMCD)(m,n,v)}$ proportionality constant to control for bias (Cerioli, 2010).

The potential of using dispersion of residuals to examine group outliers in mixed models as suggested by Cerioli (2010) has been explored in this current work to devise the method for assessing group outliers in clustered survival data.

2.5 Diagnostic statistics for leverage and influence

For the generalised linear model (2.1), a vector of fitted values $\hat{\mathbf{y}}$ is given by:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{W}\mathbf{y}. \end{aligned} \tag{2.43}$$

Therefore, a leverage, also called hat or projection matrix $\mathbf{W}$ for the fitted model is the first derivative of the vector of fitted values (2.43) with respect to $\mathbf{y}$ given by:

$$\begin{aligned} \mathbf{W} &= \frac{d\hat{\mathbf{y}}}{d\mathbf{y}^T} \\ &= \frac{d\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}{d\mathbf{y}^T} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \end{aligned} \tag{2.44}$$

Hence, the leverage of the i-th observation on the i-th fitted value, denoted by w_{ii} is the amount by which the i-th estimate $\hat{Y}_i$ would change with respect to the i-th observed response value (Nobre & Singer, 2011; Sarkar et al., 2011). This is the i-th element of the main diagonal of hat matrix $\mathbf{W}$.

The leverage w_{ii} is always a ratio, whose range of values is $[0,1]$. The value 0 means that subject i has no effect on $\hat{Y}_i$ and 1 implies the i-th subject has remarkable effect on the fitted line or that line $\hat{Y}_i$ passes through the data point (X_i, Y_i) (Sarkar et al., 2011). Thus large leverage subjects have influence on the fitted regression line. The working cutoff from which a leverage is considered large is $2p/n$ or $4/n$, where p is number of parameters in the model and n sample size (Dobson & Barnett, 2008; Nobre & Singer, 2011). The assessment is also done using graphical methods, that is, by plotting w_{ii} against subject indexes.

As for the linear mixed-effects model (2.2), leverage is defined according to the level of analysis of the data. This is based on the marginal fitted values and conditional fitted values for the model. The conditional fitted value for linear mixed-effects model (2.2) can be expanded using the ML estimators for fixed and random effects given in equations (2.17) and (2.21) as:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{b}\\ &= \mathbf{X}\hat{\beta} + \mathbf{Z}D\mathbf{Z}^T G^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y} + \mathbf{Z}D\mathbf{Z}^T G^{-1}(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y})\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y} + \mathbf{Z}D\mathbf{Z}^T\left(G^{-1} - G^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\right)\mathbf{y}\\ &= \hat{\mathbf{y}}^{*1} + \hat{\mathbf{y}}^{*2},\end{aligned} \tag{2.45}$$

where the component $\hat{\mathbf{y}}^{*1} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}$ is the conditional fitted value for fixed effects and $\hat{\mathbf{y}}^{*2} = \mathbf{Z}\hat{\mathbf{b}} = \mathbf{Z}D\mathbf{Z}^T\left(G^{-1} - G^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\right)\mathbf{y}$ the marginal fitted value for random effects.

This leads to the definition of a generalized leverage matrix for marginal fitted values or simply generalised marginal leverage matrix (Nobre & Singer, 2011), given by:

$$\begin{aligned}\mathbf{Q}_1 &= \frac{\partial \hat{\mathbf{y}}^{*1}}{\partial \mathbf{y}^T} \\ &= \frac{\partial \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}}{\partial \mathbf{y}^T} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}.\end{aligned} \tag{2.46}$$

The generalised marginal leverage matrix $\mathbf{Q}_1$ is the overall conditional leverage for the fixed effects (Fung et al., 2002; Nobre & Singer, 2011). This will measure influence of an observation or cluster on the conditional fitted value $\hat{\mathbf{y}}^{*1}$.

With the second level of the data, the generalised leverage matrix for the random effect component is given by:

$$\begin{aligned}\mathbf{Q}_2 &= \frac{\partial \hat{\mathbf{y}}^{*2}}{\partial \mathbf{y}^T} \\ &= \frac{\partial \mathbf{Z}D\mathbf{Z}^T\left(G^{-1} - G^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\right)\mathbf{y}}{\partial \mathbf{y}^T} \\ &= \mathbf{Z}D\mathbf{Z}^T\left(G^{-1} - G^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\right).\end{aligned} \tag{2.47}$$

The leverage matrix $\mathbf{Q}_2$ can estimate influence of a subject or cluster on marginal fitted values of the random effects. The part $\mathbf{Z}D\mathbf{Z}^T$ of the generalised marginal leverage $\mathbf{Q}_2$ represents proportion of within-cluster variability explained by the presence of random effects and it is referred to as generalised random component leverage matrix (Nobre & Singer, 2011). This component depends on random covariates and covariance matrix for random effects unlike the entire $\mathbf{Q}_2$, which depends on both fixed and random effects. Hence, $\mathbf{Z}D\mathbf{Z}^T$ can serve well in examining leverage of observations on fitted random effects of the model Nobre & Singer (2011). The subjects with high leverage in respect of $\mathbf{Z}D\mathbf{Z}^T$ in $\mathbf{Q}_2$ are expected to have disproportionate weight on the estimate of the variance components of the model.

Now, using the conditional fitted value (2.45) for the linear mixed-effects model, a generalized joint leverage matrix for subjects on overall fitted values is given by:

$$\begin{aligned}\mathbf{Q} &= \mathbf{Q}_1 + \mathbf{Q}_2\\ &= \frac{\partial \hat{\mathbf{y}}^{*1}}{\partial \mathbf{y}^T} + \frac{\partial \hat{\mathbf{y}}^{*2}}{\partial \mathbf{y}^T}\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1} + \mathbf{Z}D\mathbf{Z}^T\left(G^{-1} - G^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}^{-1}\right).\end{aligned} \tag{2.48}$$

The diagonal of the quantity $\mathbf{Q}$ will measure overall influence of a subject on the fitted value of the entire model (Nobre & Singer, 2011).

Further transforms of the residual (2.8) are used to examine influence of a subject in the model. Examples include the Difference in Fit Standardised ($DFFITS$) (Belsley et al., 2005) and Cook's distance (CD) (D. Cook, 1977). The $DFFITS_i$ of subject i is a scaled measure that captures the change in the fitted value $\hat{Y}_i$ for the i-th subject computed after removing subject i from the data (Belsley et al., 2005). The i-th subject $DFFITS_i$ for $\hat{Y}_i$ from the linear model (2.1) is given by:

$$\begin{aligned}DFFITS_i &= \frac{(\hat{Y}_i - \hat{Y}_{(i)})}{se(\hat{Y}_{(i)})}\\ &= \lambda_i\sqrt{\left(\frac{w_{ii}}{1-w_{ii}}\right)}.\end{aligned} \tag{2.49}$$

The values of $DFFITS_i$ larger than $2 \times \sqrt{p/n}$, in absolute sense, where p is the number of parameters in the model, are considered influential on the fitted value $\hat{Y}_i$.

While the Cook's distance CD_i for the i-th subject for the model (2.1) is the change in parameter estimates $\hat{\beta}$ following removal of i-th data record (D. Cook, 1977), measured as sum of this change for all parameters in the model, given by:

$$
\begin{aligned}
CD_i &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2_\epsilon} \\
&= \frac{\left((\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1} X_i [\mathbf{y} - X_i^T \hat{\beta}]\right)^T \mathbf{X}^T \mathbf{X} \left((\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1} X_i [\mathbf{y} - X_i^T \hat{\beta}]\right)}{p\hat{\sigma}^2_\epsilon} \\
&= \frac{\left(\frac{(\mathbf{X}^T\mathbf{X})^{-1} X_i [\mathbf{y} - X_i^T \hat{\beta}]}{1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i}\right)^T \mathbf{X}^T \mathbf{X} \left(\frac{(\mathbf{X}^T\mathbf{X})^{-1} X_i [\mathbf{y} - X_i^T \hat{\beta}]}{1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i}\right)}{p\hat{\sigma}^2_\epsilon} \\
&= \left[\frac{\mathbf{y} - X_i^T \hat{\beta}}{\hat{\sigma}_\epsilon \sqrt{\left(1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i\right)}}\right]^2 \frac{X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i}{p\left(1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i\right)} \qquad (2.50) \\
&= \frac{\lambda_i^2}{p} \frac{var(\hat{Y}_i)}{var(\hat{e}_i)} \\
&= \frac{\lambda_i^2}{p} \frac{X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i}{1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i} \\
&= \frac{\lambda_i^2}{p} \frac{w_{ii}}{1 - w_{ii}},
\end{aligned}
$$

where $\hat{\sigma}^2_\epsilon = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-p}$ is estimated variance of random error term, $\mathbf{X}_{(i)}$ is the $n-1 \times p+1$ design matrix without i-th row X_i^T of covariates, $(\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1} X_i X_i^T (\mathbf{X}^T\mathbf{X})^{-1}}{1 - X_i^T (\mathbf{X}^T\mathbf{X})^{-1} X_i}$ (D. Cook, 1977; Pregibon, 1981). The term w_{ii} is i-th diagonal element of the leverage matrix for the linear model (2.1), and p is the number of parameters in the model. The values of CD_i (2.50) that are greater than 1 are usually considered large and their corresponding subjects become targets for influence on the regression parameter estimates (D. Cook, 1977; Sarkar et al., 2011).

A different approach to case-deletion for assessing influence of the data point on the fitted value $\hat{Y}_i$ is the squared norm of a vector of forecast changes, called Pena's Statistic (Peña, 2005; Turkan & Toktamis, 2012). It estimates how deletion of each data record affects the forecast for a specific observation of interest (Türkan & Toktamis, 2013). For the general linear model (2.1), the Pena's statistic, denoted

a_i is given by:

$$\begin{aligned} a_i &= \frac{1}{p\hat{\sigma}_\epsilon^2}||\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}||^2 \\ &= \frac{\lambda_i^{*2} w_{ii}}{p(1-w_{ii})} \\ &= \frac{e_i^{*2} w_{ii}}{p\hat{\sigma}_\epsilon^2(1-w_{ii})^2} \\ &= \frac{w_{ii}^T w_{ii} e_i^{*T} e_i^*}{p\hat{\sigma}_\epsilon^2 w_{ii}(1-w_{ii})^T(1-w_{ii})}, \end{aligned} \tag{2.51}$$

where $e_i^* = \hat{Y}_i - \hat{Y}_{(i)}$ is the displacement in i-th fitted value when the observation i is deleted; w_{ii} is the leverage for subject i from the model with reduced data; $\hat{\sigma}_\epsilon^2 = \hat{\mathbf{e}}^T\hat{\mathbf{e}}/(n-p)$ is unbiased estimate of variance of the error term; and $\lambda_i^{*2} = \hat{\mathbf{e}}[\hat{\sigma}_\epsilon^2(1-w_{ii})]^{-1/2}$ is the studentised displacement e_i^* (Türkan & Toktamis, 2013).

The measure can also track a relative outlying tendency of each subject compared to the rest (Türkan & Toktamis, 2013). Peña (2005) demonstrated that the measure has asymptotic normal distribution, with capability to detect a group of high leverage similar outliers, a feature that Cook's statistic falls short of, also observed by (Türkan & Toktamis, 2013). In addition, the measure was proven to be handy in detecting model heterogeneity in large high-dimensional datasets (Peña, 2005). The cutoff proposed by Peña (2005) for outlying observations is: $|a_i - median(a_i)| \geq 4.5MAD(a_i)$, where $MAD(a_i) = median[|a_i - median(a_i)|]$, i.e. Median of Absolute Deviations from sample median.

For the Cox univariate model (2.11), leverage of i-th subject for the covariate X at time t_i is a distance between the subject's covariate value X_i and its weighted average $\bar{X}$ at time t_i, given as a component in the score function (2.25):

$$w_{ii} = X_i - \bar{X}(\hat{\beta}, t_i). \tag{2.52}$$

Large values of the leverage (2.52) indicate that a subject exerts considerable influence on the fitted hazard $\hat{h}(t_i)$. The general influence assessment for regression coefficients estimators $\hat{\beta}$ in Cox univariate model is done by the "deleted observations" method, a procedure involving computing $\hat{\beta}$ from complete data and then $\hat{\beta}_{(i)}$ from subset of the data following elimination of subject i (Nguyen & Rocke, 2002; Cleves et al., 2010; Wilson, 2013).

Then, the influence of i-th subject on $\hat{\beta}$ is measured by the statistic called Difference in Beta Standardised ($DFBetas$), also referred to as $Delta-beta$, which captures the change in the value of the coefficient $\hat{\beta}$. The $DFBetas$ is given by:

$$DFBetas_i = \frac{\hat{\beta}_i - \hat{\beta}_{(i)}}{se(\hat{\beta}_{(i)})}. \tag{2.53}$$

The large values of $DFBetas_i$ are indicative of influence of the subject i on the estimate $\hat{\beta}$.

The process for computing DFBetas (2.53) is however tedious, as it involves re-fitting the model (2.11) to the data $n+1$ times. This is a major setback of the method (2.53) (Nguyen & Rocke, 2002; Cleves et al., 2010; Wilson, 2013). An alternative and efficient measure, called Score residual is used for the univariate Cox model (2.11) (Therneau et al., 1990; Wei & Su, 1999). This technique is based on the fact that mean $\bar{X}$ of a covariate changes over time for the model (2.11), as individuals leave the risk set. The leverage (2.52) therefore, takes the form that integrates out the time-effect (Therneau et al., 1990; Wilson, 2013), as:

$$\begin{aligned} w_{ii}(\hat{\beta}, t) &= \int_0^\infty [X_i(t) - \bar{X}(\hat{\beta}, t)] d\delta_i(t) \\ &= \int_0^\infty [X_i(t) - \bar{X}_k(\hat{\beta}, t)] dm_i(t), \end{aligned} \tag{2.54}$$

where $m_i(.)$ is a martingale residual for the i-th subject. The transformed leverage (2.54) is the Score residual. It measures the contribution of subject i in the risk set to the score function for the covariate X.

Since the score function estimates one parameter at a time and treats others as constants, there is possibility of setting up a vector of score functions and hence score residuals $w_{ii}(\hat{\beta},t) = (w_{i1}(\hat{\beta},t),...,w_{ip}(\hat{\beta},t))^T$ (Nguyen & Rocke, 2002; Wilson, 2013). The estimation of DFbetas (2.52) using score residual (2.54) is thus done by multiplying the inverse of variance-covariance matrix of the parameter estimates $I(\hat{\beta})^{-1}$ with the vector of score residuals (Therneau et al., 1990) as:

$$\begin{aligned} DFbetas_i &= (\hat{\beta}_i - \hat{\beta}_{(i)})/se(\hat{\beta}_{(i)}) \\ &\approx I(\hat{\beta})^{-1}(w_{i1}(\hat{\beta},t),...,w_{ip}(\hat{\beta},t))^T. \end{aligned} \tag{2.55}$$

When the Cox model is correct, plots of score residual (2.55) against values of the covariate X_i will fluctuate around zero and any systematic deviations will suggest lack-of-fit for the independent variable X. This will at the same time spot influence of subjects on the parameter estimate (Therneau et al., 1990). Figure 2.6 is the DFBetas plots for the variables *age* and *prio* in the Cox recidivism model (2.13). It shows that most subjects did not have influence on both *age* and *prio* variables, since the DFBetas plots concentrated around the zero line. Although few subjects had their points away from the zero line, their values were very small, indicating negligible influence on regression coefficients.

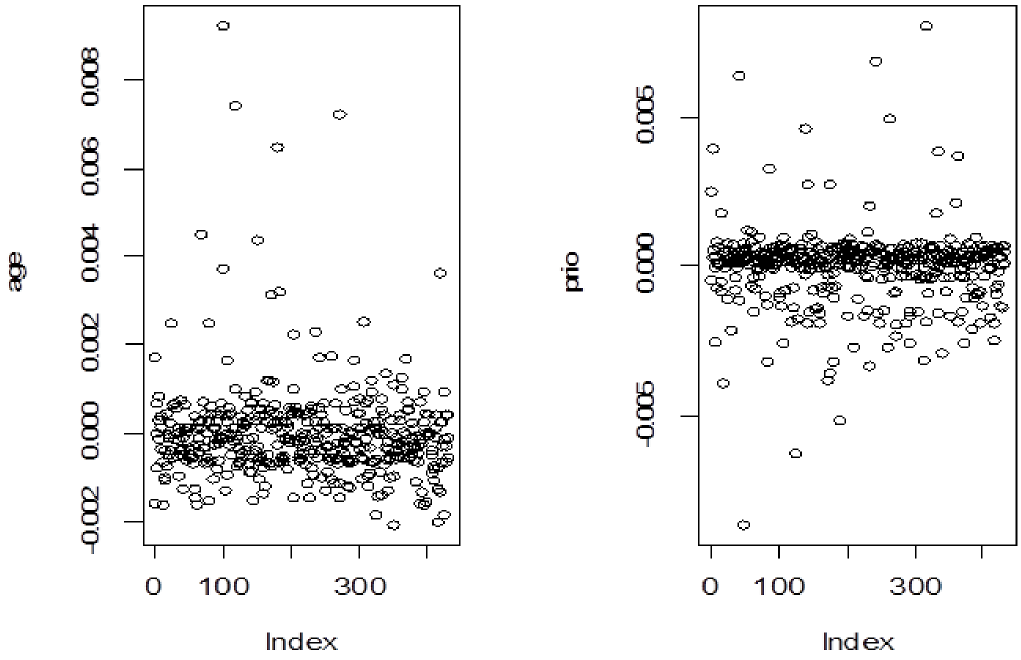


Figure 2.6: Index plots of DFbetas on age and prio for Cox regression for re-arrest data. Source: (Fox, 2002)

Alternative influence technique for the Cox model was studied by Cain & Lange (1984). They engaged concepts of influence curve from Samuels (1978) and Hampel (1974) to develop an approximation to change in parameter estimate $\hat{\beta} - \hat{\beta}_{(i)}$ using first-order Taylor series expansion, by taking the estimator $\hat{\beta}$ as a function of individual weight ϖ_i, i.e. $\hat{\beta}(\varpi_i)$. Hence, an approximation to $\hat{\beta} - \hat{\beta}_{(i)}$ following removal of i-th subject in the model is given by:

$$\hat{\beta} - \hat{\beta}_{(i)} \approx \partial\hat{\beta}/\partial\varpi_i = (-\partial U_\beta/\partial\hat{\beta})^{-1}\partial U_\beta/\partial\varpi_i, \tag{2.56}$$

where U_β is the score vector of the model, and ϖ_i the subject's weight taking values of 1 for all subjects in the model and 0 for the removed subject. The influence of a subject is assessed graphically, by plotting the measure against ranked survival time (Cain & Lange, 1984).

The work reviewed in this section has shown that methods on subjects' influence for linear, linear mixed, and univariate Cox models are based on studying leverage and outlier statistics of the subjects. Then, influence statistics are constructed as a product of the two quantities. For example, the DFFITS and Cook's distance for linear models are products of leverage statistics and studentized residuals, while DFBetas for univariate Cox survival model is a product of leverage and inverse covariance matrix of parameter estimates. The current work has exploited such approaches in the clustered survival model to develop group influence measure for this model.

