Residual Diagnostics and Statistical Inference for Shared Frailty Models

A dissertation submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Public Health
University of Saskatchewan
Saskatoon

By
Tingxuan Wu

Permission to Use

In presenting this dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my dissertation work was done. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my dissertation.

Disclaimer

Reference in this dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this dissertation in whole or part should be addressed to:

Director of School of Public Health
Health Sciences Building E-Wing, 104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 2Z4
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Frailty models are commonly used for analyzing clustered survival data and accounting for unobserved heterogeneity. Shared frailty models are random-effect models in which the frailties are shared among individuals within groups. Several R packages, such as survival, frailtyEM, fraltypack, frailtysurv, and frailtyHL, are available for fitting shared frailty models. However, little research has been conducted to compare their performances, leaving users without clear guidance in selecting an appropriate tool for analyzing clustered survival data. The first study in this thesis aims to address this gap by providing an overview of current R packages for fitting shared frailty models and comparing their performances through simulation studies. After fitting a shared frailty model, model diagnostics are an essential part of the modelling process. The use of residuals in assessing model adequacy is a conventional tool for normal regression. In the second study of this thesis, we propose to use the Z-residual for detecting the non-linearity in the shared frailty model. Through a simulation study, we investigate the power of Z residuals in detecting non-linear effects in covariates and demonstrate their effectiveness in diagnosing models using real data on the survival of acute myeloid leukemia patients. Typically, all residuals in survival analysis are calculated using the full dataset, resulting in a bias problem due to double usage of the dataset. In the third study of this thesis, we propose applying cross-validation methods to compute residuals for diagnosing a semi-parametric shared frailty model and investigate the performance of cross-validatory Z-residual for diagnosing a shared frailty model with non-parametric baseline hazards. We compare Z-residuals calculated through three methods: without cross-validation (No-CV) method which is the basic algorithm, 10-fold cross-validation (10-fold) and leave-one-out cross-validation (LOOCV). Through simulation studies, we investigate their performances in the detection of nonlinear effects in covariates and identification of the outliers in the dataset through graphical visualization and overall GOF test. We also compared No-CV Z-residual and LOOCV Z-residual in a real data application for identifying outliers for a kidney infection dataset. Finally, in the fourth study of this thesis, we extended the Z residual to diagnose the proportional hazards assumption and compare it with existing residual methods.

Acknowledgements

I would like to begin by expressing my sincere gratitude to my supervisors, Dr. Longhai Li and Dr. Cindy Feng, for their unwavering support, guidance, and patience throughout my Ph.D. study. Their academic and financial assistance, as well as their encouragement and kindness, have been instrumental in helping me complete this thesis. I am truly grateful for their mentorship, which has not only shaped my research but also inspired me for future endeavours. It has been an honour to work under their supervision.

I wish to extend a special thank you to my external appraiser, Professor Xuewen Lu. I am deeply grateful for the invaluable feedback and insights from the committee members, Professors Lloyd Balbuena, June Lim, and Li Xing. Their comments have played a significant role in shaping the quality of my thesis. I would also like to express my appreciation for the support given by the director of the Biostatistics programs, Professor Alexander Crizzle.

I am deeply thankful to the School of Public Health and the Department of Mathematics and Statistics at the University of Saskatchewan for their financial and academic support throughout my Ph.D. program. I am grateful to all the professors, graduate students, and staff in both departments for their invaluable assistance and support during my studies.

I would like to extend special thanks to my parents for their unwavering love and support throughout my life. Their encouragement has been a constant source of strength throughout my Ph.D. program. I also want to express my gratitude to my husband, Haonan Tian, for his unwavering support and care during my studies. I am truly grateful for their love, support, and understanding.

Finally, I would like to express my heartfelt appreciation to everyone who has supported me throughout my Ph.D. journey. Your encouragement, assistance, and friendship have made this experience more rewarding and unforgettable. I am deeply grateful for everything you have done. Thank you.

Contents

List of Tables

List of Figures

List of Abbreviations

AFT	Accelerated Failure Time
AUC	The Area Under the Receiver Operating Characteristic Curve
CDF	Cumulative Distribution Function
CHF	Cumulative Hazard Function
CI	Confidence Interval
CS	Cox-Snell
CV	Cross-Validation
CP	Coverage Probability
EM	Expectation-Maximization
FPR	False Positive Rate
GOF	Goodness-of-Fit
HL	Hierarchical Likelihood
KM	Kaplan-Meier
KS	Kolmogorov-Smirnov
LP	Linear Predictor
LOOCV	Leave-One-Out Cross-Validation
LOWESS	The Locally Weighted Scatterplot Smoothing
MML	Maximum Marginal Likelihood
MPL	Maximization Penalized Likelihood
MSE	Mean Square Error
NRSP	The Normally-Transformed Randomized Survival Probability
PPL	Penalized Partial Likelihood
PFL	Pseudo Full Likelihood
PH	Proportional Hazard
RSP	Randomized Survival Probability
ROC	Receiver Operating Characteristic
SP	Survival Probability
SW	Shapiro-Wilk
SF	Shapiro-Francia

1 Introduction

1.1 Background

In many public health and epidemiological studies, clustered survival data are often observed. This clustering imposes a correlation among individuals within each cluster, which is known as within-cluster correlation. For instance, within each long-term care facility, the elderly within the same care facility often have access to similar healthcare which may affect the individuals' health status. Other typical examples include matched measurements on similar organs like eyes and kidneys. These patients who received organs from the same individual can be considered as a cluster and share some common unobserved heterogeneity. As such, individuals sharing the same hidden features may correlate with each other. The shared random effect (frailty) models [1–4] were introduced as a generalization of the survival models allowing a random effect due to the unobserved heterogeneity of each cluster.