2.5.1 Influence diagnostics in mixed-effects models

For the linear mixed-effects model (2.2), cluster-deletion diagnostics are derived from partitioned regression matrices and vectors by clusters. Once a cluster is removed from the data, the updated parameter estimators are solved from the data that remain (Xiang et al., 2002; Zewotir, 2008). These help in computing measures for estimating contribution of the dropped cluster to the model. To illustrate these partitions, the linear mixed-effects model (2.2) is re-specified in stacked form as below, where the horizontal dotted lines indicate the demarcations between clusters:

$$
\begin{bmatrix}
Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_1 1} \\ \cdots\cdots \\ Y_{12} \\ Y_{22} \\ \vdots \\ Y_{n_2 2} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Y_{1M} \\ Y_{2M} \\ \vdots \\ Y_{n_M M}
\end{bmatrix}
=
\begin{bmatrix}
1 & X_{111} & X_{112} & \cdots & X_{11p} \\
1 & X_{211} & X_{212} & \cdots & X_{21p} \\
\vdots & & & & \\
1 & X_{n_1 11} & X_{n_1 12} & \cdots & X_{n_1 1p} \\
\hdotsfor{5} \\
1 & X_{121} & X_{122} & \cdots & X_{12p} \\
1 & X_{221} & X_{222} & \cdots & X_{22p} \\
\vdots & & & & \\
1 & X_{n_2 21} & X_{n_2 22} & \cdots & X_{n_2 2p} \\
\hdotsfor{5} \\
\vdots & & & & \\
\hdotsfor{5} \\
1 & X_{1M1} & X_{1M2} & \cdots & X_{1Mp} \\
1 & X_{2M1} & X_{2M2} & \cdots & X_{2Mp} \\
\vdots & & & & \\
1 & X_{n_M M1} & X_{n_M M2} & \cdots & X_{n_M Mp}
\end{bmatrix}
\cdot
\begin{bmatrix}
\beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ .. \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ .. \\ \vdots \\ .. \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p
\end{bmatrix}
+
\begin{bmatrix}
1 & Z_{111} & Z_{112} & \cdots & Z_{11q1} \\
1 & Z_{211} & Z_{212} & \cdots & Z_{21q1} \\
\vdots & & & & \\
1 & Z_{n_1 11} & Z_{n_1 12} & \cdots & Z_{n_1 1q1} \\
\hdotsfor{5} \\
1 & Z_{121} & Z_{122} & \cdots & Z_{12q2} \\
1 & Z_{221} & Z_{222} & \cdots & Z_{22q2} \\
\vdots & & & & \\
1 & Z_{n_2 21} & Z_{n_2 22} & \cdots & Z_{n_2 2q2} \\
\hdotsfor{5} \\
\vdots & & & & \\
\hdotsfor{5} \\
1 & Z_{1M1} & Z_{1M2} & \cdots & Z_{1MqM} \\
1 & Z_{2M1} & Z_{2M2} & \cdots & Z_{2MqM} \\
\vdots & & & & \\
1 & Z_{n_M M1} & Z_{n_M M2} & \cdots & Z_{n_M MqM}
\end{bmatrix}
\cdot
\begin{bmatrix}
b_{11} \\ b_{21} \\ \vdots \\ b_{q1} \\ \cdots \\ b_{12} \\ b_{22} \\ \vdots \\ b_{q2} \\ \cdots \\ \vdots \\ \cdots \\ b_{1M} \\ b_{2M} \\ \vdots \\ b_{qM}
\end{bmatrix}
+
\begin{bmatrix}
\epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{n_1 1} \\ \cdots\cdots \\ \epsilon_{12} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{n_2 2} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ \epsilon_{1M} \\ \epsilon_{2M} \\ \vdots \\ \epsilon_{n_M M}
\end{bmatrix}
.
\tag{2.57}
$$

where $j = 1,2,...,M$ clusters; $i = 1,2,...,n_j$ subjects in cluster j; p is the number of fixed covariates, with $\beta = (\beta_0\beta_1...\beta_p)^T$ the vector of fixed parameters; q_j is the number of covariates with random effects, with $b_{q_j} = (b_{1j}b_{2j}...b_{qj})^T$ is vector of random effects; Y_{ij} is the response value for subject i in cluster j; X_{ij} is the value of observed fixed covariate for subject i in cluster j; Z_{ij} is the observed value of covariate with random effect for subject i in cluster j; and ϵ_{ij} is the unknown error for subject i in cluster j. The probability distributions for the error term, ϵ and random effect $\mathbf{b}$ as well as all other model assumptions are as provided in model (2.2).

The partitioned matrices and vectors in model (2.57) can also be expressed as: $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, ..., \mathbf{y}_m^T)^T$, i.e. $n_j \times 1$ component vectors $\mathbf{y}_j$ corresponding to the j-th cluster; $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T, ..., \mathbf{X}_m^T)^T$, i.e. $n_j \times p$ component matrices $\mathbf{X}_j$, and $\mathbf{Z} = (\mathbf{Z}_1^T, \mathbf{Z}_2^T, ..., \mathbf{Z}_m^T)^T$, $n_j \times q$ component matrices $\mathbf{Z}_j$ (Xiang et al., 2002; Zewotir, 2008). If cluster j is removed from the dataset, the entire design matrix $\mathbf{X}$ can be re-written as a set of two design matrices, that is, $\mathbf{X} = (\mathbf{X}_j^T, \mathbf{X}_{(j)}^T)^T$ comprising design matrix $\mathbf{X}_{(j)}^T$ without j-th cluster, and design matrix $\mathbf{X}_j$ for covariates data in cluster j. The same applies to all other relevant matrices and vectors, for example $\hat{\mathbf{e}} = (\hat{\mathbf{e}}_j^T, \hat{\mathbf{e}}_{(j)}^T)^T$ (Xiang et al., 2002). From the partitioned matrices, such as $\mathbf{X} = (\mathbf{X}_j^T, \mathbf{X}_{(j)}^T)^T$, one can note that the log-likelihood function $l_{(j)}(\hat{\beta}_{(j)})$ for β for the model on reduced sample is the function of both full data and the data for dropped cluster j. For illustration, one can think of the log-likelihood function $l_{(j)}(\hat{\beta})$ as the difference of the log-likelihood functions from the full data and the data from cluster j, i.e. $l_{(j)}(\hat{\beta}) = l(\hat{\beta}) - l_j(\hat{\beta})$. This means that the log-likelihood $l_{(j)}$ for $\beta_{(j)}$ can provide an estimate of impact of the dropped cluster j in the model.

A number of techniques exist for estimating the parameter displacement $\hat{\beta} - \hat{\beta}_{(j)}$ when cluster j is removed from analysis. One method that is used is the first-order Taylor series expansion of the score function of conditional log-likelihood function for reduced data $U_{\hat{\beta}_{(j)}}(.)$ evaluated at $\hat{\beta}$. The updated estimator $\hat{\beta}_{(j)}$ is obtained by solving for $\hat{\beta}_{(j)}$ in the first-order Taylor series expansion of the score function when it is equated to zero, see (Pregibon, 1981; Xiang et al., 2002). A first-order Taylor- series expansion of any univariate function $f(X)$ around a point $X = \alpha$ is a linear approximation of the value of the polynomial $f(X)$ or its gradient at point α, given by: $f(X) = f(\alpha) + \frac{d}{dX} f(\alpha)(X - \alpha)$. Since gradient of a curve $f(X)$ around a point α is an instantaneous change of the curve with respect of the variable X, the first-order Taylor series expansion concept is applied on the score function from the conditional log-likelihood function $l_{(j)}(\hat{\beta}_{(j)})$ to estimate the updated formula of the regression parameter resulting from removing a subject or cluster of subjects. The conditional log-likelihood function $l_{(j)}(\hat{\beta}_{(j)})$ for reduced data for linear mixed-effects model (2.2) is given by:

$$l_{(j)}(\beta_{(j)}|\mathbf{y}_{(j)}, \mathbf{X}_{(j)}, \mathbf{b}) = -\frac{n - n_j}{2} log(2\pi) - \frac{1}{2} log|\mathbf{G}_{(j)}| - \frac{1}{2}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)})^T \mathbf{G}_{(j)}^{-1}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}). \tag{2.58}$$

Then, the first-order Taylor series expansion of the score function obtained from the conditional log-likelihood function (2.58), evaluated at $\hat{\beta}$ and treating $\mathbf{G}_{(j)}$ as nuisance parameter (Xiang et al., 2002), is given by:

$$\begin{aligned} U_{\hat{\beta}_{(j)}}(\hat{\beta}_{(j)}) &= \frac{\partial l_{(j)}(\hat{\beta})}{\partial \hat{\beta}_{(j)}} + \frac{\partial^2 l_{(j)}(\hat{\beta})}{\partial \hat{\beta}^T_{(j)} \partial \hat{\beta}_{(j)}}(\hat{\beta} - \hat{\beta}_{(j)}) \\ &= \mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}) - \mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)}(\hat{\beta} - \hat{\beta}_{(j)}). \end{aligned} \tag{2.59}$$

Therefore, the updated parameter estimator $\hat{\beta}_{(j)}$ is approximated by equating the equation (2.59) to zero and solve for $\hat{\beta}_{(j)}$ as follows:

$$\begin{aligned} \mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)}(\hat{\beta} - \hat{\beta}_{(j)}) &= \mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}) \\ \hat{\beta} - \hat{\beta}_{(j)} &= (\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}) \\ \therefore \hat{\beta}_{(j)} &= \hat{\beta} - (\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}) \\ &= \hat{\beta} - (\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\hat{\mathbf{e}}_{(j)}. \end{aligned} \tag{2.60}$$

Once the updating formulae are solved, the influence measures for clusters are developed from the usual quantities for examining influence defined in previous section, such as Cook's distance, DFBetas, and DFFits (Xiang et al., 2002; Zewotir, 2008). For example, the generalised Cook's distance for $\hat{\beta}$ for data without j-th cluster in linear mixed-effects model (Xiang et al., 2002; Zewotir, 2008) can be estimated as:

$$\begin{aligned} CD_j(\hat{\beta}) &= \frac{(\hat{\beta} - \hat{\beta}_{(j)})^T(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})(\hat{\beta} - \hat{\beta}_{(j)})}{p\hat{\sigma}^2_e} \\ &= \frac{\left((\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\hat{\mathbf{e}}_{(j)}\right)^T(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})\left((\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\hat{\mathbf{e}}_{(j)}\right)}{p\hat{\sigma}^2_e} \\ &= \frac{\hat{\mathbf{e}}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)}(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\hat{\mathbf{e}}_{(j)}}{p\hat{\sigma}^2_e} \\ &= \frac{\hat{\mathbf{e}}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)}(\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}^T_{(j)}\mathbf{G}^{-1}_{(j)}\hat{\mathbf{e}}_{(j)}}{p\hat{\sigma}^2_e} \\ &= \frac{\hat{\mathbf{e}}^T_{(j)}\mathbf{G}^{-1}_{(j)}\mathbf{Q}_{1(j)}\hat{\mathbf{e}}_{(j)}}{p\hat{\sigma}^2_e}, \end{aligned} \tag{2.61}$$

where $\mathbf{Q}_{1(j)} = \mathbf{X}_{(j)}(\mathbf{X}_{(j)}^T\mathbf{G}_{(j)}^{-1}\mathbf{X}_{(j)})^{-1}\mathbf{X}_{(j)}^T\mathbf{G}_{(j)}^{-1}$ is leverage matrix for fixed effects solved in equation (2.46) but without cluster j. Large values of $CD_j(\hat{\beta})$ in equation (2.61) show that subjects in cluster j are jointly influential on $\hat{\beta}$ (Zewotir, 2008).

A similar approach can be used to find the updating formulae for the other model parameter estimators, such as estimated variance of the error term $\hat{\sigma}_\epsilon^2$. The same results can also be found using direct application of properties of multivariate normal distribution on reduced data (Zewotir & Galpin, 2007). As for the random effects $\mathbf{b}$, the linear mixed-effects model (2.2) assumes that these are mutually independent across clusters, hence deleting one cluster will not affect the estimator $\hat{\mathbf{b}}$ for the remaining clusters (Xiang et al., 2002). This is demonstrated below, using the method of first-order Taylor-series expansion on score function obtained from the complete joint log-likelihood function (2.19). The first-order Taylor-series expansion of score function for $\hat{\mathbf{b}}_{(j)}$ resulting from the conditional log-likelihood $l_{(j)}(\hat{\mathbf{b}}_{(j)})$ for reduced data, evaluated at $\hat{\mathbf{b}}$, is given by:

$$\begin{aligned} U_{\hat{\mathbf{b}}_{(j)}}(\hat{\mathbf{b}}_{(j)}) &= \frac{\partial l_{(j)}(\hat{\mathbf{b}})}{\partial \hat{\mathbf{b}}_{(j)}} + \frac{\partial^2 l_{(j)}(\hat{\mathbf{b}})}{\partial \hat{\mathbf{b}}_{(j)}^T \partial \hat{\mathbf{b}}_{(j)}}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(j)}) \\ &= \mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2 \mathbf{I}_{(j)})^{-1}(\mathbf{y}_{(j)} - \mathbf{X}_{(j)}\hat{\beta}_{(j)}) - \left[\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}\mathbf{Z}_{(j)} + \mathbf{D}_{(j)}^{-1}\right]\hat{\mathbf{b}}_{(j)} \\ &\quad - \left[\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}\mathbf{Z}_{(j)} + \mathbf{D}_{(j)}^{-1}\right](\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(j)}). \end{aligned} \tag{2.62}$$

Therefore, the updated formula $\hat{\mathbf{b}}_{(j)}$ for $\hat{\mathbf{b}}$ will be found by equating the quantity (2.62) to zero and solving for $\hat{\mathbf{b}}_{(j)}$ as follows:

$$\begin{aligned}
&\left[\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}\mathbf{Z}_{(j)}+\mathbf{D}_{(j)}^{-1}\right]\hat{\mathbf{b}}_{(j)}+\left[\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}\mathbf{Z}_{(j)}+\mathbf{D}_{(j)}^{-1}\right](\hat{\mathbf{b}}-\hat{\mathbf{b}}_{(j)})\\
&=\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}(\mathbf{y}_{(j)}-\mathbf{X}_{(j)}\hat{\beta}_{(j)})\\
\hat{\mathbf{b}}_{(j)}+(\hat{\mathbf{b}}-\hat{\mathbf{b}}_{(j)})&=\left[\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}\mathbf{Z}_{(j)}+\mathbf{D}_{(j)}^{-1}\right]^{-1}\mathbf{Z}_{(j)}^T(\sigma_{\epsilon(j)}^2\mathbf{I}_{(j)})^{-1}(\mathbf{y}_{(j)}-\mathbf{X}_{(j)}\hat{\beta}_{(j)})\\
\therefore\hat{\mathbf{b}}&=\mathbf{D}_{(j)}\mathbf{Z}_{(j)}^T\mathbf{G}_{(j)}^{-1}(\mathbf{y}_{(j)}-\mathbf{X}_{(j)}\hat{\beta}_{(j)})\\
\hat{\mathbf{b}}&=\hat{\mathbf{b}}_{(j)}.
\end{aligned} \tag{2.63}$$

This result shows that the updated formula $\hat{\mathbf{b}}_{(j)}$ for $\hat{\mathbf{b}}$, upon dropping cluster j, is just the same formula $\hat{\mathbf{b}}$ obtained when using all available clusters. This implies that there is no change in the estimator for random effects when a cluster is dropped from analysis. This means that the prediction of random effects $\mathbf{b}$ for each of the available clusters is insensitive to any other cluster that might have been dropped from the dataset. Such is the case due to independence of the random effects across clusters (Xiang et al., 2002).

A version of Pena's statistic for influence of individual subjects on the linear mixed-effects model (2.2) follows naturally from the definition (2.51) (Turkan & Toktamis, 2012). The Pena's measure for linear mixed-effects model (2.2) is thus given by:

$$\begin{aligned}
a_i&=\frac{1}{(p+q)\hat{\sigma}_\epsilon^2}||\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(i)}||^2\\
&=\frac{\lambda_i^{*2}(1-r_{ii})}{(p+q)(r_{ii})}\\
&=\frac{e_i^{*2}(1-r_{ii})}{(p+q)\hat{\sigma}_\epsilon^2(r_{ii})^2}\\
&=\frac{(1-r_{ii})^2e_i^{*T}e_i^*}{(p+q)\hat{\sigma}_\epsilon^2r_{ii}^2(1-r_{ii})},
\end{aligned} \tag{2.64}$$

where $1-r_{ii}$ is i-th subject leverage defined in equation (2.22); $\hat{e}_i^*=r_{ii}Y_i$ is i-th subject's displacement of $\hat{Y}_i$ due to removal of subject i from analysis; $\lambda_i^*=\hat{e}_i^*/\hat{\sigma}_\epsilon\sqrt{r_{ii}}$

is the Studentised displacement. Large values of the measure (2.64) will show subjects that have influence on the fitted value. By revisiting various cutoffs, the reliability of Pena's residual in tracking outliers is also reported in the work of Das & Gogoi (2015). The implementation packages for group or individual subjects influence methods for linear mixed-effects model are available in literature. For example, Schabenberger (2005) uses the `SAS program MIXED` to compute the multivariate $DFFITS$ statistic corresponding to removal of a group of observations from the model. While Loy & Hofmann (2014) use `the R package HLMdiag` to implement the diagnostics.

Throughout this review, it is clear that the residuals for the linear mixed-effects model (2.2) are direct extensions of those for generalised linear model (2.1). This reflects the relationship of the structures and estimation procedures for both models. The review also shows that influence and outlier assessment methods for linear and linear mixed-effects model are well-studied. There is just little work done on diagnostics for non-linear mixed models. In the next section, the current application of model diagnostics in clustered survival data is reviewed.

2.6 Application to clustered survival data in Malawi

Standard outlier and influence statistics for survival data were implemented on child survival, a major indicator of health and development of a country, collected as part of 2015-16 Malawi Demographic and Health Survey (MDHS) data. Malawi is a landlocked country in south-eastern Africa in the Great Rift Valley and lies on the western shores of Lake Malawi. The country is bordering Tanzania to the north, Zambia to the west, and Mozambique to the east, south, and west. The population of Malawi was just over 17 million in 2018, representing intercensal growth rate of 2.9% per annum between the previous housing and population census of 2008 and the recent one of 2018 (Malawi National Statistical Office (NSO), 2019). Using this estimated growth rate, the population is expected to double by 2042. Over 80% of the Malawi's population is rural, and with 64% under the age of 15 years, thus the country has a young population.

Administratively, Malawi is divided into the Northern, Central and Southern regions, which are further divided into twenty-eight districts, namely: Balaka, Blantyre, Chikwawa, Chiradzulu, Machinga, Mangochi, Mulanje, Mwanza, Neno, Nsanje, Phalombe, Thyolo, and Zomba in the Southern region; Dedza, Dowa, Kasungu, Nkhotakota, Ntcheu, Ntchisi, Lilongwe, Mchinji, and Salima in the Central region; and Chitipa, Karonga, Likoma, Mzimba, Nkhatabay, and Rumphi in the Northern region. Four of the districts, namely: Blantyre, Lilongwe, Mzimba, and Zomba contain the four major cities, which themselves are further divided into rural and city locations. Figure A.1 in the Appendix shows the map of Malawi with the 28 districts and the four cities. The economy of Malawi is largely dependent on agriculture, fishing and forestry and the country's GDP is one of the lowest in sub-Saharan Africa. The very low GDP places pressure on the delivery of health care system, which is based on primary health care (PHC), largely operated within the 28 districts and 4 cities (Makaula et al., 2019).

The 2015-16 MDHS survey, which was the fifth since 1992, aimed to provide data for monitoring the population and health status of the country. The survey was held from 19 October 2015 to 18 February 2016 and collected child survival data from the women respondents and caregivers who provided birth histories. For the purpose of this work, mortality data on 17,286 children, who were born within the last 5 years of the survey, were analysed. The survey employed a two-stage stratified sampling design, with emuneration areas as primarily sampling units and households as secondary sampling units, having all women aged 15-49 years being eligible to participate in the survey. Further information on the 2015-16 MDHS can be found in the survey report (Malawi National Statistical Office (NSO) &

ICF, 2017), and the information about the DHS progam and data access are available at `www.DHSprogram.com`.

In order to balance between sufficient clusters and number of children per cluster, the rural and urban areas in each district were taken as separate groupings or clusters. Thus, we used the resulting 52 subdistricts (clusters) on which to analyze the child survival data from the 2015-16 MDHS. Child birth order and sex were used in the analysis of child survival because previous studies had indicated that these are some of the well-known predictors of child survival (Manda, 2001). The survival model was fitted to the dataset and cluster outlier or influence was assessed for each sub-district using available methods.

2.6.1 Using random effects residuals from frailty model

One of the statistics that are used for group outlier examination for clustered data is the random effect residual discussed in Section 2.4.1. A Cox frailty model was fitted to the 2015-16 MDHS data, with event of interest being death of a child from any cause before 60 months of age. The event-time was age in months as at death or censoring point. The ages-at-death that were recorded as zero months were transformed into random $Uniform(0,1)$ values to reflect proportions of month-days lived before death or censoring by the corresponding children. The data had about 5% children who experienced the event of death. Administrative censoring was used, and children who were still alive or had survived up to 60 months were censored. The covariates were birth order and sex of the child. The fitted model was as follows:

$$\begin{aligned} h_{ij}(age) = h_0(age)exp(&-0.185 \times Female - 0.214 \times Birthorder \\ &+ 0.0233 \times Birthorder_{squared} + subdistrict). \end{aligned} \tag{2.65}$$

The model results showed that female children had significantly lower risk of death than the male children (p-value < 0.0096). While higher birth order was associated with reduced risk of death (p-value < 0.0001) and birth order squared with increased risk. The relationship between birth order and logarithm of hazard of death was therefore quadratic. The results are consistent with previous findings (Manda, 1999). The variance of sub-district random-effects was 0.0419 and it was significantly different from zero (p-value < 0.001). The scatter plots in Figure 2.7 for estimates of random effects showed that *Neno* urban was an outlier to the survival mixed model based on random effects estimates. These results were reserved for comparison when applying the derived group outlier statistic to the same data.