Several R packages are available for fitting shared frailty models, including survival [5], frailtyEM [6], frailtySurv [7], frailtyHL [8], parfm [9], and frailtypack [10]. However, little research has been conducted to compare the performance of various R packages for fitting shared frailty models, making it difficult for users to decide on an appropriate tool for analyzing clustered survival data. Therefore, a thorough comparison of the performance of these packages will be useful for users who need to fit shared frailty models.

Despite the availability of many R packages for fitting shared frailty models, the examination of model assumptions is often neglected, partly due to limitations in the available diagnostic tools for survival models. A residual diagnosis is a conventional tool for assessing the validity of model assumptions in normal linear regression. However, in shared frailty models with censored observations, residual diagnostics are not as straightforward as those for a normal linear regression model. For diagnosing shared frailty models, several residuals have been proposed [11]; such as Cox-Snell (CS) [12], martingale [13], deviance [14, 15] Schoenfeld [11, 16] and scaled Schoenfeld residuals[11, 17]. Each of these residuals for diagnosing survival models has its own limitations.

The most widely used tool for checking survival regression models for failure times is Cox-Snell (CS) residual [12]. It is calculated as the negative logarithm of the estimated survival probability. When failure times are not censored, the survival probability is uniformly distributed when the model is true; therefore, the CS residual is exponentially distributed. However, when there are censored observations, CS residuals are no longer exponentially distributed since the survival probability is not uniformly distributed. To account for censored observations, diagnostics based on CS residuals compare the agreement of cumulative hazard plot of CS residuals estimated with Kaplan-Meier method [18] and the 45° straight line, which is the cumulative hazard of the standard exponential distribution. Although the cumulative hazard plot of CS residuals is widely used in the graphical overall GOF assessment of the adequacy of a fitted model, the overall GOF test reveals little information about the nature of the model inadequacies. Tailored graphical and numeric diagnostic tools are therefore needed.

Martingale and deviance residuals have been often used to check the functional form of covariates in

survival models. Martingale residuals are defined as the difference between the observed and expected values of a subject's failure indicator, integrated over the time at risk. They can help to identify outliers and assess the functional form of covariates. Deviance residuals are a normalized transform of martingale residuals that are approximately symmetrically distributed about zero with a mean of zero when the fitted model is appropriate. Both types of residuals are available in the `survival` package in R. However, martingale residuals are asymmetric, with no lower bound and an upper bound of one, making them difficult to visually inspect. Deviance residuals are less skewed and more normally distributed than martingale residuals. Although the locally weighted scatterplot smoothing (LOWESS) lines on the scatterplots of the residuals against the continuous covariates are useful for revealing patterns in the residuals that would not otherwise be perceived, visual inspection of LOWESS lines can be still subjective. Furthermore, martingale and deviance residuals lack a reference distribution due to censoring, so it is challenging to derive a numerical test to measure the statistical significance of the observed pattern in residual plots.

Schoenfeld residuals and Scaled Schoenfeld residuals are the most commonly used graphical diagnostic tools to check the proportional hazards (PH) assumption in the survival models. The PH assumption states that the hazard ratio between two groups is constant over time. Schoenfeld residuals represent the difference between the observed and expected values of the covariate(s) given the risk set at that time. Scaled Schoenfeld residuals, proposed by Grambsch and Therneau in 1994, are a variation of Schoenfeld residuals that are scaled by the variance of the covariate. Scaled Schoenfeld residuals are a modification of Schoenfeld residuals that can be more effective in detecting departures from the proportional hazards assumption, especially when the variance of the covariate changes over time or is small [17]. The graphical diagnostics based on scaled Schoenfeld residuals commonly use a plot of the Scaled Schoenfeld residuals versus the observed survival time, and a horizontal line at zero indicates that the PH assumption is satisfied. A score test is commonly used to check the proportional hazards (PH) assumption in Cox regression models by comparing the observed values of the score function to their expected values under the null hypothesis of PH. The test is used to assess whether the hazard ratio between two groups is constant over time and to determine if there is evidence of non-proportional hazards. The `survival` package in R can compute the score test [5].

However, Schoenfeld residuals-based diagnostics only consider a specific case of PH violations due to the time-varying covariate effect, and they may have low power in detecting other kinds of PH violations, such as accelerated failure time. When the survival time does not follow the Cox model form but rather an accelerated failure time model, Schoenfeld residuals-based diagnostics may fail to detect Non-PH. Therefore, the detection of Non-PH using Schoenfeld residuals and related tests is not comprehensive.