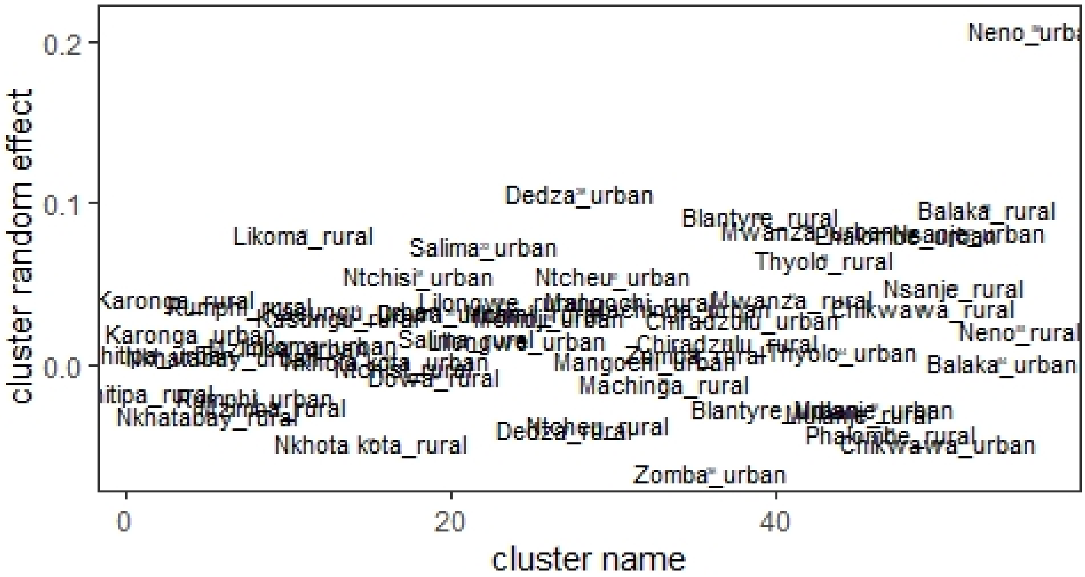


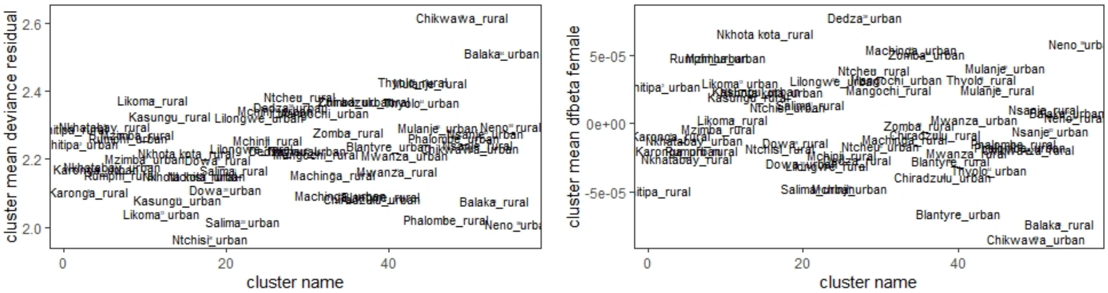
Figure 2.7: Sub-district level random effect residual from fitting a frailty Cox hazard regression model to Malawi child survival data, 2015-16 MDHS. Source: Researcher

2.6.2 Using group summary statistics of residuals from univariate Cox model

The other method for examining group outlier and influence for clustered survival data involves fitting a univariate Cox model (2.8) to the data and compute group summary statistics of the residuals such as deviance and DFBetas (Jennings, 1986; Langford & Lewis, 1998; Duchateau & Janssen, 2005; Legrand et al., 2006). The univariate Cox model was fitted to the 2015-16 MDHS data using the same event of interest and covariates as in previous section. The fitted model was:

$$
\begin{aligned}
h_{ij}(age) = h_0(age)exp(&-0.182 \times Female - 0.212 \times Birthorder \\
&+ 0.0234 \times Birthorder_{squared}).
\end{aligned}
\tag{2.66}
$$

As with frailty model (2.65), effects of female gender ($p - value = 0.011$), birth order ($p - value < 0.0001$), and birth order squared ($p - value < 0.0001$) on child risk to death were also significant. The only difference was that the sizes of the fixed effects were slightly larger in the univariate Cox PH model comapred to the multivariate Cox PH model. The subdistrict unweighted averages of the model's deviance residual and DFBetas were computed. The results in Figure 2.8 (a) show that *Chikwawa* rural and *Balaka* urban were under-five mortality outliers based on a cutoff of 2.5 for the cluster average deviance residual from the univariate survival model. While, the average DFBetas in Figure 2.8 (b) indicate that *Chikwawa* urban and *Balaka* rural were marginally influential on the effect of female gender on child survival. The results were also reserved for comparison when applying, on the same dataset, the proposed group outlier and influence statistics developed in this study.



(a) Cluster-wise average deviance residuals from fitting a univariate Cox hazard regression model to Malawi child survival data, 2015-16 MDHS.

(b) Cluster-wise average DFBetas for influence of female effect on log hazard in univariate Cox model applied on 2015-16 MDHS.

Figure 2.8: Plots of average deviance and dfbetas residuals per cluster upon fitting a univariate Cox model to 2015-16 MDHS. Source: Researcher

Chapter 3

Cluster Outliers for Survival Mixed Model

This chapter presents a method for detecting outlying clusters in grouped survival data. The chapter begins by defining the important statistics for group outlier analysis and later presents the suggested measure.

3.1 Useful definitions for studying group outliers

The review in Chapter 2 revealed that the outlier concept is to do with subject(s) not conforming to the distributional assumption of the fitted model (Langford & Lewis, 1998; Sarkar et al., 2011; Aguinis et al., 2013; Z. Zhang, 2016). This is now examined using post-estimation statistics that can capture the distribution of the model's fitted value or residual. One such statistic that is used in generalised linear models is the residual $\hat{\mathbf{e}}$ in equation (2.8) given by:

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{I}_n - \mathbf{W})\mathbf{y},\end{aligned} \tag{3.1}$$

where $W = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix, or its scaled version in equation (2.10) given by:

$$\begin{aligned}\lambda &= [var(\hat{\mathbf{e}})]^{-1/2}\hat{\mathbf{e}} \\ &= \left[(\mathbf{1}-\mathbf{w}_{ii})^T var(\mathbf{y})(\mathbf{1}-\mathbf{w}_{ii})\right]^{-1/2}\hat{\mathbf{e}} \\ &= \hat{\sigma}_\epsilon^{-1}(\mathbf{1}-\mathbf{w}_{ii})^{-1/2}\hat{\mathbf{e}},\end{aligned} \tag{3.2}$$

where $\mathbf{w}_{ii}$ is a vector consisting of diagonal elements of $\mathbf{W}$ (Loy & Hofmann, 2014).

The primary purpose of the residual (3.1) is to assess linearity and additivity assumptions of the general linear model (2.1) (Yang, 2012). However, it is also used to assess outliers due to the fact that it is an estimate of the model's error term, whose probability distribution is assumed to be normal with mean zero and constant variance. So, subjects that are in the periphery of the scatter plot of estimated errors $\hat{e}_i$ against individual indexes i are considered outliers to the model. In linear mixed-effects model (2.2), the residual $\hat{\mathbf{e}}$ given in equation (2.23) is:

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{b}. \tag{3.3}$$

This residual also serves to examine linearity and additivity assumptions of the linear mixed-effects model (2.2) as in generalised linear model (2.1) (Nobre & Singer, 2011; Turkan & Toktamis, 2012; Loy & Hofmann, 2014). For the univariate Cox PH model (2.11), a residual is defined as in equation (2.12):

$$m(t_i) = N(t_i) - \int_0^{t_i} Y_i(t)exp(X_i^T(t)\hat{\beta})d\hat{H}_0(t), \tag{3.4}$$

where $N(t_i)$ is a counting process for the i-th subject indicating number of observed events experienced over time t_i, $Y_i(t)$ is a 0-1 process indicating whether the i-th subject is at risk at time t_i and the Cox model restricts that $Y_i(t) = 1$ until the first event or censoring and 0 thereafter, $\hat{\beta}$ regression coefficients, $X_i^T(t)$ p-dimensional covariate processes, and $\hat{H}_0(t)$ the baseline cumulative hazard function that is unspecified (Therneau et al., 1990).

The residual is called martingale because of its relation with a counting process. It is interpreted as the difference over $[0, t_i]$ in the observed and expected number of events at each time t_i given the model, or as excess events (Therneau et al., 1990). Thus, positive values imply individuals failed earlier than expected and negative values means they survived longer than estimated. Just as in linear models, the residual (3.4) has also the properties of summing to zero, having an average of zero and with no correlation between any two of its values at any given time point (Therneau et al., 1990). For this reason, the residual (3.4) is also used to assess the linearity and additivity assumptions in survival models. When plotted against each covariate X, the values are expected to average around zero where a covariate has correct linear specification. Once consideration is on the Cox model with time-independent covariates, the martingale residual (3.4) reduces to:

$$m(t_i) = \delta_i - \hat{H}_0(t) exp(X_i^T \hat{\beta}), \tag{3.5}$$

where δ_i is the final status of subject i and t_i the observation time for subject i.

The natural extension of martingale residual (3.5) to clustered survival model (1.1) is defined as:

$$m(t_{ij}) = \delta_{ij} - \hat{H}_0(t)exp(X_{ij}^T\hat{\beta} + Z_{ij}^T\hat{b}_j)$$

$$\Rightarrow \begin{bmatrix} m(t_{11}) \\ \vdots \\ m(t_{n_1 1}) \\ m(t_{12}) \\ \vdots \\ m(t_{n_2 2}) \\ \vdots \\ m(t_{1M}) \\ \vdots \\ m(t_{n_M M}) \end{bmatrix} = \begin{bmatrix} \delta_{11} - \hat{H}_0(t)exp(X_{11}^T\hat{\beta} + Z_{11}^T\hat{b}_1) \\ \vdots \\ \delta_{n_1 1} - \hat{H}_0(t)exp(X_{n_1 1}^T\hat{\beta} + Z_{n_1 1}^T\hat{b}_1) \\ \delta_{12} - \hat{H}_0(t)exp(X_{12}^T\hat{\beta} + Z_{12}^T\hat{b}_2) \\ \vdots \\ \delta_{n_2 2} - \hat{H}_0(t)exp(X_{n_2 2}^T\hat{\beta} + Z_{n_2 2}^T\hat{b}_2) \\ \vdots \\ \delta_{1M} - \hat{H}_0(t)exp(X_{1M}^T\hat{\beta} + Z_{1M}^T\hat{b}_M) \\ \vdots \\ \delta_{n_M M} - \hat{H}_0(t)exp(X_{n_M M}^T\hat{\beta} + Z_{n_M M}^T\hat{b}_M) \end{bmatrix}. \qquad (3.6)$$

The properties of the residual (3.5) may not apply to the extended version (3.6) due to correlation of subjects resulting from shared random effect in a cluster. In both univariate and multivariate cases, the martingale residual is negatively-skewed because δ_i has values 0 or 1 while $\hat{H}_0(t)exp(X_i^T\hat{\beta})$ has values in the interval $[0, \infty)$. Due to this skewed distribution, the martingale statistics (3.5) and (3.6) may not ably serve to examine outliers.

The review in Section 2.4 showed that apart from Studentised residual (3.2), a deviance residual is also used for outlier assessments in generalised linear models. As indicated in the stated section, the deviance residual measures the disagreement between an element of the log-likelihood of the fitted model and the corresponding element of the log-likelihood that would result if each point were fitted exactly (Sarkar et al., 2011). For example, a deviance statistic for logistic regression presented as equation (2.35) is given by:

$$d_i = sign(Y_i - \hat{\theta}(X_i))\left[-2(Y_i log\hat{\theta}(X_i)) + (1 - Y_i)log(1 - \hat{\theta}(X_i))\right]^{\frac{1}{2}} \qquad (3.7)$$

where $\hat{\theta}(X_i) = \frac{exp(X^T\hat{\beta})}{1+exp(X^T\hat{\beta})}$ is the fitted conditional probability of success given covariate X, Y_i is binary response taking values 0 or 1, and 'sign' the sign of raw residual $Y_i - \hat{\theta}(X_i)$, plus or minus. The deviance residual (3.7) is expected to be symmetric around the mean zero, hence marginal points in its distribution are regarded as outliers.

A similar version of deviance residual was suggested by Therneau et al. (1990) for Cox PH model, it is a transformation of a martingale residual given in equation (2.36) as:

$$d_i = sgn(m(t_i))\,[-2(m(t_i) + \delta_i log(\delta_i - m(t_i))]^{\frac{1}{2}}\,. \tag{3.8}$$

From this version of the deviance residual, an extension for the clustered survival model (1.1) is defined in stacked vector form as:

$$\begin{aligned} d_{ij} &= sgn(m(t_{ij}))\,[-2(m(t_{ij}) + \delta_{ij} log(\delta_{ij} - m(t_{ij}))]^{\frac{1}{2}} \\ \Rightarrow \begin{bmatrix} d_{11} \\ \vdots \\ d_{n_1 1} \\ d_{12} \\ \vdots \\ d_{n_2 2} \\ \vdots \\ d_{1M} \\ \vdots \\ d_{n_M M} \end{bmatrix} &= \begin{bmatrix} sgn(m(t_{11}))[-2(m(t_{11}) + \delta_{11} log(\delta_{11} - m(t_{11})))]^{1/2} \\ \vdots \\ sgn(m(t_{n_1 1}))[-2(m(t_{n_1 1}) + \delta_{n_1 1} log(\delta_{n_1 1} - m(t_{n_1 1})))]^{1/2} \\ sgn(m(t_{12}))[-2(m(t_{12}) + \delta_{12} log(\delta_{12} - m(t_{12})))]^{1/2} \\ \vdots \\ sgn(m(t_{n_2 2}))[-2(m(t_{n_2 2}) + \delta_{n_2 2} log(\delta_{n_2 2} - m(t_{n_2 2})))]^{1/2} \\ \vdots \\ sgn(m(t_{1M}))[-2(m(t_{1M}) + \delta_{1M} log(\delta_{1M} - m(t_{1M})))]^{1/2} \\ \vdots \\ sgn(m(t_{n_M M}))[-2(m(t_{n_M M}) + \delta_{n_M M} log(\delta_{n_M M} - m(t_{n_M M})))]^{1/2} \end{bmatrix}. \end{aligned} \tag{3.9}$$

Once again, the deviance residual (3.9) does not have same properties as its counterpart (3.8) for univariate Cox model because subjects in a cluster are correlated. Hence, assessment of individual outliers within a cluster for the mixed survival model is not a straightforward task. However, the values of the residual (3.9) are uncorrelated across clusters. The concern of this work was on group outlier analysis. By utilising the independence of values of the measure (3.9) across clusters, a statistic for assessing group outliers in multivariate Cox model (1.1) is developed and presented in the next section.

3.2 Proposed outlier statistic for multivariate survival data

There are a number of techniques that are used to assess group outliers in mixed models. One way is through graphically assessing the homogeneity of the distribution of standardised residuals of single observations in a linear mixed-effects model plotted against each cluster (Langford & Lewis, 1998), given by:

$$\lambda_{ij} = \hat{\mathbf{e}}/stdev(\hat{\mathbf{e}}), \tag{3.10}$$

where $stdev(\hat{\mathbf{e}})$ is Studentized residual of a subject. The clusters with highly skewed Studentized residuals compared to others, are considered outliers to the linear mixed model (Langford & Lewis, 1998).

A similar method is the re-weighted minimum covariance determinant (RMCD) (Cerioli, 2010), given by:

$$d^2_{i(RMCD)} = (Y_i - \hat{\mu}_{(RMCD)})^T \hat{\Sigma}^{-1}_{(RMCD)} (Y_i - \hat{\mu}_{(RMCD)}) \tag{3.11}$$

where $\hat{\mu}_{(RMCD)} = \frac{1}{m}\sum_{i \in Y_{MCD}} \varpi_i Y_i$ is re-weighted MCD estimate of location; $\hat{\Sigma} = \frac{k_{(RMCD)(m,n,v)}}{m-1} \times \sum_{i \in Y_{MCD}} \varpi_i (Y_i - \hat{\mu}_{(RMCD)})^T (Y_i - \hat{\mu}_{(RMCD)})$ is re-weighted MCD estimate of scatter; $m = \sum_{i \in Y_{MCD}} \varpi_i$; v dimension of covariance matrix $\mathit{\Sigma}$; and $k_{(RMCD)(m,n,v)}$ proportionality constant to control for bias. This method tests whether or not a group of subjects belongs to a subsample of homogeneous units with constant variability, referred to as 'good' observations. The null hypothesis is $H_{0i} : Y_i \sim N(\mu, \Sigma)$ and $d^{*2}_{(RMCD)}$ is test statistic. When H_{0i} is not rejected, it means the subsample Y_i is a 'good' observation, otherwise the group of subjects being assessed is deemed outlier (Cerioli, 2010).

The overalaps of scatter plots of residuals from different clusters is a major setback for application of method (3.10), as one may not reliably conclude on outlierness of a cluster when plots of its standardised residuals overlap with those of another cluster. Similarly, application of method (3.11) on regrouped subsamples of data of size greater than half of the total sample size implies that the technique ignores natural groups in the data, some of which may have lower sample sizes than half of the total sample. Outlier detection methods that can be applied on clusters of data are crucial in studying how behaviours of subjects in the clusters affect the modelling. Nonetheless, both methods (3.10) and (3.11) transform some known single observations residuals into distance quantity that can examine grouped outliers to the mixed-effects model. Schall & Dunne (1988) demonstrated that when a linear model is fitted to any residual $\hat{\mathbf{e}}$ that is normally distributed, the model's diagnostic assessments can reveal outliers to the main model with response $\mathbf{y}$. The deviance residual (3.9) is one of the diagnostic statistic that is symmetric about zero and has asymptotic mean of zero (Therneau et al., 1990), so this study suggests an outlier statistic for model (1.1) by manipulating further the extended deviance residual (3.9).

Following from the ideas developed for linear mixed-effects models, this study proposes a statistic computed from a ratio of within-cluster variance of deviance residual (3.9) to between-cluster variance for examining outlying clusters to model (1.1). If observations in model (1.1) were independent, the total variation of d_{lj} would have been the sum of within-cluster variation and between-cluster variation given by:

$$\frac{\sum_{i=1}^{M}\sum_{j=1}^{n_i}(d_{ij}-\bar{\bar{d}})^2}{n-1} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{n_i}(d_{ij}-\bar{d}_i)^2}{n-M} + \frac{\sum_{i=1}^{M} n_i(\bar{d}_i-\bar{\bar{d}})^2}{M-1}, \qquad (3.12)$$

where $\bar{\bar{d}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{n_i} d_{ij}}{n}$ is the grand mean of the deviance residual d_{ij}; $\bar{d}_i = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i}$ is the mean of d_{ij} for any fixed i; $n = n_1 + n_2 + ... + n_M$ is number of subjects in entire dataset.

However, the correlations of observations in model (1.1) will yield biased estimate of within-cluster variance of d_{ij} in equation (3.12) for entire dataset. Since the clusters are independent and assuming conditional independence of observations in each cluster, the respective within-cluster variances of residual d_{ij} will be unbiased estimators of variance of d_{ij} in each cluster. These will consequently measure how distant the survival times of subjects in each cluster are from the fitted survival curve. Therefore, the proposed group outlier statistic for model (1.1) is an $M \times 1$ vector, denoted $\mathbf{k}_i$, which is the ratio of within-cluster to between-cluster variances of d_{il} given by:

$$\begin{aligned}\mathbf{k} &= \frac{1}{L}(k_1, ..., k_M)^T \\ &= \frac{1}{L}\left(\frac{\sum_{i=1}^{n_1}(d_{1j}-\bar{d}_1)^2}{n_1-1}, ..., \frac{\sum_{i=1}^{n_M}(d_{Mj}-\bar{d}_M)^2}{n_M-1}\right)^T,\end{aligned} \qquad (3.13)$$

where $L = \frac{\sum_{j=1}^{M} n_j(\bar{d}_j-\bar{\bar{d}})^2}{M-1}$ is between-cluster variance of d_{ij}.

Since the fitted survival curve is expected to pass through all available clusters of observations, the proposed statistic (3.13) will separate homogeneous clusters from the outlying clusters in view of the fitted survival mixed model (Rousseeuw & Hubert, 2011). The small values of k_j will correspond to well-fitted clusters of observations, that is, those units that closely span the fitted survival curve. While large values of (3.13) will correspond to clusters whose observations have been poorly fitted by model (1.1), and hence outliers.