Li et al. [19] proposed using randomized survival probabilities (RSPs) to define residuals for checking the assumptions of accelerated failure time (AFT) models without random effects. The key idea of RSP is to replace the survival probability of a censored failure time with a uniform random number between zero and the survival probability of the censored time. The RSPs are uniformly distributed under the true model, hence, can then be transformed into normally distributed residuals with the normal quantile function. The resulting residual is called the normally-transformed RSP (NRSP) residual. Statistical tests can be derived based on NRSP residuals for checking model assumptions, such as distributional assumptions and the functional form of covariates, given the normally distributed reference distribution for NRSP residuals. However, NRSP residuals have not been extended to diagnose Cox proportional hazard models or semi-parametric shared frailty models.

1.2 Literature Review

1.2.1 Shared Frailty Models

A shared frailty model is a frailty model where the frailties are common or shared among individuals within groups. The formulation of a frailty model for clustered failure survival data is defined as follows. Suppose there are g groups of individuals with n_i individuals in the ith group, $i = 1, 2, \ldots, g$. If the number of subjects n_i is 1 for all groups, then the univariate frailty model is obtained [3]. Otherwise, the model is called the shared frailty model [2, 20, 21] because all subjects in the same cluster share the same frailty value z_i. Suppose t_{ij} is the true failure time for the jth individual of the ith group, which we assume to be a continuous random variable in this article, where $j = 1, 2, \ldots, n_i$. Let t^*_{ij} denote the realization of t_{ij}. In many practical problems, we may not be able to observe t^*_{ij} exactly, but we can observe that t_{ij} is greater than a value c_{ij}, where c_{ij} be the corresponding censoring time. The observed failure times are denoted by the pair (y_{ij}, δ_{ij}), where $y_{ij} = \min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} < c_{ij})$. The observed data can be written as $y = (y_{11}, \cdots, y_{gn_g})$ and $\delta = (\delta_{11}, \cdots, \delta_{gn_g})$. This is called right-censoring. Since we will consider only the right-censoring in this article, we will use "censoring" as a short for "right-censoring".

For the shared frailty models, the hazard of the event at time t for the jth individual, $j = 1, 2, \ldots, n_i$, in the ith group, is then

$$h_{ij}(t) = z_i \exp(x_{ij}\beta) h_0(t); \tag{1.1}$$

and the survival function for the jth individual of the ith group at time t follows:

$$S_{ij}(t) = \exp\left\{ - \int_0^t h_{ij}(t)\, \mathrm{d}t \right\} = \exp\left\{ - z_i \exp(x_{ij}\beta) H_0(t) \right\}, \tag{1.2}$$

where x_{ij} is a row vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \cdots, x_{gn_g})$; β is the column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, $H_0(t)$ is the baseline cumulative hazard function, and z_i is the frailty term that is common for all n_i individuals within the ith group, let $z = (z_1, \cdots, z_g)$. The hazard and survival functions with frailty effect can also be written as:

$$h_{ij}(t) = \exp(x_{ij}\beta + u_i) h_0(t), \tag{1.3}$$

and

$$S_{ij}(t) = \exp\left\{ - \exp(x_{ij}\beta + u_i) H_0(t) \right\}, \tag{1.4}$$

where $u_i = \log(z_i)$ is a random effect in the linear component of the proportional hazards model. Note that z_i cannot be negative, but u_i can be any value. If u_i is zero, the corresponding to z_i is one, which means the model does not have frailty. The form of the baseline hazard function may be assumed to be unspecified as a semi-parametric model or fully specified to follow a parametric distribution.

In our study, we focus mainly on the shared gamma frailty model, since gamma distribution is the most common distribution for modelling the frailty effect. It is easy to obtain a closed-form representation of the observable survival, cumulative distribution, and hazard functions due to the simplicity of the Laplace transform. The gamma distribution is a two-parameter distribution with a shape parameter k and scale parameter θ. It takes a variety of shapes as k varies: when $k = 1$, it is identical to the well-known exponential distribution; when k is large, it takes a bell-shaped form reminiscent of a normal distribution; when k is less than one, it takes exponentially shaped and asymptotic to both the vertical and horizontal axes. Under

the assumption $k = \frac{1}{\theta}$, the two-parameter gamma distribution turns into a one-parameter distribution. The expected value is one and the variance is equal to θ.

1.2.2 Review of Existing Residual Diagnostic Methods for Survival Models

Cox-Snell residuals

Cox-Snell (CS) residual [12] is the most widely used tool for checking survival regression models for failure times. CS residuals are transformed from the survival probabilities with the quantile function of the exponential distribution, which is defined as

$$r_{ij}^{c}(t_{ij}) = -\log(S_{ij}(t_{ij})) \tag{1.5}$$

where t_{ij} is the true failure time. In the absence of censored observations, the survival probability is uniformly distributed when the model is true; therefore, the CS residual is exponentially distributed. A plot of the CHF against the true failure time will give a straight line through the origin with a unit slope when the residuals have a unit exponential distribution, which is expected when the survival model is correctly specified. In addition to the graphical checking, we can apply numerical GOF testing methods such as Kolmogorov-Smirnov (KS) test to CS residuals. When there are censored failure times, the distribution of $S_{ij}(y_{ij})$ is no longer uniformly distributed under the true model, which means the CS residuals are no longer exponentially distributed. The CS residuals r_{ij}^{c} can be regarded as a dataset with censoring. The Kaplan-Meier (KM) estimate of the survivor function can still be computed for CS residuals. Hence, the most widely used diagnostics tool is to apply the KM method to get an estimate of the CHF of CS residuals and compare the CHF against the 45^{o} straight line.