We explored properties of $k_j = f(K_j, L) = K_j/L$, where K_j is the within-cluster variance component of (3.13). Clearly, $\mathbf{k}_j \in [0, \infty)$ and it is a non-linear function, since K_j and L, being variances, have support $[0, \infty)$. A common method to estimate expected value of ratio estimator is through second order Taylor series expansion about $\mu = (\mu_{k_j}, \mu_l)$ (Van Kempen & Van Vliet, 2000). Thus,

$$\begin{aligned}
E(k_j) &= E(K_j/L) \\
&= E(f(K_j, L)) \\
&\approx E[f(\mu) + f'_{k_j}(\mu)(k_j - \mu_{k_j}) + f'_l(\mu)(l - \mu_l) + \frac{1}{2}\{f''_{k_jk_j}(\mu)(k_j - \mu_{k_j})^2 \\
&+ 2f''_{lk_j}(\mu)(k_j - \mu_{k_j})(l - \mu_l) + f''_{ll}(\mu)(l - \mu_l)^2\}] \\
&= f(\mu) + \frac{1}{2}\left\{f''_{k_jk_j}(\mu)Var(K_j) + 2f''_{lk_j}(\mu)Cov(L, K_j) + f''_{ll}(\mu)Var(L)\right\} \\
&= \frac{\mu_{k_j}}{\mu_l} - \frac{1}{\mu_l^2}Cov(K_j, L) + \frac{\mu_{k_j}}{\mu_l^3}Var(L),
\end{aligned} \tag{3.14}$$

where $f(\mu) = \mu_{k_j}/\mu_l$, $f''_{k_jk_j}(\mu) = 0$, $f''_{lk_j}(\mu) = -1/(\mu_l)^2$, and $f''_{ll}(\mu) = 2\mu_{k_j}/(\mu_l)^3$, since $f(K_j, L) = K_j/L$ and $E(K_j/L) = E(f(K_j, L))$. Also, $E(k_j - \mu_{k_j}) = E(l - \mu_l) = 0$; $Var(K_j) = E(k_j - \mu_{k_j})^2$, and $Cov(K_j, L) = E[(k_j - \mu_{k_j})(l - \mu_l)]$. For variance of k_j, it follows from the equation of mean above and from first order Taylor series expansion of $f(K_j, L)$ around $\mu = (\mu_{K_j}, \mu_l)$ that

$$\begin{aligned}
Var(k_j) &= Var(K_j/L) \\
&= Var(f(K_j, L)) \\
&= E\{[f(K_j, L) - f(\mu)]^2\} \\
&\approx E\left\{\left[f(\mu) + f'_{k_j}(\mu)(k_j - \mu_{k_j}) + f'_l(\mu)(l - \mu_l) - f(\mu)\right]^2\right\} \\
&= f'^2_{k_j}(\mu)Var(K_j) + 2f'_{k_j}(\mu)f'_l(\mu)Cov(K_j, L) + f'^2_l(\mu)Var(L) \\
&= \frac{1}{\mu_l^2}Var(K_j) - 2\frac{\mu_{k_j}}{\mu_l^3}Cov(K_j, L) + \frac{\mu_{k_j}^2}{\mu_l^4}Var(L).
\end{aligned} \tag{3.15}$$

These properties and others such as estimates of third and fourth moments of k_j can help in characterising the distribution of k_j, which can in turn provide a basis for formal tests about outliers to model (1.1). Nonetheless, graphical methods also provide reliable alternative to formal tests of model residuals (Yang, 2012). Searching for a fixed critical point for determining an outlier using a residual becomes relevant when there are few isolated cases in the dataset (Zewotir & Galpin, 2007). Where the data has many outlying cases, use of multiple comparisons of the residual values relative to one another, and through graphical displays, is recommended (Zewotir & Galpin, 2007). The graphical methods are known to provide reliable alternative to formal tests of model residuals (Yang, 2012). For these reasons, this study engaged graphical assessments were engaged in this study to analyse outlying clusters to model (1.1). In practice, relative comparisons of values of a group outlier statistic suffice to isolate outlying groups to mixed models (Zewotir & Galpin, 2007). Hence, this study applied graphical techniques on values of the proposed outlier statistic $\mathbf{k}$ to assess the outlying clusters in relevant datasets.

3.3 Simulation study

In order to evaluate performance of the proposed outlier statistic, a simulation study was carried out. There are many examples in literature for survival-times data simulation techniques (Bender et al., 2005; M. J. Crowther & Lambert, 2012, 2013; Cho et al., 2009; Moriña & Navarro, 2014; Montez-Rath et al., 2017; Wan, 2017; Brilleman et al., 2018). A shared frailty survival model was assumed in order to generate survival times T. Two covariates were used, X_1 generated from $Bernoulli(0.7)$ and X_2 from $N(0,1)$. The cluster random effects $\mathbf{b}$ were generated from $N(0, 0.4^2)$. The survival time data T were generated from the Exponential(1) distribution, using the cumulative hazard inversion method (Brilleman et al., 2018) on the model:

$$h_{ij}(t|b_j, X_i) = h_0(t) exp(\beta_1 X_{ij1} + \beta_2 X_{ij2} + b_j) \tag{3.16}$$

where $h_0(t) \sim Weibull(0.1, 1)$, i.e. $h_0(t) = act^{c-1}$, with $a = 0.1$ and $c = 1$ making $h_0(t) = 0.1$ a constant; $\beta_1 = 0.5$ and $\beta_2 = 1$. The inversion method derived t_{ij} from $t^*_{ij} = H^{-1}_{ij}(-log(S(t_{ij})))$, where $S(t_{ij}) \sim Uniform(0, 1)$ and hence making $H_{ij}(t) = -log(Uniform(0, 1)) \sim Exponential(1)$ (Brilleman et al., 2018).

The random censoring variable Δ was generated from $Bernoulli(0.4)$, giving a censoring rate of 60%. This rate was chosen because few cluster surveys of various populations in Africa have reported an average failure rate of 40%, when the event-time is death (Manda & Meyer, 2005). Other methods for generating censoring variable are available in literature (Montez-Rath et al., 2017; Wan, 2017). For instance, administrative censoring, where the study end point is defined and a censoring variable is created that gets a value of 0 for subjects' survival times that cannot be observed beyond that end point and 1 for those that can be observed. Another example is the traditional censoring where two sets of event-times data are generated in parallel; survival times and censoring times and the minimum of the two is picked for study, and the censoring variable gets a 1 if this minimum is from survival time and 0 when it is from censoring times (Montez-Rath et al., 2017; Wan, 2017).

The `R package simsurv` (Moriña & Navarro, 2014; Brilleman et al., 2018) was used to set up and draw the clustered survival data from the exponential distribution. Samples of size 10, 20, and 50 clusters each, having 80 and 500 subjects per cluster were generated. Each case was replicated 100 and 1000 times. This tested effect of cluster and simulation sizes on performance of the proposed method. A common approach that is used to evaluate performance of newly proposed diagnostic measure is to simulate regular data set based on the model of interest and introduce various scenarios of aberrant cases so as to check if the diagnostic statistic can detect these (Zewotir & Galpin, 2006). In that regard, two clusters in each of the three cases were perturbed so as to have different survival-times. At first, the survival times of the first two clusters were generated from a model (3.16) with perturbed random-effects parameter values as $b_{1,2} \sim N(10, 2.5^2), N(15, 5.5^2)$, while the parameters of X_1 and X_2 remained intact. This generated random effects in clusters 1 and 2 that were outside the 95% confidence range of the expected average value of zero, i.e. $[-0.784, 0.784]$, which was expected to generate survival times T in cluster 1 or 2 with some degree of outlying. That was done on assumption that values of random effects will contribute to outlying behaviour of survival times variable T in a cluster.

Secondly, the survival times in the first two clusters were generated from a model with $(\beta_1 = 1.8, 2.7)$, leaving the other parameters fixed as in model (3.16). This assessed how β_1 influenced outlying tendency of survival times in the first two clusters. Thirdly, data of the first two clusters were generated with $(\beta_2 = 2.0, 2.5)$, leaving the rest of the parameters fixed. In all other clusters than 1 and 2, data were generated using parameter values defined along with model (3.16) without any adjustment to ensure that the outlier measure should only detect cluster 1 or 2 as outlying when applied on the dataset involving all clusters.

Nonetheless, a cluster can have outlying effects on survival times T due to an interplay of values of fixed- (β_1, β_2) and random-effects $\mathbf{b}$ parameters (Zewotir & Galpin, 2006). Hence, joint perturbations of fixed- or random-effects were also performed, i.e. $\beta_1 = 1.8, 2.7$ and $b_{1,2} \sim N(10, 2.5^2), N(15, 5.5^2)$, leaving β_2 intact. Then, $\beta_2 = 2.0, 2.5$ and $b_{1,2}$
$simN(10, 2.5^2), b_{1,2}, N(15, 5.5^2)$, leaving β_1 unchanged, likewise $\beta_1 = 1.8, 2.7$ and $\beta_2 = 2.0, 2.5$, leaving random-effects b_j intact. Finally, data were generated with all the three effects perturbed, i.e. $\beta_1 = 1.8, 2.7$, $\beta_2 = 2.0, 2.5$ and $b_{1,2}N(10, 2.5^2), N(15, 5.5^2)$.

A decision about the effectiveness of the proposed method in identifying cluster 1 or 2 as outlier was made using proportion, among simulations, of correct identification of the outlying clusters 1 and 2 by the proposed outlier measure (3.13) at a given cutoff (Xiang et al., 2002). The cutoff used is $k_{1,2} > mean\,[maximum(k_i : i = 3, 4, ..., M)]$ or $k_{1,2} < mean\,[minimum(k_i : i = 3, 4, ..., M)]$ out of 100 or 1000 simulations (Xiang et al., 2002). When a newly proposed statistical method is for estimating a parameter, performance of the method is assessed using coverage probability (CP), also called Type I error, which is defined as the proportion of confidence intervals that contains the hypothetical value of the parameter in a given simulation (Kontopantelis & Reeves, 2012; Trikalinos et al., 2013; Montez-Rath et al., 2017). Where a 95% confidence interval is used, a CP close to 0.95 is desirable, and CP above 0.95 is indicative of inefficient method, while CP below 0.95 implies the new method is inaccurate (Kontopantelis & Reeves, 2012; Trikalinos et al., 2013).

In addition, power probability, also known as Type II error, for the parameter being estimated is used (Kontopantelis & Reeves, 2012). Further, bias or standardized bias is used, this is the difference between the true or simulated parameter value and its estimate, as a percentage of the estimate's standard error. Finally, the mean squared error (MSE) is also used, this is the squared difference between the true or simulated parameter value and its estimate, averaged over number of simulations. The bias and MSE close to zero are preferred for a good estimator (Trikalinos et al., 2013; Montez-Rath et al., 2017).

The following steps summarise the process used to simulate data in R:

Step 1: Set up data frame for $j = 10, 20, 50$ clusters, each with $n_j = 80$ and 500, and $n = j \times n_j$, Step 2: Sample X_1 from $Binomial(N,1,0.7)$, X_2 from $N(0,1)$, and b_j from $N(0,0.4^2)$,

Step 3: Multiply X_1, X_2 with respective coefficients $\beta_1 = 0.5$ and $\beta_2 = 1$,

Step 4: Sample survival times t^*_{ij} from Exponential(1) using model (3.16) and ‘simsurv’ package,

Step 5: Sample δ_{ij} from $Binomial(N,1,0.4)$,

Step 6: Merge and save the dataset $(X_1, X_2, \delta_{ij}, cluster_j, b_j, t_{ij})$,

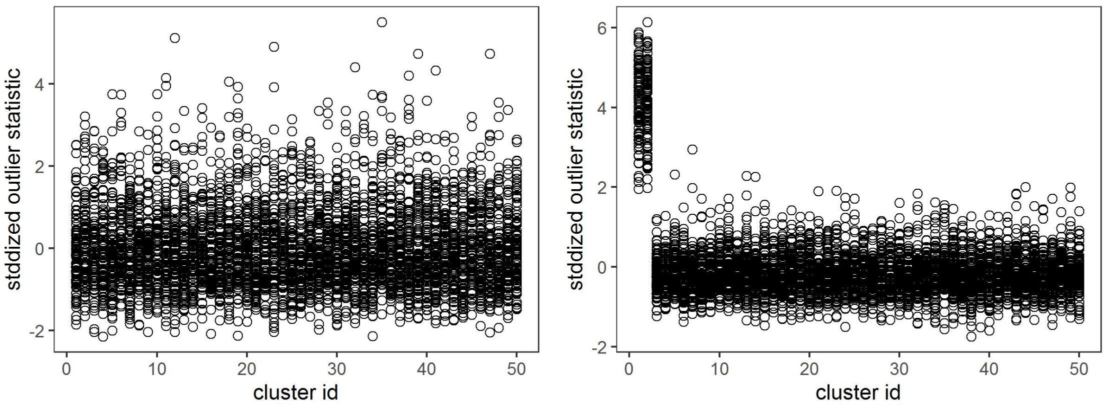
Step 7: Replicate the data in Steps 1 to 6 by 100 and 1000 simulations,

Step 8: Repeat Steps 1 to 7 with first two clusters having T generated from model with perturbed parameters as described before.

The clustered survival model (3.16) was fitted to each of the simulated dataset and the proposed outlier measure was computed for each cluster. The performance of the proposed statistic was evaluated as per criterion indicated in preceding paragraphs. The results are presented in the following section. The R codes that were used are given as appendices.

3.3.1 Simulation results when separate perturbations were done to β_1 or β_2 or b_j in first two clusters

The plots in Figure 3.1 for selected cases of the simulations indicate that the outlier statistic detected clusters 1 and 2, as outliers, as per the cutoff criterion given in previous section, when the perturbations involved fixed- and not random-effects. The plots of the statistic were out of range in the first two clusters for the case that involved β_1, and they remained consistent with the rest clusters for samples with perturbed random effects. The rest of the results on success rates of the proposed statistic are given in Tables 3.1 and 3.2.



(a) Plots of outlier statistic for a case of data with perturbed $b \sim (15, 5.5^2)$ in 2 of 50-clusters sample, each with 80 subjects and with 100 replications

(b) Plots of outlier statistic for a case of data with perturbed $\beta_1 = 2.7$ in 2 of 50-clusters sample, each with 500 subjects and with 100 replications

Figure 3.1: Plots of the proposed outlier statistic when perturbed models were used in first two clusters. Source: Researcher.

The results in Table 3.1 show that the proposed outlier statistic was effective, when the perturbations involved fixed and not random effects. When β_1 was perturbed, the residual correctly identified the affected two clusters a minimum of 0.4% and up to 100% of the simulations. Where the adjustments concerned β_2, the statistic correctly identified the two clusters at least 57% and up to 100% of the times. Adjusting random effects in the model did not cause the cluster to be outlier, the success rates of the statistic were all zero.

Further, performance of the statistic improved with cluster sample size and fixed effect size. In addition, the success rates of statistic converged to the same values between 100 and 1000 replications, for scenarios with large cluster sample size. There was a slight drop in the rates at 1000 simulations in cases of low cluster sizes. The results also show that the outlier statistic performed equally across different number of clusters per data set, holding constant the cluster sample size and fixed effect size.

Table 3.1: Percentage of times per 100 or 1000 simulations in which cluster 1 or 2 was detected as outlier by proposed statistic; a case of separate perturbations to b_j, β_1 or β_2, under 10, 20 or 50 clusters per dataset, each with 80 or 500 subjects

					100 replicates		1000 replicates	
M	n_j	β_1	β_2	$b_{1,2}$	%Cluster1	%Cluster2	%Cluster1	%Cluster2
10	80	0.5	1	$N(10,2.5^2)$	0	0	0	0
	80	0.5	1	$N(15,5.5^2)$	0	0	0	0
10	500	0.5	1	$N(10,2.5^2)$	0	0	0	0
	500	0.5	1	$N(15,5.5^2)$	0	0	0	0
20	80	0.5	1	$N(10,2.5^2)$	0	0	0	0
	80	0.5	1	$N(15,5.5^2)$	0	0	0	0
20	500	0.5	1	$N(10,2.5^2)$	0	0	0	0
	500	0.5	1	$N(15,5.5^2)$	0	0	0	0
50	80	0.5	1	$N(10,2.5^2)$	0	0	0	0
	80	0.5	1	$N(15,5.5^2)$	0	0	0	0
50	500	0.5	1	$N(10,2.5^2)$	0	0	0	0
	500	0.5	1	$N(15,5.5^2)$	0	0	0	0
10	80	1.8	1	$N(0,0.4^2)$	0	0	0	0
	80	2.7	1	$N(0,0.4^2)$	20	17	22	22
10	500	1.8	1	$N(0,0.4^2)$	50	50	25.8	24.9
	500	2.7	1	$N(0,0.4^2)$	100	100	88.9	89.6
20	80	1.8	1	$N(0,0.4^2)$	3	3	0.6	0.4
	80	2.7	1	$N(0,0.4^2)$	17	7	2.2	1.7
20	500	1.8	1	$N(0,0.4^2)$	84	83	17.4	18.9
	500	2.7	1	$N(0,0.4^2)$	96	100	95.2	95.5
50	80	1.8	1	$N(0,0.4^2)$	11	15	8.0	8.2
	80	2.7	1	$N(0,0.4^2)$	6	2	0.7	0.7
50	500	1.8	1	$N(0,0.4^2)$	57	59	31.7	34.6
	500	2.7	1	$N(0,0.4^2)$	100	98	98.5	97.5
10	80	0.5	2.0	$N(0,0.4^2)$	95	92	57	59.6
	80	0.5	2.5	$N(0,0.4^2)$	100	100	92.7	93.8
10	500	0.5	2.0	$N(0,0.4^2)$	100	100	99.9	99.6
	500	0.5	2.5	$N(0,0.4^2)$	100	100	100	100
20	80	0.5	2.0	$N(0,0.4^2)$	82	83	72.8	74.7
	80	0.5	2.5	$N(0,0.4^2)$	99	98	92.5	93
20	500	0.5	2.0	$N(0,0.4^2)$	100	100	99.7	100
	500	0.5	2.5	$N(0,0.4^2)$	100	100	99.7	99.7
50	80	0.5	2.0	$N(0,0.4^2)$	93	89	79.6	80.4
	80	0.5	2.5	$N(0,0.4^2)$	98	97	87.2	87.8
50	500	0.5	2.0	$N(0,0.4^2)$	100	100	99.1	99.1
	500	0.5	2.5	$N(0,0.4^2)$	100	100	99.9	99.8

3.3.2 Simulation results when joint perturbations were done to β_1, β_2 and b_j in first two clusters

The results in Table 3.2 show that the proposed statistic correctly detected the two outlying clusters a minimum of 39.1% and up to 100% of the times, when the joint perturbations involved β_1 and β_2. Where the joint adjustments were done to β_1, β_2 and b_j, the success rates ranged from 38.3% and up to 100% of the times. Thus the ranges of the rates were not different, with or without random effects in the joint perturbations, implying the contribution of random effects in offsetting cluster 1 or 2 was negligible.

As with cases of separate perturbations in Table 3.1, there was no tangible differences in performance of the statistic between 100 and 1000 simulations for cases with large cluster sizes. Again, the performance of the statistic improved with cluster sample size and fixed effect sizes. Once again, the proposed outlier statistic performed equally between different number of clusters per dataset, controlling for cluster sample size and fixed effect size.