Martingale residuals

Martingale residuals [13] can be viewed as the difference between the observed value of a subject's failure indicator and its expected value, integrated over the time for which that patient was at risk. The residuals are defined as

$$r_{ij}^{M} = \delta_{ij} - r_{ij}^{c} \tag{1.6}$$

where δ_{ij} is the event indicator for the jth individual of the ith group observation, δ_{ij} is equal to 1 if that observation is an event; otherwise zero if censored, and r_{ij}^{c} is the Cox-Snell residual.

The martingale residuals sum to zero but are not symmetrically distributed about zero when the fitted model is appropriate. The scatterplot of the residual against the survival time or the continuous covariates is used for assessing the functional form covariates and identifying outliers in the survival data. The locally weighted scatterplot smoothing (LOWESS) lines on the scatterplots can be used for revealing patterns in the residuals that would not otherwise be perceived.

Deviance residuals

The deviance residuals [14, 15] can be regarded as an attempt to make the martingale residuals symmetrically distributed about zero, and are defined as

$$r_{ij}^{D} = sgn(r_{ij}^{M})[-2(r_{ij}^{M} + \delta_{ij}\log\big(\delta_{ij} - r_{ij}^{M}\big))]^{\frac{1}{2}} \tag{1.7}$$

where r_{ij}^M is the martingale residual, the function $sgn(.)$ is the sign function [11].

Deviance residuals are less skewed and more normally distributed than martingale residuals, the plots based on deviance residuals tend to be easier to interpret. An index plot of Deviance residual can be used to identify individuals with unusual survival times. Additionally, scatterplots of deviance residuals against continuous covariates are also used in assessing the functional form of covariates, while LOWESS lines on the scatterplots can reveal patterns in the residuals.

Schoenfeld residuals

Schoenfeld residuals [11, 16] were originally termed partial residuals. It can overcome two disadvantages of the Cox Snell (CS) residuals [12], which depend heavily on the observed survival time and require an estimate of the cumulative hazard function. In [16], Schoenfeld residual is perceived as the difference between the observed and expected values of the x_{ijp} given the risk set at the time y_{ij}, one for each covariate in the fitted Cox regression model, defined as

$$r_{S_{ijp}} = \delta_{ij}(x_{ijp} - \hat{a}_{ijp}), \tag{1.8}$$

where x_{ijp} is the value of the pth covariate, $p = 1, 2, \cdots, P$, for the jth individual of the ith group in the study,

$$\hat{a}_{ijp} = \frac{\sum_{lh \in R(y_{ij})} x_{lhp} z_l \exp\left(x_{lh}\hat{\beta}\right)}{\sum_{lh \in R(y_{ij})} z_l \exp\left(x_{lh}\hat{\beta}\right)}, \tag{1.9}$$

and $R(t_{ij})$ is the set of all individuals at risk at time y_{ij}.

Schoenfeld residuals are a commonly used graphical diagnostic method to evaluate the proportional hazards assumption of a Cox PH model. If the PH assumption is met, the scatterplot of Schoenfeld residuals versus survival time should be flat, centred around zero without exhibiting any pattern.

Scaled Schoenfeld residuals, proposed by Grambsch and Therneau (1994) [17], are a variation of Schoenfeld residuals that are scaled by the variance of the covariate [11, 17]. The purpose of scaling is to make the residuals more sensitive to departures from the PH assumption, especially in cases where the variance of the covariate changes over time or is very small. By scaling, the residuals are put on a similar scale, which makes it easier to compare them and identify patterns of deviation from the PH assumption. The widely used scaled Schoenfeld residuals are defined as

$$r^{\star}_{S_{ij}} = d\text{var}(\hat{\beta}) r_{S_{ij}}, \tag{1.10}$$

where d is the number of events among all observed failure times, and $\text{var}(\hat{\beta})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. Diagnostics based on scaled Schoenfeld residuals commonly use a plot of the scaled Schoenfeld residuals against the observed survival time. A plot that shows a random pattern around a horizontal line without showing a trend indicates the plausibility of the PH assumption.

1.3 Research Contributions

Contribution I: A Comparative Study of R Packages for Shared Frailty Models

Several R packages are available for fitting shared frailty models, including `survival` [5], `frailtyEM` [6], `frailtySurv` [7], `frailtyHL` [8], `parfm` [9], and `frailtypack` [10]. In this study, we provided a general

overview of these R packages and comprehensively assess their algorithms and applications, creating a summary for each. We compare their performance through simulation studies, including the bias and variance of the parameter estimates, rate of convergence, and computational time of the packages. We found that all the R packages produced similar and unbiased parameter estimates for fixed-effect regression coefficients, regardless of the cluster sizes and censoring rates. However, differences were found with respect to the estimation of the variance parameter for the frailty term, convergence rate and computational time. In addition, we note that estimating standard errors for frailty variance is not straightforward, and not all R packages provide standard errors for the frailty variance. Furthermore, many packages assume that the estimated frailty variance follows a roughly normal distribution, despite the fact that it is positively skewed, which makes a symmetrical confidence interval inappropriate. To address this issue, we developed a confidence interval for frailty variance, which demonstrated superior performance compared to the conventional confidence intervals provided by some R packages. We thoroughly examined the strengths and limitations of the R packages.