Table 3.2: Percentage of times per 100 or 1000 simulations in which cluster 1 or 2 was detected as outlier by proposed statistic; a case of joint perturbations among b_j, β_1 and β_2, under 10, 20 or 50 clusters per dataset, each with 80 or 500 subjects

					100 replicates		1000 replicates	
M	n_j	β_1	β_2	$b_{1,2}$	%Cluster1	%Cluster2	%Cluster1	%Cluster2
10	80	1.8	2.0	$N(0,0.4^2)$	55	68	39.1	41.8
	80	2.7	2.5	$N(0,0.4^2)$	90	89	74.6	72.7
10	500	1.8	2.0	$N(0,0.4^2)$	100	100	99.4	99.1
	500	2.7	2.5	$N(0,0.4^2)$	100	100	100	100
20	80	1.8	2.0	$N(0,0.4^2)$	58	58	43.4	44.8
	80	2.7	2.5	$N(0,0.4^2)$	86	88	70.2	69.6
20	500	1.8	2.0	$N(0,0.4^2)$	99	99	98.9	98.6
	500	2.7	2.5	$N(0,0.4^2)$	100	100	100	100
50	80	1.8	2.0	$N(0,0.4^2)$	59	54	41.2	39.7
	80	2.7	2.5	$N(0,0.4^2)$	86	87	69	66.8
50	500	1.8	2.0	$N(0,0.4^2)$	100	100	99	98.5
	500	2.7	2.5	$N(0,0.4^2)$	100	100	97.7	98
10	80	1.8	2.0	$N(10,2.5^2)$	75	74	38.7	40.9
	80	2.7	2.5	$N(15,5.5^2)$	85	86	74.6	72.6
10	500	1.8	2.0	$N(10,2.5^2)$	100	100	99.3	99.2
	500	2.7	2.5	$N(15,5.5^2)$	100	100	100	100
20	80	1.8	2.0	$N(10,2.5^2)$	48	54	38.3	35.4
	80	2.7	2.5	$N(15,5.5^2)$	82	76	73.4	73.3
20	500	1.8	2.0	$N(1,2.5^2)$	100	100	97.9	98.4
	500	2.7	2.5	$N(15,5.5^2)$	100	100	99.5	99.3
50	80	1.8	2.0	$N(10,2.5^2)$	65	51	41.6	40.8
	80	2.7	2.5	$N(15,5.5^2)$	79	80	68.9	71.8
50	500	1.8	2.0	$N(10,2.5^2)$	100	99	96.1	95.2
	500	2.7	2.5	$N(15,5.5^2)$	100	100	98.7	99.4

3.4 Application to Malawi child survival data

The proposed outlier statistic was applied along with the standard method of visual inspection of studentized residual (Langford & Lewis, 1998) on child survival data, that were collected as part of 2015-16 Malawi Demographic and Health Survey (MDHS) data. The 2015-16 MDHS, held from 19 October 2015 to 18 February 2016, collected child survival data from women respondents and caregivers aged 15-49 years who provided birth histories. The data set is described in Section 2.6 and summarised in Table B.1. The survey employed a two-stage stratified sampling design, with emuneration areas as primary and households as secondary sampling units. Further information on the 2015-16 MDHS can be found in the survey report (Malawi National Statistical Office (NSO) & ICF, 2017), and the information about the DHS progam and data access are available at `www.DHSprogram.com`.

In order to balance between sufficient clusters and number of children per cluster, the rural and urban areas in each district were taken as separate clusters, resulting into 52 sub-districts. Child birth order and sex were used as covariates in the analysis based on previous studies (Manda, 2001). The Cox frailty model was fitted to the data and cluster outliers were assessed. The event of interest was death of a child from any cause before 60 months of age, as in Section 2.6. The event-time was age in months as at death or censoring point. The ages-at-death that were recorded as zero months were transformed into random $Uniform(0,1)$ values to reflect proportions of month-days lived by a child before death or censoring. Administrative censoring was used, and children who were still alive or had survived up to 60 months were censored. The fitted model was as follows:

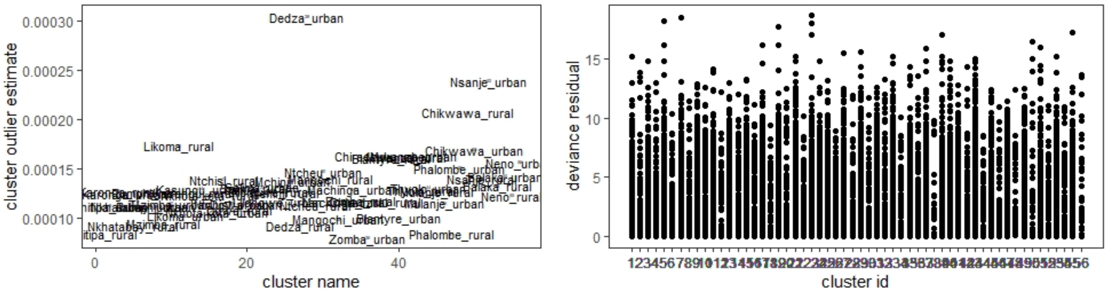
$$\begin{aligned} h_{ij}(age) = h_0(age)exp(&-0.185 \times Female - 0.214 \times Birthorder \\ &+ 0.0233 \times Birthorder_{square} + subdistrict). \end{aligned} \tag{3.17}$$

The model results showed that female children had significantly lower risk of death than the male children (p-value = 0.0096). While children with higher birth order had significantly reduced risk of death as in Section 2.6.

3.4.1 Under-Five Mortality Outlier Sub-Districts in Malawi

The computations used the national under-five mortality rate of 63 deaths per 1000 live births (Malawi National Statistical Office (NSO) & ICF, 2017) as baseline hazard. The application of the proposed statistic was analysed in comparison with the visual inspection method for standardised residuals suggested in (Langford & Lewis, 1998) to identify outlier clusters.

The results in Figure 3.2(a) indicate that the proposed statistic had detected *Dedza* urban, *Nsanje* urban, and *Chikwawa* rural as under-five mortality outlier subdistricts. This means that these clusters had poorly fitted survival times of the children compared to the other clusters. On the other hand, the visual inspection on individual deviance residuals in Figure 3.2 (b) could not conclusively determine an outlier cluster, as the plots of the deviance residuals highly overlapped across clusters.



(a) Estimates of proposed outlier statistic per cluster upon fitting the frailty Cox model on child survival data

(b) Plots of deviance residuals for children in each cluster following a frailty model on child survival data

Figure 3.2: Outlier assessment results using the proposed group outlier statistic in comparison with method of visual inspection of standardised residuals (Langford & Lewis, 1998) applied on Malawi child survival data, 2015-16 MDHS. Source: Researcher

Chapter 4

Cluster Influence for Survival Mixed Model

This chapter presents a method for assessing group influence based on the clustered semiparametric survival model. Techniques that are used in univariate survival model are revisited, before deriving the influence measure for the multivariate survival model.

4.1 Background to influence analysis for survival data

Suppose $\hat{\theta}$ is a set of maximum likelihood estimators of model parameters θ, with θ consisting of β, b_j, D, and other parameters, and let $\hat{\theta}_{(ij)}$ denotes the estimator of θ obtained from the data without i-th observation from j-th cluster. Then, the influence of i-th data record from j-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 (D. 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 (D. R. Cox, 1972).

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 $\hat{\beta(\varpi_{ij})}$ 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:

$$\begin{aligned} &(\partial U/\partial \hat{\beta})(\partial \hat{\beta}/\partial \varpi_{ij}) + \partial U/\partial \varpi_{ij} = 0 \\ &\qquad\qquad \therefore \partial \hat{\beta}/\partial \varpi_{ij} = (-\partial U/\partial \hat{\beta})^{-1} \partial U/\partial \varpi_{ij}. \end{aligned} \tag{4.1}$$

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 (4.1) 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:

$$v_{ij}(\hat{\beta}) = \int_0^\infty \left[X_{ijp}(t) - \bar{X}_p(\hat{\beta},t)\right] dm(t_{ij}), \tag{4.2}$$

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 $R(t_{lj})$ risk sets. The measure (4.2) 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 (4.2) is similar to the jackknife measure (4.1) 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:

$$\begin{aligned} \hat{\beta}_1 &= \hat{\beta}_0 + I^{-1}(\hat{\beta}_0)\dot{l}(\hat{\beta}_0) \\ \Rightarrow \hat{\beta}_1 - \hat{\beta}_0 &= I^{-1}(\hat{\beta}_0)\dot{l}(\hat{\beta}_0) \\ &= \frac{-I^{-1}(\hat{\beta}_0)\xi_{ij}}{\pi_{ij} - \xi_{ij}^T I^{-1}(\hat{\beta}_0)\xi_{ij}} m(t_{ij}) \end{aligned} \tag{4.3}$$

where $m(t_{ij})$ is the martingale residual defined along with equation (4.2), $\xi_{ijp} = \hat{H}_0(X_{ijp} - \bar{X}_p(\hat{\beta}))exp(X_{ij}^T\hat{\beta})$ 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).

As it may be appreciated, these methods are all related because they are based on rate of change of the maximum likelihood estimators, as a result of removal of one record from the data (Therneau et al., 1990). The influence measure approximation techniques are also supported in the Bayesian framework for parameter estimation for survival semiparametric model. In Bayesian set up for a survival model, the approximation of case-deletion influence measure is computed by the Kullback-Leibler divergence, denoted by $K(P, P_{(ij)})$, between the posterior distributions P of parameter θ for full data $D = \{t, \sigma, X\}$ and $P_{(ij)}$ for the data without ij-th subject, $D = \{t_{(ij)}, \sigma_{(i)}, X_{(ij)}\}$ (Cho et al., 2009; Suzuki et al., 2013). The Bayesian influence measure of a subject on posterior probability is given by:

$$K(P, P_{(ij)}) = \int p(\beta|D) log \frac{p(\beta|D)}{p(\beta|D_{(ij)})} d\beta, \tag{4.4}$$

where $p(\beta|D) \alpha L(\beta|D) f(\beta)$ and $p(\beta|D_{(ij)}) \alpha L(\beta|D_{(i)}) f(\beta)$ and with $f(\beta)$ the posterior distribution of β.

Depending on the choice of prior distribution, the computation of the influence diagnostic measure (4.4) is obtained from the product of the likelihood and prior distribution. The computation of the values of the measure is numerically done using Markov chain Monte Carlo samples from the full data posterior distribution (Cho et al., 2009; Suzuki et al., 2013).

This study considered the influence methods that result from the parametric estimation process, and not the Bayesian estimation. Further, Therneau et al. (1990) demonstrated that using the score residual, jackknife, and augmented model approaches yield similar conclusions about influence of a subject, but the score residual has a number of advantages including simplicity of interpretation. It is for this reason that this study applied the method of score residual to derive counterpart influence statistic for the clustered survival data. The extension is derived and presented in the next section.

4.2 Proposed influence statistic for multivariate survival data

Consider a case of shared frailty model for model (1.1), then the joint partial likelihood function (1.8) will be simplified to:

$$L(\beta,\sigma^2) = \prod_{j=1}^{M}\prod_{i=1}^{n_j}\left[\frac{exp(X_{lj}^T\beta + b_j)}{\sum_{s\in R(t_{lj})} exp(X_{sj}^T\beta + b_j)}\right]^{\delta_{ij}} \times \prod_{j=1}^{M}\left[(2\pi\sigma^2)^{\frac{-1}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{M} b_j^2\right)\right]. \tag{4.5}$$

The full joint partial log-likelihood function is:

$$\begin{aligned} l(\beta,\sigma^2) = &\sum_{j=1}^{M}\sum_{i=1}^{n_j}\delta_{ij}\left[(X_{ij}^T\beta + b_j) - ln\sum_{s\in R(t_{lj})} exp(X_{sj}^T\beta + b_j)\right] \\ &+ log[(2\pi\sigma^2)^{\frac{-n}{2}}] - \frac{1}{2\sigma^2}\sum_{j=1}^{M} b_j^2. \end{aligned} \tag{4.6}$$

The score functions for β and b_j follow from the log-likelihood (4.6) and are, respectively, given by:

$$U_\beta = \frac{\partial l(\beta, \sigma^2)}{\partial \beta} = \sum_{j=1}^{M}\sum_{i=1}^{n_j} \delta_{ij} \left[X_{ij} - \frac{\sum_{s\in R(t_{lj})} X_{sj} exp(X_{sj}^T\beta + b_j)}{\sum_{s\in R(t_{lj})} exp(X_{sj}^T\beta + b_j)} \right], \tag{4.7}$$

and

$$U_{\mathbf{b}} = \frac{\partial l(\beta, \sigma^2)}{\partial b_j} = \sum_{j=1}^{M}\sum_{i=1}^{n_j} \delta_{ij} \left[1 - \frac{\sum_{s\in R(t_{lj})} exp(X_{sj}^T\beta + b_j)}{\sum_{s\in R(t_{lj})} exp(X_{sj}^T\beta + b_j)} \right] - \frac{1}{\sigma^2}\sum_{i=1}^{M} b_j. \tag{4.8}$$

The estimates for β and b_j are found by solving the score functions (4.7) and (4.8) simultaneously, when they are equated to zero. The values of estimates are computed through numerical algorithms, such as Newton-Raphson method, since the equations (4.7) and (4.8) are not in closed forms (Ripatti & Palmgren, 2000). Therefore, 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 as stated before, because it requires refitting the model for each removal of a cluster.

This study therefore proposes an extension of the score residual (4.2) (Therneau et al., 1990), that results from fitting the model to data once, to study influence of clusters on fixed effects estimators from model (1.1). As for estimates of random effects $\mathbf{b}$, model (1.1) assumes that b_j 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). Hence, this study focuses on deriving group influence statistic for fixed effect estimators $\hat{\beta}$, that depend on observations from all clusters.

To analyse influence for grouped observations, the study first defines a leverage and a residual for a single unit ij at a given time t_{ij}. The score process (4.7) derived for the model (1.1) 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; Z. Zhang, 2016). For individual ij, let $r_{ij} = exp(X_{ij}^T\hat{\beta} + \hat{b}_j)$ be its risk score. Then, at the lj-th event time t_{lj}, the Schoenfeld residual (or leverage) (Schoenfeld, 1982), denoted by w_{lj}, is given by:

$$\begin{aligned} w_{li} &= X_{lj} - \frac{\sum_{s \in R(t_{lj})} r_{sj} X_{si}}{\sum_{s \in R(t_{lj})} r_{sj}}, \\ &= X_{lj} - \bar{X}(\hat{\beta}, \hat{b}_j, t_{lj}), \end{aligned} \tag{4.9}$$

where $r_{sj} = exp(X_{sj}^T\hat{\beta} + \hat{b}_j)$ is the risk score for unit ij in the risk set $R(t_{lj})$, and X_{lj} is the covariate vector of the individual experiencing the event at time t_{lj}. Further, $\hat{\beta}$ and $\hat{b}_j$ are, respectively, fixed and random effects terms estimated from the log-likelihood (4.6). 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 (4.9) is $1 \times p$ vector corresponding to each ij-th unit in the risk set.

The quantity (4.9) is also a residual proposed by Schoenfeld (1982) that sums the score processes (4.7) of units with failure time at each unique event, assuming no ties. Denote $\mathbf{W}_{lj}$ as leverages w_{lj} for all n_l data points in the risk set and p covariates, then $\mathbf{W}_{lj}$ will be $n_l \times p$ matrix. Furthermore, $w_{lj} \in [-\infty, +\infty]$, with mean $E(w_{lj}) = E(X_{lj}) - E[\bar{X}(\hat{\beta}, \hat{b}_j, t_{il})] = E(X_{lj}) - E(X_{lj}) = 0$. The value 0 of w_{lj} corresponds to observations with intermediate covariates values and are thus close to the weighted average for covariate X_{lj}, and hence their leverage on the fitted survival curve is negligible. While large negative and positive values of w_{lj} correspond to observations that have unusual covariates values, that are far from the weighted average of X_{lj}, and hence they have high leverage on the fitted survival curve (Z. 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; Z. Zhang, 2016). For survival data, one of the residuals is the martingale, defined along equation (4.2), 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 (1.1), we obtain an $n_l \times 1$ stacked vector of residuals for units in the risk set $R(t_{lj})$ given by:

$$
\begin{aligned}
&m(t_{lj}) = N(t_{lj}) - \hat{H}_0(t)exp(X_{lj}^T\hat{\beta} + \hat{b}_j)\\
\Rightarrow &\begin{bmatrix} m(t_{11}) \\ \vdots \\ m(t_{n_11}) \\ m(t_{12}) \\ \vdots \\ m(t_{n_22}) \\ \vdots \\ m(t_{1M}) \\ \vdots \\ m(t_{n_MM}) \end{bmatrix} = \begin{bmatrix} N(t_{11}) - \hat{H}_0(t)exp(X_{11}^T\hat{\beta} + \hat{b}_1) \\ \vdots \\ N(t_{n_11}) - \hat{H}_0(t)exp(X_{n_11}^T\hat{\beta} + \hat{b}_1) \\ N(t_{12}) - \hat{H}_0(t)exp(X_{12}^T\hat{\beta} + \hat{b}_2) \\ \vdots \\ N(t_{n_22}) - \hat{H}_0(t)exp(X_{n_22}^T\hat{\beta} + \hat{b}_2) \\ \vdots \\ N(t_{1M}) - \hat{H}_0(t)exp(X_{1M}^T\hat{\beta} + \hat{b}_M) \\ \vdots \\ N(t_{n_MM}) - \hat{H}_0(t)exp(X_{n_MM}^T\hat{\beta} + \hat{b}_M) \end{bmatrix},
\end{aligned}
\tag{4.10}
$$

where $\hat{H}_0(t) = \int_{-\infty}^{t} h_0(s)ds$ is the estimated cumulative baseline hazard. The residual (4.10) has values in the range $(-\infty, 1]$, because $N(t_{il})$ is either 0 or 1 and $\hat{H}_0(t)exp(X_{lj}^T\hat{\beta} + \hat{b}_j)$ has values in the interval $[0, \infty)$. In addition, $E(m(t_{lj})) = E(N(t_{lj})) - E(\hat{\Lambda}_0(t)exp(X_{lj}^T\hat{\beta} + \hat{b}_j)) = E(N(t_{lj})) - E(N(t_{lj})) = 0$, since the off-minus quantity in (4.10) is the average number of events.

Both leverage quantity (4.9) and residual (4.10) 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 (D. 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 (4.10) and leverage in (4.9) for model (1.1), we propose an analogue of the score residual (4.2) (Therneau et al., 1990) to measure influence of a cluster on $\hat{\beta}$ for the model (1.1) as a vector product of values of vector (4.10) and those of columns of matrix (4.9) for subjects under risk set $R(t_{il})$ in the same cluster i, given by:

$$v_j(\hat{\beta}) = [m(t_{lj})]^T \times \mathbf{W}_{lj}. \tag{4.11}$$

The extended score residual (4.11) 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 (4.11) 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 $\mathbf{W}_{lj}$ in (4.11) has elements $w_{lj} \in [-\infty, +\infty]$ and $m(t_{lj}) \in (-\infty, 1]$, both with mean 0, then the proposed influence statistic (4.11) is expected to have mean 0.

Large positive value of the proposed statistic (4.11) means a cluster has majority of subjects that have high positive values in w_{lj} that coincide with high positive values in $m(t_{lj})$, or large negative values in w_{lj} coinciding with large negative values in $m(t_{lj})$. 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 (4.11) implies that a cluster has majority of subjects that have large positive leverage w_{lj} that coincide with large negative values of the residual $m(t_{lj})$ 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.

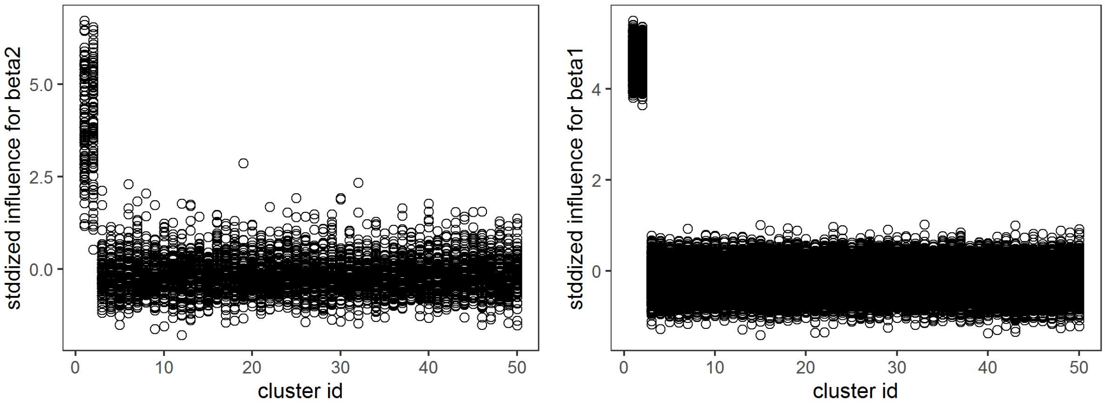
The values of (4.11) that are close to zero imply most subjects of the corresponding clusters have either leverage close to zero or residual close to zero, hence such clusters have no issues for follow up investigation. To decide on influential groups, some studies in linear mixed-effects models have 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 chapter to examine influential clusters to the fixed-effects estimates in the fitted semiparametric survival mixed models using the proposed influence statistic (4.11).

Chapter 5

Simulation Results and Application of the Influence Statistic

The data generated from a simulation study described in Section 3.3 were utilised to evaluate performance of the proposed influence statistic developed in Chapter 4. The two clusters in which the generated survival data from model (3.16) involved perturbed β_1 and β_2 were subjected to examination to observe whether they would be identified by the proposed influence statistic. The same assessment criterion described in Section 3.3 was used, that is, through percentage of simulations for which the proposed influence statistic correctly identified the two target clusters as having influence on β_1 or β_2 using the cutoff given in Section 3.3. Upon fitting the model (3.16) to the simulated data, the proposed influence statistic was computed and its performance evaluated.