Contribution II: Z-residual Diagnostics for Detecting Misspecification of the Functional Form of Covariates for Shared Frailty Model

We have extended the concept of randomized survival probabilities (RSPs) and created a residual diagnostic tool that can provide both graphical and numerical tests to assess the functional form of covariates in semi-parametric shared frailty models. We have renamed this residual "Z-residuals" for simplicity, as the letter Z is commonly used to represent a standard normal random variable. To implement this diagnostic tool, we developed a general function that computes Z-residuals for semi-parametric shared frailty models using the output from the `coxph` function in the `survival` package in R. We also proposed a non-homogeneity test to determine if there is a trend in Z-residuals. To evaluate the performance of this diagnostic tool in detecting misspecification of the functional form of covariates, we conducted simulation studies. Our results showed that the non-homogeneity test based on Z-residuals has greater power and satisfactory type I error compared to the overall goodness-of-fit (GOF) tests in detecting misspecification of the covariate functional form. In a real data application involving modelling the survival time of acute myeloid leukemia patients, the Z-residual diagnosis revealed that a model with log-transformation is not suitable for modelling the survival time, a finding was not detected by other diagnostic methods.

Contribution III: Cross-validatory Z-Residual for Diagnosing Shared Frailty Models

Computing residuals based on the full dataset can result in a conservative bias that reduces the power of detecting model misspecification, as the same dataset is used for both model fitting and validation. Although cross-validation is a potential solution to this problem, it has not been commonly used in residual diagnostics due to computational challenges and a lack of awareness of the severity of the bias caused by the double use of the dataset. We proposed a cross-validation approach for computing Z-residuals in the context of shared frailty models and compare them to the Z-residual without cross-validation. Specifically, we developed a general function that calculates cross-validated Z-residuals using the output from the `coxph` function in the `survival` package in R. We also built an R function for splitting data into K-fold to ensure adequate representations of groups and other covariates in each fold. In our study design, the Z-residuals are calculated using three methods: the full dataset (No-CV), 10-fold cross-validation (10-fold) and leave-one-out cross-validation (LOOCV). We conduct simulation studies to investigate the performances of the three types of

Z-residuals in detecting nonlinear covariate effects and identifying outliers through graphical visualization and overall GOF tests. Our simulation studies demonstrate that the Shapiro-Wilk (SW) tests based on cross-validatory Z-residuals are significantly more powerful and more discriminative than No-CV Z-residuals for detecting non-linear covariate effects and identifying outliers. We also compare the performance of the No-CV Z-residuals and LOOCV Z-residuals in identifying outliers in a real application that models the recurrence time of kidney infection patients. Our findings suggest that LOOCV Z-residuals can identify outliers that are missed by Z-residuals without cross-validation.

Contribution IV: Z-Residual Diagnostics for Detecting Non-Proportional Hazards for Survival Models

Schoenfeld residuals-based diagnostics only consider a specific case of violations of PH due to the time-varying covariate effect. Therefore, it may have low power in detecting other kinds of violations of the PH assumption, for example, accelerated failure time. The accelerated failure time (AFT) model incorporates a wide variety of survival time distributions, which is used in many fields as an alternative to the Cox PH model if the PH assumption is not tenable. In this study, we propose using the Z-residual diagnostics for detecting Non-PH in a survival model. Our simulation studies show that, compared to the score tests related to Schoenfeld residuals, the tests based on Z-residuals have similar powers and type I error rates in time-varying covariate effect scenarios but they have significantly higher powers in AFT scenarios. In a real data application, using Z-residual diagnostics, we identify a severe violation of the PH assumption in the blindness time of a Diabetic Retinopathy study, which was not detected by Schoenfeld residuals and related tests.

1.4 Organization of the Thesis

The remaining chapters of this thesis are structured as follows. Chapters 2 through 5 describe four research studies in detail. Chapter 2 presents a comparative study of R packages for fitting shared frailty models, in a complete publishable article format. Chapter 3, also in a publishable article format, examines Z-residual diagnostics for detecting misspecification of the functional form of covariates for shared frailty models. Chapter 4 provides a complete publishable article on cross-validatory Z-residual for diagnosing shared frailty models. In Chapter 5, we investigate the performance of Z-residual diagnostic tools for detecting non-proportional hazards for survival models. Chapter 6 summarizes the findings and includes a discussion of the limitations of each study, as well as potential future research directions.

2 A Comparative Study of R Packages for Shared Frailty Models [1]

Abstract: Frailty models are often used to model the unobserved heterogeneity and clustered survival data. A shared frailty model is a random-effect model where the frailties are common or shared among individuals within groups. Different R packages are available for fitting shared frailty models such as survival, frailtyEM, fraltypack, frailtysurv, and frailtyHL. However, little research has been conducted to compare the performance of various R packages for fitting shared frailty models, making it difficult for users to decide on an appropriate tool for analyzing clustered survival data. We aim to compare the performance of the R packages via a series of simulation studies. The bias and variance of the parameter estimates, rate of convergence, and computational time of the packages are compared. The advantages and limitations of the software are discussed in detail.

2.1 Introduction

In survival analysis, conventional Cox proportional hazard models [22] and accelerated failure time models [23] assume that subjects are independent of one another. However, data with a multilevel structure occur frequently in a wide range of research problems. For example, patients are often nested within hospitals. The hazard of the event differs from one cluster to another cluster induced by unobserved cluster-level factors. Random effects can be incorporated into conventional survival models to account for cluster-level heterogeneity. Such heterogeneity is often called frailty in the context of survival analysis. Frailty models extend the classic survival models by incorporating random effects (frailties) acting multiplicatively on the baseline hazard function [24]. In cases where the frailty is greater than one, subjects experience an increased hazard of failure. A shared frailty model is a frailty model where the frailties are common or shared among individuals within a cluster or group [1–4].

Different R packages are available for fitting shared frailty models, such as `survival` [5], `frailtyEM` [6], `frailtySurv` [7], `frailtyHL` [8], `parfm` [9], and `frailtypack` [10]. The `survival`, `frailtyEM`, `frailtySurv`, `frailtyHL` packages can be used to implement semi-parametric survival models with frailties. The top downloaded package for fitting a shared frailty model is `survival` package, which estimates the parameters by maximizing the penalized partial likelihood [25, 26]. Since the frailty term can be assumed as a latent variable, the general expectation-maximization (EM) algorithm [27, 28] is implemented by `frailtyEM` package. The `frailtySurv` package implements the parameter estimation via a pseudo full likelihood approach [29, 30] and `frailtyHL` estimates the parameters through a hierarchical-likelihood approach [31]. The `parfm` package can be used to fit parametric shared frailty models. Examples of such distribution are exponential, Weibull, inverse Weibull, Gompertz, lognormal, log-skewNormal, or loglogistic, among others. The parameter estimation is performed based on the maximum marginal likelihood (MML) approach [32, 33].

[1] This paper has been submitted for possible publication in Statistics in Biosciences

The `frailtypack` package fits flexible parametric frailty models for the gamma and log-normal distributions (including correlated random effects, nested random effects and other scenarios). The maximization of the penalized log-likelihood [34, 35] is used in `frailtypack` for estimating parameters.

Despite the wide range of R packages for fitting frailty models, it remains unclear if the R packages have similar or different performances in terms of precision and efficiency of parameter estimates, computational speed and convergence rate. Early research [36] included three R packages for fitting shared frailty models in their comparison via simulation studies. However, in recent years, more R packages have been developed for fitting shared frailty models. An updated comparison of R packages is therefore needed. The current study aims at filling this gap by providing a general overview of current R packages for fitting shared frailty models and comparing their performances via simulation studies. Our simulation studies showed that the current R packages with the default parameter settings considered for fitting the shared frailty models yielded very similar and unbiased parameter estimates for the fixed-effect regression coefficients, regardless of the cluster sizes and censoring rates. However, differences were found with respect to the estimation of the variance parameter for the frailty term, convergence rate and computational time. In addition, inference for frailty variance is not straightforward. Not all the currently available R packages provide an estimation of standard errors for the frailty variance. Most packages assume the distribution of the estimated frailty variance is approximately normally distributed. However, frailty variance is positively skewed, so a symmetric confidence interval is not ideal. To circumvent this problem, we developed a confidence interval for frailty variance and showed its better performance compared to the conventional confidence intervals provided in some of the R packages.

The rest of the article is structured as follows. Section 2.2 gives a brief review of shared frailty models. Section 2.3 introduces the parameter estimation methods and corresponding R packages. Section 2.4 presents the design and results of the simulation study for comparing the performance of the R packages. Finally, we conclude the paper with a discussion of the advantages and limitations of each R package for fitting a shared frailty model in Section 2.5. The recommendations for the selection of a package for fitting shared frailty models are also presented in Section 2.5.

2.2 Shared Frailty Models

A shared frailty model is a frailty model where the frailties are common or shared among individuals within groups. The formulation of a frailty model for clustered failure survival data is defined as follows. Suppose there are g groups of individuals with n_i individuals in the ith group, $i = 1, 2, \ldots, g$. If the number of subjects n_i is 1 for all groups, then the univariate frailty model is obtained [3]. Otherwise, the model is called the shared frailty model [2, 20, 21] because all subjects in the same cluster share the same frailty value z_i. Suppose t_{ij} is the true failure time for the jth individual of the ith group, which we assume to be a continuous random variable in this article, where $j = 1, 2, \ldots, n_i$. Let t^*_{ij} denote the realization of t_{ij}. In many practical problems, we may not be able to observe t^*_{ij} exactly, but we can observe that t_{ij} is greater than a value c_{ij}, where c_{ij} be the corresponding censoring time. The observed failure times are denoted by the pair (y_{ij}, δ_{ij}), where $y_{ij} = \min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} < c_{ij})$. The observed data can be written as $y = (y_{11}, \cdots, y_{gn_g})$ and $\delta = (\delta_{11}, \cdots, \delta_{gn_g})$. This is called right-censoring. Since we will consider only the right-censoring in this article, we will use "censoring" as a short for "right-censoring".