An inspection of performance of the statistic displayed in Figure 5.1 indicates that the residual detected influence of the first two clusters on $\hat{\beta}_1$ and $\hat{\beta}_2$. The values of the statistic were outstandingly higher in the first two clusters than in the other clusters. This study therefore assessed success rates of the proposed influence statistic under each simulation scenario using the cutoff presented in Section 3.3.



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

5.1 Simulation results for influence of cluster 1 or 2 on $\hat{\beta}_1$

Table 5.1 shows success rates of the proposed influence statistic in detecting impact of cluster 1 or 2 on $\hat{\beta}_1$. 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 performance of the proposed influence residual improved with increasing 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 increasing fixed effect size, 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 5.1: 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 (3.16) had $\beta_1 = 0.5$, $\beta_2 = 1$.

5.2 Simulation results for influence of cluster 1 or 2 on $\hat{\beta}_2$

The results in Table 5.2 are for success rates of 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 on influence of cluster 1 or 2 on $\hat{\beta}_2$ improved with increasing cluster sample size, as the rates were consitently higher for cluster sizes of 500 and lower with cluster sizes of 80 subjects. Again, the performance of the statistic improved with increasing fixed effect size, a situation that was also noticeable in low cluster sizes like before. Likewise, there was no difference in performance of the proposed influence 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 well in different number of clusters per sample.

Table 5.2: 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 (3.16) had $\beta_1 = 0.5$, $\beta_2 = 1$.

5.3 Application to Malawi child survival data

Using the 2015-16 MDHS child survival data described in Section 3.4, the proposed influence statistic was computed. The fitted frailty model in equation (3.17) was used:

$$\begin{aligned} h_{ij}(age) = h_0(age)exp(&-0.185 \times Female - 0.214 \times Birthorder \\ &+ 0.023 \times Birthorder_{squared} + subdistrict). \end{aligned} \tag{5.1}$$

The study analysed influence of each cluster on effect of being female on child mortality for better comparison with findings from Section 2.6. The national under-five mortality rate of 63 deaths per 1000 live births (Malawi National Statistical Office (NSO) & ICF, 2017) was used as baseline hazard rate. Upon identifying the influential clusters to the model, their impact on fixed regression parameter estimates was analysed through re-fitting the model to data without the detected clusters and observe the changes in the parameter estimates for effect of being female on child mortality.

5.3.1 Results for influential clusters on effect of being female on child mortality

The results in Figure 5.2 show that 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 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. While *Phalombe* urban, *Karonga* rural and *Salima* urban clusters were identified as having negative borderline influence on effect of being female on child mortality. This implies that the three clusters had majority of children with high leverage on estimated mortality who had also survived longer than predicted by the model, such that removing these three clusters from analysis would impact on estimated effect of female gender on child mortality.

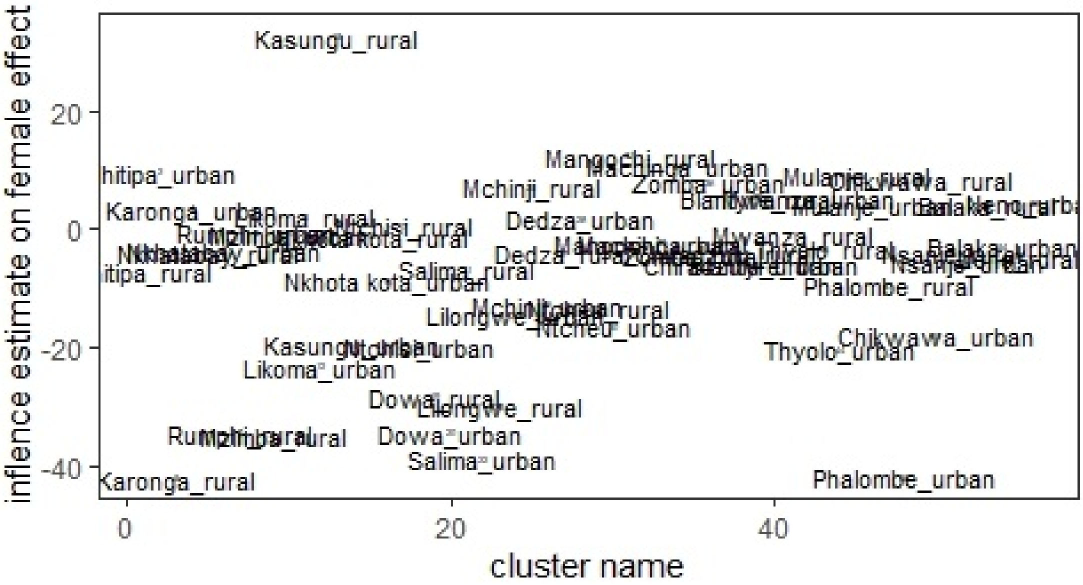


Figure 5.2: Sub-district level estimates of the proposed influence statistic for effect of female gender upon fitting a frailty Cox hazard regression model to Malawi child survival data, 2015-16 MDHS. Source: Researcher

5.3.2 Impact of the identified influential clusters on model estimate for effect of female gender on mortality

Table 5.3 shows results of model estimates using full Malawi child survival dataset and also the data without two of the identified influential clusters; *Kasungu* rural and *Salima* urban. The findings indicate that removal of *Kasungu* rural cluster from analysis resulted in further reduction in logarithm hazard of death for female children by 0.0163. Thus, the survival model was better off without data from *Kasungu* rural cluster. This was also noticed with the reduction in p-value by 0.0042. While dropping *Salima* urban cluster increased the hazard of death in female children by 0.0015. Thus, the data from *Salima* urban cluster were required in the model. Again, this is reflected in the p-value that got higher upon removing this cluster.

Removing both clusters from analysis resulted in reduction in logarithm of hazard of death in female children, but not as much 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, while *Salima* urban negative influence on estimate of effect of female gender on child mortality. The standard errors of the parameter estimates slightly increased in each case, implying that the original model parameter 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 the proposed statistic in the previous section. It is shown in Table 5.3 that impact of *Kasungu* rural cluster on the estimate of effect of female gender on mortalirt was so huge compared to that of *Salima* urban cluster.

Table 5.3: 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)
$\hat{\beta}$	-0.1848	-0.2011 (0.0163)	-0.1833 (-0.0015)	-0.1996 (0.0148)
$se(\hat{\beta})$	0.0713	0.0722 (-0.0009)	0.0715 (-0.0002)	0.0723 (-0.0010)
p-value	0.0096	0.0054 (0.0042)	0.0100 (-0.0004)	0.0058 (0.0038)
$var(re)$	0.0419	0.0399 (0.0020)	0.0418 (0.0001)	0.0397 (0.0022)

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.

Chapter 6

Discussion and Conclusions

This chapter discusses the findings from the evaluation of the proposed outlier and influence statistics for multivariate survival data model. It also discusses the findings and implications arising from the analysis of clustered child survival data in Malawi based on a nationally representative health survey. In addition, the main biostatistical contribution of this PhD work in the broader topics of outlier and influence statistics for multivariate models has been discussed.

6.1 Discussion of findings

This study set out to develop statistics for detecting outlying and influential groups of data points in a multivariate survival data model. The group outlier and influence statistics have been derived. The proposed outlier statistic extends methods that are developed for the linear mixed-effects model. While the proposed influence statistic extends the score residual that is developed for the univariate survival model. In each case, the proposed statistics utilise the model postestimation quantities that have correlated values within clusters, but uncorrelated across clusters. The proposed statistics have proved to be very effective in identifying outlying and influential groups of observations in the analysis of multivariate survival data. For example, when a fixed effect coefficient was perturbed in one cluster, the proposed outlier statistic correctly identified the affected cluster 99.8% of the time and the influence statistic 100% of the time.

The findings have shown that performance of both the proposed outlier and influence statistics improves with increasing cluster sample size. This results from lowered uncertainty in repeated sampling that any statistic gains from large sample size (Hemez et al., 2010). So, it is likely for an outlying or influential cluster to be detected as such using the proposed statistics, when cluster sample size is large enough compared to small cluster sizes. This is also the reason why the success rates of the statistics were observed to be stable in large simulation sizes compared to small simulation sizes, for cases of large cluster sample sizes, compared to small cluster sizes. With large cluster sizes, the performance of both outlier and influence statistics was not affected by the number of repetition of sampling. This means that the proposed statistics fulfill the property of robustness required for any statistical tool (Hemez et al., 2010). The success rates of a 'good' residual must converge to the same range of values in repeated experimentations. Furthermore, both proposed statistics were effective regardless of the number of clusters per dataset.

Evaluation of the proposed statistics has not supported a definitive cutoff for their application. Thus, relative comparisons of values of the statistics across clusters suffice to examine outlying or influential clusters of observations to the mixed survival model (Zewotir & Galpin, 2006). With linear mixed-effects models, Nieuwenhuis et al. (2012) suggest using a cutoff of $2/\sqrt{M}$ to assess influential groups, but this cannot be suggested as a standard for clustered survival data, as demonstrated in this study. A search for proper cutoff for examining outlying and influential clusters in mixed-effects models is still being debated in literature. Since, each fitted model to data may not be entirely correct for every dataset, outlier and influence residual cutoffs have to be applied in a flexible manner using relative comparisons of clusters in a particular dataset as observed by Zewotir &

Galpin (2006).

With the frailty survival model that was used in this study, in which observations in one cluster share a random effect that is additive to fixed covariates in a model and hence in the extended martingale residual, the outlying tendency of a cluster based on the proposed outlier statistic is largely influenced by the fixed covariates values of subjects in the model and not values of random effects. Previously, it has been viewed that the random effects part of a hierarchical model may have contribution to making clusters outliers (Langford & Lewis, 1998). But simulation studies conducted as part of this study have shown contrary findings. Since estimates of random-effects are considered as best linear unbiased predictors (BLUPs) of the random effects, they may serve to assess correct specification of the random effects part of the mixed model Schabenberger (2005); Zewotir & Galpin (2007); Loy & Hofmann (2014).

It might be necessary to identify individual outlying and influential subjects in the identified outlying and influential clusters to understand their contribution in making the clusters as such (Langford & Lewis, 1998; Zewotir & Galpin, 2006). Ordinarily, Xiang et al. (2002) observed that group-level diagnostics are well applicable to individual-level data through assessing observations nested within each cluster. This study did not perform subject-level outlier or influence analyses, as the assumed dependence of observations within a cluster paused a challenge for such analyses. Likewise, a score residual has potential to diagnose the proportional hazard assumption in a survival model (Therneau et al., 1990); this too was not explored for the proposed influence statistic for clustered survival data.

Perturbing regression parameters to introduce unusual observations in the data has been a standard practice for evaluating newly introduced diagnostic statistics (Xiang et al., 2002; Zewotir & Galpin, 2006; Kontopantelis & Reeves, 2012; Montez-

Rath et al., 2017). However, for clustered survival data, deciding a threshold for which a parameter value can cause the survival times in a cluster to be outlying or the cluster to be influential on regression coefficients would be purely guess-work of the analyst. The absence of literature on such reference values contributed to uncertainty in deciding effect sizes for perturbed parameter values during simulations in this study. As an alternative, some studies have used direct mechanical imputations of unusual values for the response variable for few target subjects or groups of observations in the already-generated dataset. This is also done on covariates data by using a different probability distribution to generate covariate values of the target subjects (Zewotir & Galpin, 2006; Cho et al., 2009).

When applied to child survival data from Malawi with 56 clusters, the proposed outlier statistic identified two urban clusters: *Dedza* urban and *Nsanje* urban and one rural cluster: *Chikwawa* rural as outliers to child survival in Malawi, based on a mixed survival model that had covariates: sex and birth order of a child. This meant the three clusters had majority of children that were poorly fitted by the model. The three districts are located off the major cities in the country or they are rural-based. Although, this study did not estimate predicted survival probabilities of children in each subdistrict, it might be that children in the detected outlying subdistricts are at high risk of death since rural settings are subjected to low access to health services due to long distance to clinics (Ustrup et al., 2014).

The results from applying the influence statistic on under-five mortality data identified four clusters that had influence on the model, one with outright positive influence and the other three with borderline negative influence. Upon investigating the clusters, it was confirmed that deleting one with outright positive influence caused a huge change on regression estimates, while the other with borderline influence impacted a small change. This confirmed relevance and usability of the proposed influence statistic for the multivariate survival model. Given survival data with large number of clusters, performing the exact delete-one analysis becomes computationally demanding and tedious for assessment of cluster influence, as it involves fitting and re-fitting the model to the data for each removed cluster. Therefore, the proposed influence statistic becomes practical and time-effective method for the group influence examination for the clustered survival data model, as it results from fitting the model once.

Furthermore, none of the influential subdistricts detected through the residual 'averaging' method (Jennings, 1986; Duchateau & Janssen, 2005; Legrand et al., 2006) for univariate survival model reported in Section 2.6 were commonly identified by the application of the proposed influence statistic in Section 5.3 in this study. Similarly, none of the outlier clusters reported by the random effects residual in Section 2.6 were commonly identified by the application of the proposed outlier statistic in Section 3.4 in this study. But *Chikwawa* rural subdistrict was detected as an outlier by both the averaging method of unrivariate survival model residual in Section 2.6 and the proposed outlier statistic in Section 3.4. The advantage of using mixed survival model when the data have apparent clustering is in ensuring the unbiased estimates of regression coefficients (Liang & Zeger, 1993), which can in turn put value to the proposed diagnostic statistics in this study, that are based on fitting a multivariate survival model to clustered survival data.

Overall, the application of this study to child survival data from Malawi showed that female children were associated with lower risk of mortality in the first five years of age compared to their male counterparts. It was not the intention of this study to discuss reasons for this effect beyond evaluating the applicability of the derived outlier and influence statistics on the data. However, studies have attributed the 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).

6.2 Limitations and challenges of the study

The simulation and application studies done in this PhD work are subject to some limitations. This section highlights some of these challenges.

The proposed outlier and influence statistics apply to a clustered survival model with time-independent covariates and multivariate normal random effects for analytical convenience. The derivation of the outlier and influence statistics could not have been straightforward had the model used assumed time-dependent covariates and non-normally distributed random effects, although parameter estimation methods exist for multivariate survival model with time-dependent covariates and non-normally distributed random effects (Manda, 2011).

The evaluation of the proposed outlier and influence statistics would have been enriched had this study accessed a variety of multivariate survival datasets, for example, recurrent events data from some longitudinal study. The application data used had 95% censoring rate, which might have affected the regression parameter estimates obtained from fitting the multivariate survival model to the data. The high censoring rate in survival data causes biased estimates in Cox survival models, especially in low sample-sized data (Lin et al., 2013).

The derived outlier and influence statistics are based on the conditional multivariate survival model, in which estimates of fixed- and random-effects model parameters are simultaneously solved. The techniques may not apply to marginal multivariate survival model, in which the estimators for fixed- and random-effects are solved separately (Liang & Zeger, 1993), although some studies have worked on diagnostics for marginal models (Russo et al., 2009).

As earlier alluded to, the dependence of observations within clusters for clustered survival data hampered efforts to think of follow-up outlier and influence statistics for individual observations within the identified outlying and influential clusters. The derived methods for cluster level outlier and influence analysis in this study as well as the existing methods for individual level survival data diagnostic analyses (Therneau et al., 1990) are based on assumption of independence of groups and individual observations, respectively.

6.3 Directions for future research

As highlighted in the previous sections, the model outlier and influence statistics for clustered survival data are not a widely-studied field. The following areas are recommended for future work:

- The methods studied in this work are for clustered survival model with time-constant covariates. However, various formulations of the survival model exist, for example stratified and time-dependent survival mixed models. Each choice of the specification has implications on parameter estimation, and hence on derivations for model diagnostic statistics. Future work could develop outlier and influence statistics for stratified or time-dependent multivariate survival model.

- The proposed outlier and influence statistics are post-estimation functions of model parameter estimators, in which the estimation was done using penalised joint partial likelihood method (Ripatti & Palmgren, 2000) supported by the Newton-Raphson maximisation. The model could have been estimated using marginal likelihood construction supported by the EM algorithm as in Manda (2011), all within a frequentist estimation paradigm. Alternatively, Bayesian estimation could also have been used (Manda & Meyer, 2005; Cho et al., 2009; Suzuki et al., 2013). In addition, normal random effects were assumed for the frailty effect in the multivariate survival model. Thus one could look at either having a marginal or Bayesian construction for the model or using a different assumption for random effect distribution. It could be worthwhile to derive similar group outlier and influence statistics based on different ways of estimating the multivariate survival model.

- The use of graphical methods to display results of the proposed outlier and influence statistics could not ascertain the degree of outlying or influence of a cluster to the multivariate survival model. The formal diagnostic hypothesis tests about the outlying or influential clusters, using the proposed statistics, have not been developed. This could be an area for future research.

- The proposed influence statistic has been derived for regression parameters. However, influence statistics cover assessing impact of a subject on likelihood estimate, fitted values, and other model inferences. Other than assessing the impact of clusters on covariates via the regression parameters, one could assess the impact on quantities such as the likelihood function and fitted values.

6.4 Contribution of the thesis to statistics field

From the outset, it has been stated that while diagnostic statistics are well known in linear and linear mixed models, there is a paucity of equivalent statistics for multivariate survival data models. This study was set out in this context, where appropriate diagnostic statistics for multivariate survival data models have been derived. Specifically;

- The study contributes to research on model diagnostic statistics for the clustered survival data, by developing the statistics for assessing outlier and influential groups of observations in multivariate survival model. The outlier statistic derived in this study is capable of showing a cluster that has measurements that are not consistent with the rest of the data in the other clusters used in the multivariate survival model. Similarly, the proposed influence statistic has the ability of showing impact of a cluster on the estimate of regression coefficient when the concerned cluster is dropped from the modelling. Both statistics will be applied upon fitting a model to clustered survival data once, hence they are efficient. This contribution will inform further critique of the knowledge by future researchers in the area of outlier and influence statistics for clustered survival model. The proposed statistics will also enhance the analysis of the clustered survival data by the users of statistics.

- By adapting some limited theory used in linear, linear mixed-effects, and univariate survival models, this study has shown that a researcher can innovate some statistics in an area that has less developed statistics. Both the outlier and influence statistics proposed in this study adapted methods that are available for linear, linear mixed-effects, and univariate survival models using appropriate mathematical principles. The extensions add to the efforts made by previous researchers in diagnostic statistics, which will in turn help future researchers to develop the knowledge further.

6.5 Concluding remarks

Multivariate survival data are commonly encountered in many disciplines, including biomedical studies. Statistical software packages are now available to fit a survival model to such data. However, due to lack of or limited statistics to use in assessing outlier and influential data points or groups of data points as it is mostly done in linear models, such an undertaking is seldom done when the analysis of multivariate survival data is carried out. It was in this context that this PhD study was set to fill the gap. The outlier and influence statistics for the analysis of multivariate survival data have been derived. Both statistics have been developed from adapting and combining similar statistics for univariate survival data and those derived in linear mixed-effect models. The derived statistics were able to correctly identify outlying and influential clusters based on simulation studies.

It is recommended that when an analysis of multivariate survival data is done, it should be accompanied by an assessment of unusual clusters to avoid having biased and spurious findings.

References

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.

Andersen, E. (1992). Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 781–791.

Andrews, D., & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 85–93.

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

Bell, J. F., & Malacova, E. (2004). Outliers and multilevel models. In *Sixth International Conference on Social Science Methodology: Recent Developments and Applications in Social Research Methodology, Amsterdam, the Netherlands.*

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.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723.

Bienias, J. L., Hall, C. B., & Bang, W. (2002). Diagnostics for random coefficient mixed models with unweighted and weighted data. In *Proceedings of the Annual Meeting of the American Statistical Association, Biometrics Section [CD-ROM]. Alexandria, VA: American Statistical Association.*

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.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, *105*(489), 147–156.

Chiou, S., Kang, S., & Yan, J. (2014). Fitting accelerated failure time models in routine survival analysis with r package aftgee. *Journal of Statistical Software*, *61*(11), 1–23.

Cho, H., Ibrahim, J. G., Sinha, D., & Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, *65*(1), 116–124.

Christensen, R., Pearson, L., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, *34*(1), 38–45.

Claeskens, G., & Hart, J. D. (2009). Goodness-of-fit tests in mixed models. *Test*, *18*(2), 213–239.

Cleves, M., Gutierrez, R. G., Gould, W., & Marchenko, Y. V. (2010). *An introduction to survival analysis using stata, 3rd ed.* Texas: Stata Press.