For the shared frailty models, the hazard of the event at time t for the jth individual, $j = 1, 2, \ldots, n_i$,

in the ith group, is then

$$h_{ij}(t) = z_i \exp(x_{ij}\beta) h_0(t); \tag{2.1}$$

and the survival function for the jth individual of the ith group at time t follows:

$$S_{ij}(t) = \exp\Big\{ - \int_0^t h_{ij}(t)\, \mathrm{d}t \Big\} = \exp\Big\{ - z_i \exp(x_{ij}\beta) H_0(t) \Big\}, \tag{2.2}$$

where x_{ij} is a row vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \cdots, x_{gn_g})$;$\beta$ is the column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, $H_0(t)$ is the baseline cumulative hazard function, and z_i is the frailty term that is common for all n_i individuals within the ith group, let $z = (z_1, \cdots, z_g)$. The hazard and survival functions with frailty effect can also be written as:

$$h_{ij}(t) = \exp(x_{ij}\beta + u_i) h_0(t), \tag{2.3}$$

and

$$S_{ij}(t) = \exp\Big\{ - \exp(x_{ij}\beta + u_i) H_0(t) \Big\}, \tag{2.4}$$

where u_i= $\log(z_i)$ is a random effect in the linear component of the proportional hazards model. Note that z_i cannot be negative, but u_i can be any value. If u_i is zero, the corresponding to z_i is one, which means the model does not have frailty. The form of the baseline hazard function may be assumed to be unspecified as a semi-parametric model or fully specified to follow a parametric distribution.

In our study, we focus mainly on the shared gamma frailty model, since gamma distribution is the most common distribution for modelling the frailty effect. It is easy to obtain a closed-form representation of the observable survival, cumulative density, and hazard functions due to the simplicity of the Laplace transform. The gamma distribution is a two-parameter distribution with a shape parameter k and scale parameter θ. It takes a variety of shapes as k varies: when $k = 1$, it is identical to the well-known exponential distribution; when k is large, it takes a bell-shaped form reminiscent of a normal distribution; when k is less than one, it takes exponentially shaped and asymptotic to both the vertical and horizontal axes. Under the assumption $k = \frac{1}{\theta}$, the two-parameter gamma distribution turns into a one-parameter distribution. The expected value is one and the variance is equal to θ.

2.3 Estimation and Inference for Shared Frailty Models

2.3.1 Baseline Hazard Estimation

Breslow Method

In the Cox proportional hazards regression, the Breslow estimator [37] is the nonparametric maximum likelihood estimation for the cumulative baseline hazard function. It has been implemented in all major statistical software packages. The baseline cumulative hazard function is $H_0(t) = \int_0^t h_0(s)\, ds$. Breslow (1972) suggested estimating the cumulative baseline hazard via maximizing likelihood function. After getting the estimators $\hat{\beta}'$ and $\hat{u_i}$, it can provide the nonparametric maximum likelihood estimator of $\hat{H}_0(t)$:

$$\hat{H}_0(t) = \sum_{\{v: y_{(v)} \le t\}} \left\{ \frac{d_{(v)}}{\sum\limits_{(i,j) \in R(y_{(v)})} \exp\Big(x_{ij}\hat{\beta} + \hat{u_i} \Big)} \right\}, \tag{2.5}$$

where $y_{(1)} < \cdots < y_{(r)}$ are the ordered distinct event time among the y_{ij}'s and $R(y_{(v)}) = \{(i,j) : y_{ij} \geq y_{(v)}\}$ is the risk set at $y_{(v)}$, i.e., $d_{(v)}$ is the number of events at $y_{(v)}$.

Splines Method

The splines estimator [38] is an alternative solution to estimate the baseline hazard function $h_0(t)$. The basis of splines with a specified number of knots Q for the baseline hazard can be approximated as: $\hat{h}_0(t) = \sum_{i=1}^{m} \lambda_i M_i(t)$, with $m = Q + 2$, where λ_i's are positive parameters and $M_i(t)$'s are called the M-spline basis functions [39]. M-splines are considered a variant of the normalized version of B-splines [38] within boundary knots. Cubic M-splines is a polynomial function of 3^{rd} order that approximates a function on an interval by combining linearly. The sum of the polynomial functions of 1^{st} order can be approximated to the second derivative of the baseline hazard function $h_0''(t;\lambda)$. This approximation can reduce the number of parameters and allows flexible shapes of the hazard function. The results are closer to the true hazard function if more knots can be used.