Cook, A. (2008). Cox proportional hazards model. *Lecture notes: National University of Singapore.*

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

Cook, D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Cook, D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman and Hall.

Cox, D., & Snell, J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Crowther, M. (2017). Multilevel mixed effects parametric survival analysis. *arXiv preprint arXiv:1709.06633*.

Crowther, M. J., & Lambert, P. C. (2012). Simulating complex survival data. *The Stata Journal*, *12*(4), 674–687.

Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, *32*(23), 4118–4134.

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.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models.* Boca Raton: CRC press.

Donner, A., & Klar, N. (1994). Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *American Journal of Epidemiology*, *140*(3), 279–289.

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

Duchateau, L., & Janssen, P. (2005). Understanding heterogeneity in generalized mixed and frailty models. *The American Statistician*, *59*(2), 143–146.

Fitrianto, A., & Jiin, R. (2013). Several types of residuals in cox regression model: an empirical study. *International Journal of Mathematical Analysis*, *7*(53), 2645–2654.

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

Fung, W., Zhu, Z.-Y., Wei, B.-C., & He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 565–579.

Galbraith, S., Daniel, J., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *Journal of Neuroscience*, *30*(32), 10601–10608.

Gharibvand, L., & Liu, L. (2009). Analysis of survival data with clustered events. In *SAS Global Forum* (Vol. 2009, pp. 1–11).

Glidden, D., & Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, *23*(3), 369–388.

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

Grambsch, P., & Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

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

Ha, I., Sylvester, R., Legrand, C., & MacKenzie, G. (2011). Frailty modelling for survival data from multi-centre clinical trials. *Statistics in Medicine*, *30*(17), 2144–2159.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, *33*(2), 1–22.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Hemez, F., Atamturktur, H. S., & Unal, C. (2010). Defining predictive maturity for validated numerical simulations. *Computers & Structures*, *88*(7-8), 497–505.

Hosmer Jr, D. W., Lemeshow, S., & May, S. (2011). *Applied survival analysis: regression modeling of time-to-event data* (Vol. 618). New York: John Wiley & Sons.

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564.

Jennings, D. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association*, *81*(396), 987–990.

Kontopantelis, E., & Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*, *21*(4), 409–426.

Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S., & Rondeau, V. (2017). Tutorial in joint modeling and prediction: a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *arXiv preprint arXiv:1701.03675*.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

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

Lee, S., & Xu, L. (2004). Influence analyses of nonlinear mixed-effects models. *Computational Statistics & Data Analysis*, *45*(2), 321–341.

Legrand, C., Duchateau, L., Sylvester, R., Janssen, P., van der Hage, J. A., Van de Velde, C., & Therasse, P. (2006). Heterogeneity in disease free survival between centers: lessons learned from an eortc breast cancer trial. *Clinical Trials*, *3*(1), 10–18.

Leucuţa, D., & Cadariu, A. A. (2008). Statistical graphical user interface plug-in for survival analysis in r statistical and graphics language and environment. *Applied Medical Informatics*, *23*(3, 4), 57–62.

Liang, K., Self, S., Bandeen-Roche, K., & Zeger, S. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis*, *1*(4), 403–415.

Liang, K., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, *14*(1), 43–68.

Lin, N. X., Logan, S., & Henley, W. E. (2013). Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics*, *69*(4), 850–860.

Louis, T. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

Loy, A., & Hofmann, H. (2014). Hlmdiag: A suite of diagnostics for hierarchical linear models in r. *Journal of Statistical Software*, *56*(5), 1–28.

Maia, R., Madsen, P., & Labouriau, R. (2014). Multivariate survival mixed models for genetic analysis of longevity traits. *Journal of Applied Statistics*, *41*(6), 1286–1306.

Makaula, P., Funsanani, M., Mamba, K. C., Musaya, J., & Bloch, P. (2019). Strengthening primary health care at district-level in malawi-determining the coverage, costs and benefits of community-directed interventions. *BMC Health Services Research*, *19*(1), 509.

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.

McGilchrist, C. (1993). Reml estimation for survival models with frailty. *Biometrics*, *49*(1), 221–225.

McGilchrist, C., & Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, *47*(2), 461–466.

Mehrotra, D., Su, S.-C., & Li, X. (2012). An efficient alternative to the stratified cox model analysis. *Statistics in Medicine*, *31*(17), 1849–1856.

Montez-Rath, M., Kapphahn, K., Mathur, M., Mitani, A., Hendry, D., & Manisha, D. (2017). Guidelines for generating right-censored outcomes from a cox model extended to accommodate time-varying covariates. *Journal of Modern Applied Statistical Methods*, *16*(1), 1538–9472.

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.

Munda, M., Rotolo, F., & Legrand, C. (2012). Parfm: parametric frailty models in r. *Journal of Statistical Software*, *51*(11), 1–20.

Nguyen, D., & Rocke, D. (2002). Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, *18*(12), 1625–1632.

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.

Nyangoma, S., Fung, W., & Jansen, R. (2006). Identifying influential multinomial observations by perturbation. *Computational Statistics & Data Analysis*, *50*(10), 2799–2821.

Palmgren, J., & Ripatti, S. (2002). Fitting exponential family mixed models. *Statistical Modelling*, *2*(1), 23–38.

Pan, J., Fei, Y., & Foster, P. (2014). Case-deletion diagnostics for linear mixed models. *Technometrics*, *56*(3), 269–281.

Pan, Z., & Lin, D. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, *61*(4), 1000–1009.

Peña, D. (2005). A new statistic for influence in linear regression. *Technometrics*, *47*(1), 1–12.

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.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*(4), 705–724.

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.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73–79.

Russo, C. M., Paula, G. A., & Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics & Data Analysis*, *53*(12), 4143–4156.

Samuels, S. J. (1978). *Survival analysis from the viewpoint of hampel's theory for robust estimation.* North Carolina: University of North Carolina at Chapel Hill.

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.

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI 29: Statistics and Data Analysis, Paper 189* (Vol. 29).

Schall, R., & Dunne, T. (1988). A unified approach to outliers in the general linear model. *Sankhyā: The Indian Journal of Statistics, Series B*, *50*(2), 157–167.

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

Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 659–687.

Song, P., Fan, Y., & Kalbfleisch, J. (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, *100*(472), 1145–1158.

Song, P., Zhang, P., & Qu, A. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica*, *17*, 929–943.

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.

Suzuki, A. K., Louzada, F., & Cancho, V. G. (2013). On estimation and influence diagnostics for a bivariate promotion lifetime model based on the fgm copula: A fully bayesian computation. *TEMA (São Carlos)*, *14*(3), 441–461.

Tang, N.-S., Wei, B.-C., & Wang, X.-R. (2000). Influence diagnostics in nonlinear reproductive dispersion models. *Statistics & Probability Letters*, *46*(1), 59–68.

Therneau, T. (2015). coxme: mixed effects cox models. r package version 2.2-3. *URL: http://CRAN. R-project. org/package= coxme.*

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

Thomas, L., & Reyes, E. (2014). Tutorial: survival estimation for cox regression models with time-varying coefficients using sas and r. *Journal of Statistical Software*, *61*(c1), 1–23.

Trikalinos, T. A., Hoaglin, D. C., & Schmid, C. H. (2013). Empirical and simulation-based comparison of univariate and multivariate meta-analysis for binary outcomes: retrieved from https://www.ncbi.nlm.nih.gov/books/nbk132562/.

Turkan, S., & Toktamis, O. (2012). Influence analysis in the linear mixed model. *Pakistan Journal of Statistics*, *28*(3), 341–349.

Türkan, S., & Toktamis, Ö. (2013). Detection of influential observations in semiparametric regression model. *Revista Colombiana de Estadística*, *36*(2), 271–284.

Ustrup, M., Ngwira, B., Stockman, L., Deming, M., Nyasulu, P., Bowie, C., … Bresee, J. (2014). Potential barriers to healthcare in malawi for under-five children with cough and fever: a national household survey. *Journal of Health, Population, and Nutrition*, *32*(1), 68.

Vaida, F., & Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, *19*(24), 3309–3324.

Van Kempen, G., & Van Vliet, L. (2000). Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry: The Journal of the International Society for Analytical Cytology*, *39*(4), 300–305.

Vaupel, J., Manton, K., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

Wan, F. (2017). Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in Medicine*, *36*(5), 838–854.

Wei, W., & Su, J. (1999). Model choice and influential cases for survival studies. *Biometrics*, *55*(4), 1295–1299.

Wilson, M. (2013). Assessing model adequacy in proportional hazards regression. In *Statistics and Data Analysis, SAS Global Forum, Paper 431.*

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. (2004). Proportional hazards mixed models: a review with applications to twin models. *Metodoloski Zvezki*, *1*(1), 205–212.

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.

Xue, Y., & Schifano, E. D. (2017). Diagnostics for the cox model. *Communications for Statistical Applications and Methods*, *24*(6), 583–604.

Yang, H. (2012). Visual assessment of residual plots in multiple linear regression: A model-based simulation perspective. *Multiple Linear Regression Viewpoints*, *38*(2), 24–37.

Yau, K. (2001). Multilevel models for survival analysis with random effects. *Biometrics*, *57*(1), 96–102.

Zare, K., & Rasekh, A. (2011). Diagnostic measures for linear mixed measurement error models. *SORT-Statistics and Operations Research Transactions*, *35*(2), 125–144.

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, D., Sun, J. L., & Pieper, K. (2016). Bivariate mixed effects analysis of clustered data with large cluster sizes. *Statistics in Biosciences*, *8*(2), 220–233.

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

Zhao, Y., Lee, A., Yau, K., & McLachlan, G. (2011). Assessing the adequacy of weibull survival models: a simulated envelope approach. *Journal of Applied Statistics*, *38*(10), 2089–2097.

Zhu, H., Lee, S.-Y., Wei, B.-C., & Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, *88*(3), 727–737.

Ziegler, A., Kastner, C., & Blettner, M. (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal*, *40*(2), 115–139.

Appendix A: Map of the 28 districts in Malawi, 2020

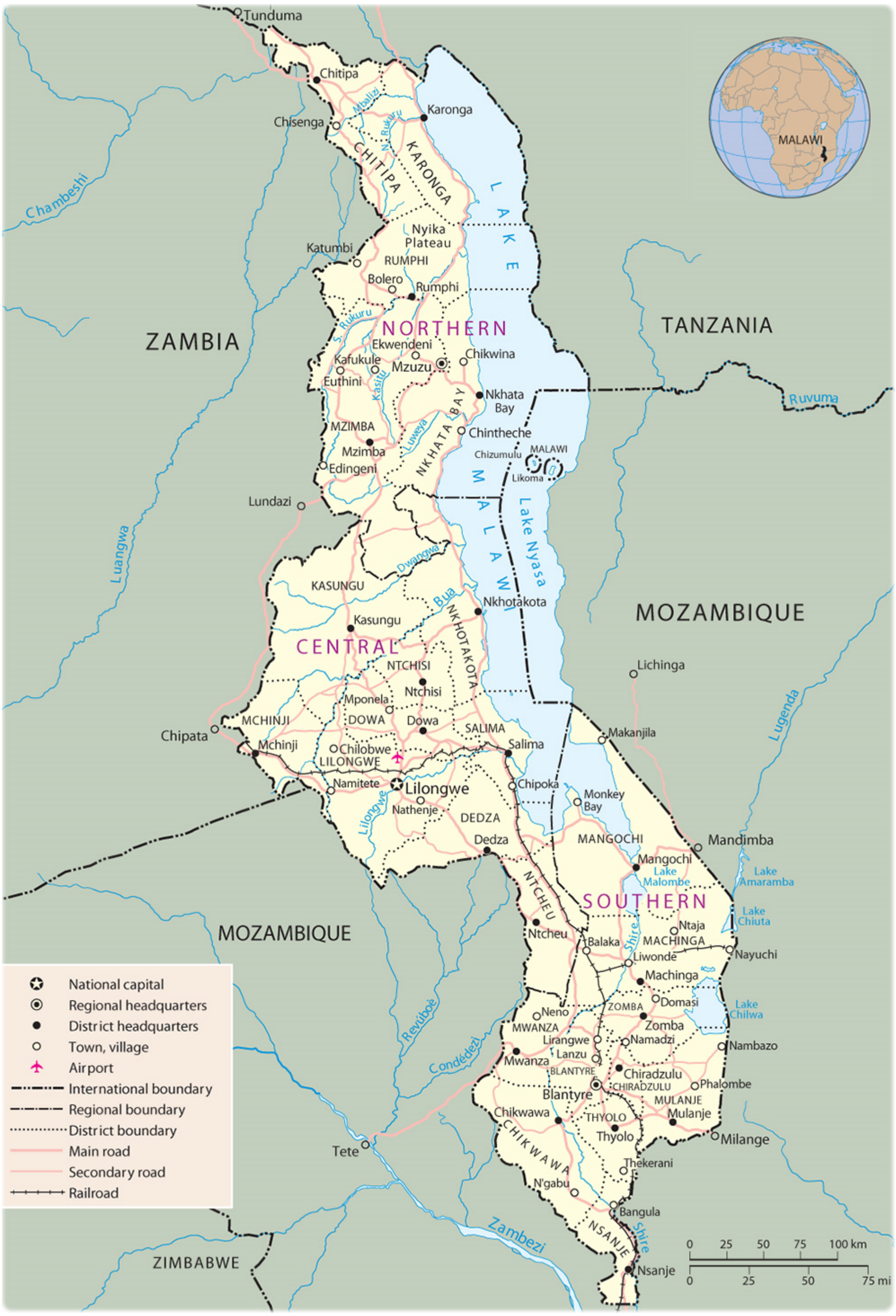


Figure A.1: Map of Malawi districts. Source: www.mw.one.un.org

Appendix B: Child characteristics per sub-district

Table B.1: District-specific child sample, proportion of female children, median birth order (medBO), and under-five mortality per 1000 live births

sub-District	N	%Female	medBO	U-5Mort	sub-District	N	%Female	medBO	U-5Mort
Chitipa-rural	411	50.4	3	36.5	Ntcheu-rural	597	52.3	3	57.0
Chitipa-urban	75	56.0	2	66.7	Ntcheu-urban	67	38.8	3	74.6
Karonga-rural	426	49.1	3	46.9	Mangochi-rural	709	50.6	3	29.6
Karonga-urban	112	42.9	2	35.7	Mangochi-urban	118	45.8	2	33.9
Nkhatabay-rural	480	51.9	3	35.4	Machinga-rural	695	51.1	3	54.7
Nkhatabay-urban	72	52.8	3	27.8	Machinga-urban	82	39.0	2	0.0
Rumphi-rural	451	51.9	3	46.6	Zomba-rural	535	50.1	3	33.6
Rumphi-urban	99	50.5	2	70.7	Zomba-urban	148	52.7	2	20.3
Mzimba-rural	533	46.9	3	31.9	Chiradzulu-rural	484	49.6	3	47.5
Mzimba-urban	157	50.3	2	31.8	Chiradzulu-urban	21	33.3	2	95.2
Likoma-rural	337	48.1	3	29.7	Blantyre-rural	314	46.2	2	57.3
Likoma-urban	55	45.5	3	54.5	Blantyre-urban	319	47.3	2	53.3
Kasungu-rural	575	48.7	3	33.0	Mwanza-rural	366	56.1	3	43.7
Kasungu-urban	119	51.3	2	67.2	Mwanza-urban	91	57.1	3	22.0
Nkhotakota-rural	529	49.5	3	37.8	Thyolo-rural	476	49.6	3	37.8
Nkhotakota-urban	127	46.5	3	23.6	Thyolo-urban	57	35.1	2	70.2
Ntchisi-rural	531	51.0	3	52.7	Mulanje-rural	534	50.6	3	69.3
Ntchisi-urban	54	55.6	2	55.6	Mulanje-urban	66	57.6	2	75.8
Dowa-rural	540	44.8	3	50.0	Phalombe-rural	636	50.2	3	69.2
Dowa-urban	69	44.9	2	14.5	Phalombe-urban	48	50.0	2	41.7
Salima-rural	587	53.5	3	59.6	Chikwawa-rural	557	51.5	3	46.7
Salima-urban	108	47.2	2	27.8	Chikwawa-urban	37	51.4	3	27.0
Lilongwe-rural	500	52.0	3	56.0	Nsanje-rural	462	48.3	3	41.1
Lilongwe-urban	259	50.2	2	50.2	Nsanje-urban	105	51.4	2	28.6
Mchinji-rural	643	48.7	3	84.0	Balaka-rural	521	47.0	3	46.1
Mchinji-urban	85	47.1	2	11.8	Balaka-urban	99	45.5	2	20.2
Dedza-rural	556	51.1	3	57.6	Neno-rural	535	47.9	3	65.4
Dedza-urban	77	49.4	2	26.0	Neno-urban	40	50.0	2	0.0

Appendix C: Sample R codes for simulating data with perturbed parameters

```
rm(list = ls(all.names = TRUE))
library(survival)
library(simsurv)
library(foreign)

#A1
for (h in 1:1)
{
  for (LL in 1:100)
  {
    set.seed(5557*LL+7552)
    n <- 80
    j <- 10*h
    N <- j*n
    covariates1 <- data.frame(id=1:160,cluster=rep(1:2,each=n),x1=
       ↪ rbinom(160,1,0.70),x2=rnorm(160,0,1),b=rep(rnorm(2,6.5,1.
       ↪ 5),each=n))
    covariates2 <- data.frame(id=161:N,cluster=rep(3:j,each=n),x1=
       ↪ rbinom(N-160,1,0.70),x2=rnorm(N-160,0,1),b=rep(rnorm(j-2,
       ↪ 0,0.4),each=n))
    covariates = rbind.data.frame(covariates1,covariates2)
    parameter <- data.frame(x1=rep(0.5,each=N),x2=rep(1,each=N))
    survtimes <- simsurv(dist='weibull',lambdas=0.1,gammas=1,x=
       ↪ covariates,betas=parameter)
    censoring <-data.frame(id=1:N,censoring=rbinom(N,1,0.4))
    mydata <-merge(survtimes,covariates)
    mydata2 <-merge(mydata,censoring)
    w=dir.create(file.path(dirname("c:/"), paste0("Scene10b6.5-1.5",
       ↪ 100)))
    write.dta(mydata2, paste0("c:/","Scene10b6.5-1.5",100,"/Data",LL
       ↪ ,".dta"))
  }
}

#B3
for (c in 1:1)
{
  for (LLLLT in 1:100)
  {
    set.seed(4188812*LLLL+60000215)
  n <- 500
  j <- 10*c
  N <- j*n
  covariates <- data.frame(id=1:N,cluster=rep(1:j,each=n),x1=rbinom
     ↪ (N,1,0.70),x2=rnorm(N,0,1),b=rep(rnorm(j,0,0.4),each=n))
  parameter1 <- data.frame(x1=rep(2.7,each=1000),x2=rep(1,each=1000
     ↪ ))
  parameter2 <- data.frame(x1=rep(0.5,each=N-1000),x2=rep(1,each=N-
     ↪ 1000))
  parameter = rbind.data.frame(parameter1,parameter2)
  survtimes <- simsurv(dist='weibull',lambdas=0.1,gammas=1,x=
     ↪ covariates,betas=parameter)
  censoring <-data.frame(id=1:N,censoring=rbinom(N,1,0.4))
  mydata <-merge(survtimes,covariates)
  mydata2 <-merge(mydata,censoring)

  w=dir.create(file.path(dirname("c:/"), paste0("Case10X1-2.7",100)
      ↪ ))

  write.dta(mydata2, paste0("c:/","Case10X1-2.7",100,"/Data",LLLLT,
      ↪ ".dta"))

}


}


#C3

for (ca in 1:1)

{

  for (LLLL in 1:100)

  {

    set.seed(2715*LLLL+55213)

    n <- 80

    j <- 10*ca

    N <- j*n

    covariates <- data.frame(id=1:N,cluster=rep(1:j,each=n),x1=
        ↪ rbinom(N,1,0.70),x2=rnorm(N,0,1),b=rep(rnorm(j,0,0.4),
        ↪ each=n))

    parameter1 <- data.frame(x1=rep(0.5,each=160),x2=rep(2.5,each=16
        ↪ 0))

    parameter2 <- data.frame(x1=rep(0.5,each=N-160),x2=rep(1,each=N-
        ↪ 160))

    parameter = rbind.data.frame(parameter1,parameter2)

    survtimes <- simsurv(dist='weibull',lambdas=0.1,gammas=1,x=
        ↪ covariates,betas=parameter)

    censoring <-data.frame(id=1:N,censoring=rbinom(N,1,0.4))
    mydata <-merge(survtimes,covariates)
    mydata2 <-merge(mydata,censoring)
    w=dir.create(file.path(dirname("c:/"), paste0("Scene10X2-2.5",10
       ↪ 0)))
    write.dta(mydata2, paste0("c:/","Scene10X2-2.5",100,"/Data",LLLL
       ↪ ,".dta"))
  }
}
```

Appendix D: R code for computing group outlier statistic values from simulated data

```
#A. Fitting a clustered survival model and computing extended
   ↪ martingale, deviance, score residuals

rm(list=ls())
library(survival)
args(coxph)
library(foreign)
library(data.table)
library(dplyr)
library(readstata13)

dir <- setwd("c:/SimulationsII/Case50X1-2.7-X2-2.5-b15-5.51000") #
   ↪ data directory
dat <- list.files(dir, full.names = T) #dir should contain all your
   ↪  data1 to data1000
listdat <- lapply(dat,read.dta13) #change to read.dta if using
   ↪ earlier versions of stata than 13


coefmat <- matrix(NA,nrow =1000,ncol = 50+2) #define matrix that
   ↪ will take the 1000 simulations as rows and it will have
   ↪ columns taking 50 random effect estimates and 2 covariates
   ↪ coef of data
dim(coefmat)
```

```
pb <- txtProgressBar(min=1,max=1000,style = 3)
tmo<- Sys.time()


for(k in 1:1000)

{
  ntimes <-data.frame(listdat[[k]] %>%count(listdat[[k]]$cluster))$
     ↪ n #for displaying cluster sample sizes
  model <- coxph(Surv(as.numeric(eventtime),as.numeric(censoring))~
     ↪ x1+x2+frailty(cluster, distribution="gaussian",sparse=F,
     ↪ method="reml"),data=listdat[[k]])

  for (j in 1:50)
  {
    dt <- data.frame(cbind(newcluster=1:50,coefx1=rep(coef(model)[1
       ↪ ],50),coefx2=rep(coef(model)[2],50),randeffect=coef(model
       ↪ )[-2:-1]))
    dt2 <- as.data.frame(dt[rep(1:nrow(dt),ntimes),]) #ntimes
       ↪ defined above just before the loop
  }

  dt2$martingale <- listdat[[k]]$censoring - (0.1*listdat[[k]]$
     ↪ eventtime*exp(
    listdat[[k]]$x1*dt2$coefx1+listdat[[k]]$x2*dt2$coefx2+dt2$
       ↪ randeffect))
  dt2$sign=ifelse(dt2$martingale>0,1,-1)
  dt2$deviance <- dt2$sign*dt2$martingale*sqrt(-2*(dt2$martingale+
     ↪ listdat[[k]]$censoring*log(listdat[[k]]$censoring - dt2$
      ↪ martingale)))
  dt2$grandmean <- setDT(dt2)[,lapply(.SD,mean,na.rm=TRUE),.SDcols=
      ↪ "deviance"]
  listdat[[k]]<- data.frame(cbind(listdat[[k]],dt2))

  write.dta(listdat[[k]],file = paste0("ddat",k,".dta"))
  setTxtProgressBar(pb,k)
}
tm1<- Sys.time()
tm1 - tmo

#B. Computing group outlier statistic

outliermat <- matrix(NA,nrow = 50,ncol =8)
outliermat <- data.frame(outliermat)
colnames(outliermat) <- c("ID","meanclusdev","wtngrpVar","grandavg"
    ↪ ,"btwngrpVar","ratiovar","sqrtratio","stdratio")
pb <- txtProgressBar(min=1,max=1000,style = 3)
tmo<- Sys.time()
outliermat_all = matrix(NA,nrow = 50,ncol =8)
colnames(outliermat_all) <- c("ID","meanclusdev","wtngrpVar","
    ↪ grandavg","btwngrpVar","ratiovar","sqrtratio","stdratio")
for(k in 1:1000)
{
  outliermat[,1]<- 1:50
  outliermat[,2] <- setDT(listdat[[k]])[,lapply(.SD,mean,na.rm=TRUE
      ↪ ),by=cluster,.SDcols="deviance"][,2]
  outliermat[,3]<- setDT(listdat[[k]])[,lapply(.SD,var,na.rm=TRUE),
      ↪ by=cluster,.SDcols="deviance"][,2]
  outliermat[,4]<- setDT(listdat[[1]])[,lapply(.SD,mean,na.rm=TRUE)
     ↪ ,by=cluster,.SDcols="grandmean"][,2]
  outliermat[,5] <- sum(ntimes*(outliermat$meanclusdev - outliermat
     ↪ $grandavg))^{2}/(50-1)
  outliermat[,6]<- outliermat$wtngrpVar/outliermat$btwngrpVar
  outliermat[,7] <- sqrt(outliermat$ratiovar)
  outliermat[,8]<- (outliermat$sqrtratio - mean(outliermat$
     ↪ sqrtratio))/sqrt(var(outliermat$sqrtratio))

  if (k==1) {outliermat_all = outliermat}
  else {outliermat_all = rbind.data.frame(outliermat_all,outliermat
     ↪ )}



  setTxtProgressBar(pb,k)
}
write.dta(outliermat_all,file = paste0("outliermat_all",1000,".dta"
   ↪ ))

tm1<- Sys.time()
tm1 - tmo
```

Appendix E: R code for computing group influence statistic values from simulated data

```
rm(list=ls())
library(survival)
args(coxph)
library(foreign)
library(data.table)
library(dplyr)
library(readstata13)

# A. Fitting clustered survival model and computing extended
   ↪ martingale and leverage residuals

dir <- setwd("c:/SimulationsIII/Case10X1-1.8100") #data directory
dat <- list.files(dir, full.names = T) #dir should contain all your
   ↪  data1 to data100
listdat <- lapply(dat,read.dta13) #change to read.dta if in stata
   ↪ earlier versions than 13

pb <- txtProgressBar(min=1,max=100,style = 3) #set process system
   ↪ to go thru 100 data files
tmo<- Sys.time() #set start time

for(k in 1:100)
```

```
{
 ntimes <-data.frame(listdat[[k]] %>%count(listdat[[k]]$cluster))$
   ↪ n #count sample size in each cluster

 model <- coxph(Surv(as.numeric(eventtime),as.numeric(censoring))~
   ↪ x1+x2+frailty(cluster, distribution="gaussian",sparse=F,
   ↪ method="reml"),data=listdat[[k]])

 for (j in 1:10)
 {
  dt <- data.frame(cbind(newcluster=1:10,coefx1=rep(coef(model)[1
    ↪ ],10),coefx2=rep(coef(model)[2],10),randeffect=coef(model
    ↪ )[-2:-1]))
  dt2 <- as.data.frame(dt[rep(1:nrow(dt),ntimes),]) #ntimes
    ↪ defined above just before the loop
 }

 dt2$martingale <- listdat[[k]]$censoring - (0.1*listdat[[k]]$
   ↪ eventtime*exp(
  listdat[[k]]$x1*dt2$coefx1+listdat[[k]]$x2*dt2$coefx2+dt2$
    ↪ randeffect)) #0.1 is chosen baseline hazard

 listdat[[k]]<- data.frame(cbind(listdat[[k]],dt2))

 listdat[[k]]$numerator_x1 <- listdat[[k]]$x1 *exp(listdat[[k]]$
   ↪ coefx1*listdat[[k]]$x1 +listdat[[k]]$coefx2*listdat[[k]]$x2
   ↪  +listdat[[k]]$randeffect)
 listdat[[k]]$denominator_x1 <- exp(listdat[[k]]$coefx1*listdat[[k
   ↪ ]]$x1 +listdat[[k]]$coefx2*listdat[[k]]$x2 +listdat[[k]]$
```

```
    ↪ randeffect)
  listdat[[k]]$numerator_x2 <- listdat[[k]]$x2 *exp(listdat[[k]]$
    ↪ coefx1*listdat[[k]]$x1 +listdat[[k]]$coefx2*listdat[[k]]$x2
    ↪  +listdat[[k]]$randeffect)
  listdat[[k]]$sum_numx1 <- setDT(listdat[[k]])[,lapply(.SD,sum,na.
    ↪ rm=TRUE),by=cluster,.SDcols="numerator_x1"][,2][rep(1:nrow(
    ↪ dt),ntimes),]
  listdat[[k]]$sum_denx1 <- setDT(listdat[[k]])[,lapply(.SD,sum,na.
    ↪ rm=TRUE),by=cluster,.SDcols="denominator_x1"][,2][rep(1:
    ↪ nrow(dt),ntimes),]
  listdat[[k]]$sum_numx2 <- setDT(listdat[[k]])[,lapply(.SD,sum,na.
    ↪ rm=TRUE),by=cluster,.SDcols="numerator_x2"][,2][rep(1:nrow(
    ↪ dt),ntimes),]

  listdat[[k]]$leverage_x1 <- listdat[[k]]$x1 - (listdat[[k]]$sum_
    ↪ numx1/listdat[[k]]$sum_denx1)
  listdat[[k]]$leverage_x2 <- listdat[[k]]$x2 - (listdat[[k]]$sum_
    ↪ numx2/listdat[[k]]$sum_denx1)

  listdat[[k]]$scoresd_x1 <- listdat[[k]]$martingale * listdat[[k
    ↪ ]]$leverage_x1
  listdat[[k]]$scoresd_x2 <- listdat[[k]]$martingale * listdat[[k
    ↪ ]]$leverage_x2

  write.dta(listdat[[k]],file = paste0("ddat",k,".dta"))
  setTxtProgressBar(pb,k)
}
tm1<- Sys.time()
tm1 - tmo
```

```
# B. Computing group score residual

influmat <- matrix(NA,nrow =10,ncol =5) #generate matrix where rows
    ↪  are 5 clusters and 5 columns for keeping influence values
influmat <- data.frame(influmat)
colnames(influmat) <- c("ID","Influ_x1","Influ_x2","stdInflu_x1","
    ↪ stdInflu_x2") #assign names to influence columns for x1 and x
    ↪ 2
pb <- txtProgressBar(min=1,max=100,style = 3)
tmo<- Sys.time()
influmat_all = matrix(NA,nrow =10,ncol =5)
colnames(influmat_all) <- c("ID","Influ_x1","Influ_x2","stdInflu_x1
    ↪ ","stdInflu_x2")
for(k in 1:100)
{
  influmat[,1]<- 1:10
  influmat[,2] <- setDT(listdat[[k]])[,lapply(.SD,sum,na.rm=TRUE),
      ↪ by=cluster,.SDcols="scoresd_x1"][,2]
  influmat[,3] <- setDT(listdat[[k]])[,lapply(.SD,sum,na.rm=TRUE),
      ↪ by=cluster,.SDcols="scoresd_x2"][,2]
  influmat[,4] <- (influmat$Influ_x1-mean(influmat$Influ_x1))/sd(
      ↪ influmat$Influ_x1,na.rm = T)
  influmat[,5] <- (influmat$Influ_x2-mean(influmat$Influ_x2))/sd(
      ↪ influmat$Influ_x2,na.rm = T)

  if (k==1) {influmat_all = influmat}
  else {influmat_all = rbind.data.frame(influmat_all,influmat)}
  setTxtProgressBar(pb,k)
}
write.dta(influmat_all,file = paste0("influmat_all",100,".dta")) #
   ↪ save matrix of 100 influence values for each of 10 clusters

tm1<- Sys.time()
tm1 - tmo
```

Appendix F: R code for applying derived outlier statistic on child survival data used

```
rm(list=ls())
library(foreign)
library(survival)
args(coxph)
library(data.table)
library(dplyr)
library(readstata13)

# 1. Read stata data in R, fit model and compute univariate
   ↪ outliers

mydata2 = read.dta("c:/Users/User/Desktop/2015␣DHS/cleaned2DHS.dta"
   ↪ ,convert.factors=T)
tmo<- Sys.time()

clustersize <-data.frame(mydata2 %>%count(v023))$n #count cluster (
   ↪ v023) sizes
clustersize

for(k in 1:1)
{
       model <- coxph(Surv(as.numeric(time_death),as.numeric(death_
          ↪ status))~as.factor(Female_Child)+birth_order+birth_
          ↪ order2+frailty(v023, distribution="gaussian",sparse=F,
          ↪ method="reml"),data=mydata2)

       for (j in 1:56)
       {
              dt <- data.frame(cbind(newCluster=1:56,coefSex=rep(
                 ↪ coef(model)[1],56),coefbord=rep(coef(model)[2],
                 ↪ 56),coefbord2=rep(coef(model)[3],56),randeffect
                 ↪ =coef(model)[-3:-1]))
              dt2 <- as.data.frame(dt[rep(1:nrow(dt),clustersize)
                 ↪ ,])
       }

       dt2$martingale <- mydata2$death_status - (0.063*mydata2$time
          ↪ _death*exp(
       mydata2$Female_Child*dt2$coefSex++mydata2$birth_order*dt2$
          ↪ coefbord+mydata2$birth_order2*dt2$coefbord2+dt2$
          ↪ randeffect))

       dt2$sign=ifelse(dt2$martingale>0,1,-1)
       dt2$deviance <- dt2$sign*dt2$martingale*sqrt(-2*(dt2$
          ↪ martingale+mydata2$death_status*log(mydata2$death_
          ↪ status - dt2$martingale)))
       dt2$grandmean <- setDT(dt2)[,lapply(.SD,mean,na.rm=TRUE),.
          ↪ SDcols="deviance"]

       mydata20 <- data.frame(cbind(mydata2,dt2))

       write.dta(mydata20, paste0("c:/Users/User/Desktop/mydata20.
          ↪ dta"))
```

```
}

tm1<- Sys.time()

tm1 - tmo


#B. Computing cluster outliers


outliermat <- matrix(NA,nrow =56,ncol = 6)

outliermat20 <- data.frame(outliermat)

colnames(outliermat20) <- c("ID","meanclusdev","wtngrpVar","
   ↪ grandavg","btwngrpVar","ratiovar")


tmo<- Sys.time()


for(k in 1:1)

{

        outliermat20[,1]<- 1:56

        outliermat20[,2]<- setDT(mydata20)[,lapply(.SD,mean,na.rm=
           ↪ TRUE),by=v023,.SDcols="deviance"][,2]

        outliermat20[,3]<- setDT(mydata20)[,lapply(.SD,var,na.rm=
           ↪ TRUE),by=v023,.SDcols="deviance"][,2]

        outliermat20[,4]<- setDT(mydata20)[,lapply(.SD,mean,na.rm=
           ↪ TRUE),by=v023,.SDcols="grandmean"][,2]

        outliermat20[,5]<- sum(clustersize*(outliermat20$meanclusdev
           ↪  - outliermat20$grandavg))^{2}/(56-1)

        outliermat20[,6]<- outliermat20$wtngrpVar/outliermat20$
           ↪ btwngrpVar


        outliermat20 = cbind.data.frame(outliermat20)

}
```

```
write.dta(outliermat20, paste0("c:/Users/User/Desktop/outlierd20.
   ↪ dta"))


tm1<- Sys.time()

tm1 - tmo
```

Appendix G: R code for applying derived influence statistic on child survival data used

```
rm(list=ls())
library(survival)
args(coxph)
library(foreign)
library(ggplot2)
library(ggrepel)
library(dplyr)
library(data.table)
library(readstata13)

#A. Fitting clustered survival model and computing extended
   ↪ martingale and leverage residuals

ourdata = read.dta("C:/Users/User/Desktop/r␣codes␣compiled/
   ↪ influence3.dta",convert.factors=F)
tmo<- Sys.time()
ntimes <-data.frame(ourdata %>%count(v023))$n #picks sample size in
   ↪  each cluster v023
ntimes

for(k in 1:1)
{
       model <- coxph(Surv(as.numeric(time_death),as.numeric(death_
          ↪ status))~as.factor(Female_Child)+birth_order+birth_
          ↪ order2+frailty(v023, distribution="gaussian",sparse=F,
          ↪ method="reml"),data=ourdata)

       for (j in 1:56)
       {
              dt <- data.frame(cbind(newDist=1:56,coefSex=rep(coef(
                 ↪ model)[1],56),coefbord=rep(coef(model)[2],56),
                 ↪ coefbord2=rep(coef(model)[3],56),randeffect=
                 ↪ coef(model)[-3:-1]))
              dt2 <- as.data.frame(dt[rep(1:nrow(dt),ntimes),])
       }

       dt2$martingale <- ourdata$death_status - (0.063*ourdata$time
          ↪ _death*exp(ourdata$Female_Child*dt2$coefSex+ourdata$
          ↪ birth_order*dt2$coefbord+ourdata$birth_order2*dt2$
          ↪ coefbord2+dt2$randeffect))

       ourdata <- data.frame(cbind(ourdata,dt2))

       ourdata$numerator_Sex <- ourdata$Female_Child *exp(ourdata$
          ↪ Female_Child*ourdata$coefSex+ourdata$birth_order*
          ↪ ourdata$coefbord+ourdata$birth_order2*ourdata$coefbord
          ↪ 2+ourdata$randeffect)

       ourdata$numerator_bord <- ourdata$birth_order *exp(ourdata$
          ↪ Female_Child*ourdata$coefSex+ourdata$birth_order*
          ↪ ourdata$coefbord+ourdata$birth_order2*ourdata$coefbord
          ↪ 2+ourdata$randeffect)

       ourdata$numerator_bord2 <- ourdata$birth_order2 *exp(ourdata
          ↪ $Female_Child*ourdata$coefSex+ourdata$birth_order*
          ↪ ourdata$coefbord+ourdata$birth_order2*ourdata$coefbord
          ↪ 2+ourdata$randeffect)

       ourdata$denominator_Sex <- exp(ourdata$Female_Child*ourdata$
          ↪ coefSex+ourdata$birth_order*ourdata$coefbord+ourdata$
          ↪ birth_order2*ourdata$coefbord2+ourdata$randeffect)

       ourdata$sum_numSex <- setDT(ourdata)[,lapply(.SD,sum,na.rm=
          ↪ TRUE),by=v023,.SDcols="numerator_Sex"][,2][rep(1:nrow(
          ↪ dt),ntimes),]
       ourdata$sum_numbord <- setDT(ourdata)[,lapply(.SD,sum,na.rm=
          ↪ TRUE),by=v023,.SDcols="numerator_bord"][,2][rep(1:nrow
          ↪ (dt),ntimes),]
       ourdata$sum_numbord2 <- setDT(ourdata)[,lapply(.SD,sum,na.rm
          ↪ =TRUE),by=v023,.SDcols="numerator_bord2"][,2][rep(1:
          ↪ nrow(dt),ntimes),]

       ourdata$sum_denSex <- setDT(ourdata)[,lapply(.SD,sum,na.rm=
          ↪ TRUE),by=v023,.SDcols="denominator_Sex"][,2][rep(1:
          ↪ nrow(dt),ntimes),]

       ourdata$leverage_Sex <- ourdata$Female_Child - (ourdata$sum_
          ↪ numSex/ourdata$sum_denSex)
       ourdata$leverage_bord <- ourdata$birth_order - (ourdata$sum_
          ↪ numbord/ourdata$sum_denSex)
       ourdata$leverage_bord2 <- ourdata$birth_order2 - (ourdata$
          ↪ sum_numbord2/ourdata$sum_denSex)
```

```
       ourdata$scoresdSex <- ourdata$martingale * ourdata$leverage_
          ↪ Sex
       ourdata$scoresdbord <- ourdata$martingale * ourdata$leverage
          ↪ _bord
       ourdata$scoresdbord2 <- ourdata$martingale * ourdata$
          ↪ leverage_bord2

       write.dta(ourdata,file = "data44.dta")

}
tm1<- Sys.time()
tm1 - tmo

# B. Computing group score residual

influmat <- matrix(NA,nrow =56,ncol =4)
influmat <- data.frame(influmat)
colnames(influmat) <- c("ID","Influ_Sex","Influ_bord","Influ_bord2"
   ↪ )

tmo<- Sys.time()
influmat_all = matrix(NA,nrow =56,ncol =4)
colnames(influmat_all) <- c("ID","Influ_Sex","Influ_bord","Influ_
   ↪ bord2")
for(k in 1:1)
{
       influmat[,1]<- 1:56
       influmat[,2] <- setDT(ourdata)[,lapply(.SD,sum,na.rm=TRUE),
          ↪ by=v023,.SDcols="scoresdSex"][,2]
```

```
        influmat[,3] <- setDT(ourdata)[,lapply(.SD,sum,na.rm=TRUE),
           ↪ by=v023,.SDcols="scoresdbord"][,2]
        influmat[,4] <- setDT(ourdata)[,lapply(.SD,sum,na.rm=TRUE),
           ↪ by=v023,.SDcols="scoresdbord2"][,2]
}
write.dta(influmat,file = "influence3.dta")


tm1<- Sys.time()
tm1 - tmo
```