2.3.2 Parameter Estimation Methods and R Packages

Penalized Partial Likelihood (PPL) Algorithm (R package: `survival`)

The penalized partial likelihood (PPL) approach can be used to estimate parameters in a shared frailty model [40]. This estimation is based on maximizing the penalized partial log-likelihood, which consists of two parts. The first part is the conditional likelihood of the data given the frailties. The second part corresponds to the frailties distribution in which the likelihood is considered a penalty term. The PPL for the frailty model is then given by

$$l_{ppl}(\beta, u, \theta; y, \delta) = l_{part}(\beta, u; y, \delta) + l_{pen}(\theta; u), \tag{2.6}$$

over both β and u. Here $l_{part}(\beta, u)$ is the partial log-likelihood for the Cox model that includes the random effects.

$$l_{part}(\beta, u; y, \delta) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \delta_{ij} \left\{ \eta_{ij} - \log \left[\sum_{(q,w) \in R(y_{ij})} \exp(\eta_{qw}) \right] \right\}, \tag{2.7}$$

where $\eta_{ij} = x_{ij}\beta + u_i$ and $\eta = (\eta_{11}, \ldots, \eta_{gn_g})$. In the penalty function $l_{pen}(\theta; u)$, θ is the parameter for frailty, chosen by the investigator to restrict the values of u. The random effect u_i is equal to $\log(z_i)$, where z_i is usually taken to have either a lognormal or a gamma distribution. The penalty function can be written as,

$$l_{pen}(\theta; u) = \sum_{i=1}^{g} \log f_U(u_i \mid \theta), \tag{2.8}$$

where $f_U(u_i)$ denotes the density function of the random effect u_i.

The maximisation of the PPL consists of an inner and an outer loop. For the log-normal frailty effects with mean zero and variance θ, the penalized likelihood can be maximized with the Newton-Raphson algorithm in the inner loop. The maximisation process proceeds iteratively by starting with a provisional θ and finding the estimates of the β's and the u's that maximise $l_{ppl}(\beta, u, \theta)$. In the outer loop, the restricted maximum likelihood estimator for the θ is obtained using the best linear unbiased predictors. The process is iterated until convergence. For the gamma frailty effects with unit mean and variance θ, the estimates of β's and the u's are first taken to be values that maximize $l_{ppl}(\beta, u, \theta)$ for a given value of the θ. The outer loop is

based on the maximisation of a profiled version of the marginal likelihood for θ, at estimates, $\hat{\beta}$ and $\hat{u}$, is then used to obtain a revised estimate of θ. The Breslow approximation is the first option to estimate the baseline hazard function in nearly the currently available R packages for fitting Cox regression models with or without frailties.

Arguably the most popular R package for fitting semiparametric shared frailty models is the `survival` package [5]. The function `coxph` in the `survival` package offers a way of fitting shared frailty models via the PPL method. The arguments are the terms including fixed effects of the model, random effects and the data. The frailty distribution can be gamma, Gaussian, or t distribution. It accommodates the clustered failures and recurrent events data with right, left, and interval censoring types. When `coxph` function fits shared frailty models with clustered failures data, cluster size should be above five. Otherwise, the random effects will be treated as fixed effects.

Expectation-maximization (EM) Algorithm (R package: `frailtyEM`)

The expectation-maximization (EM) algorithm [2, 27] is an iterative method for performing maximum likelihood estimation when the model involves latent variables (missing values). The expectation (E) step attempts to estimate the latent variables via the expectation of the log-likelihood evaluated based on the observed data. The maximization (M) step attempts to optimize the parameters of the model, which computes parameters by maximizing the expected log-likelihood found in the E step. If we consider the frailty effect z as missing data in the frailty model, the problem can be approached by using EM Algorithm. In the E step, compute the unobserved frailties as the expected values conditional on the observed information and the current parameter estimates are obtained. In the maximization step, we treat these expected values as true information, and new estimates of the parameters of interest are obtained by maximization of the likelihood, given the expected values.

We first consider the complete data log-likelihood in which the frailties z_i are regarded as another set of parameters:

$$l_{full}(\theta, \beta, z) = \log f(y, \delta, z \mid h_0, \beta, \theta) = l_{full,1}(\beta; y, \delta, \hat{h}_0) + l_{full,2}(\theta; z), \tag{2.9}$$

where,

$$l_{full,1}(\beta; y, \delta, \hat{h}_0) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left\{ \delta_{ij} \log \left[\hat{h}_0(y_{ij}) z_i \exp(x_{ij}\beta) \right] - \hat{H}_0(y_{ij}) z_i \exp(x_{ij}\beta) \right\} \tag{2.10}$$

and

$$l_{full,2}(\theta; z) = \sum_{i=1}^{g} \log f_Z(z_i \mid \theta). \tag{2.11}$$

We use $l_{full,1}(\beta; y, \delta, \hat{h}_0)$ to estimate β, and $l_{full,2}(\theta; z)$ to estimate θ. Within the framework of the EM algorithm, the expected value of the full log-likelihood needs to be maximised. In the E step, the "posterior" distribution of the frailties $p(z_i \mid y_i, \delta_i, \beta^{(k-1)}, \theta^{(k-1)})$ can be obtained. Then, the $E^{(k)}(z_i)$ and $E^{(k)}(\log z_i)$ can be calculated. In the M step, the loglikelihood in (2.9) is profiled to a partial loglikelihood by considering the frailties as fixed offset terms, then the $E^{(k)}(z_i)$ and $E^{(k)}(\log z_i)$ are considered to be the true value to replace the z_i's and $\log z_i$'s in the partial loglikelihood leading to

$$l^{(k)}_{part,1}(\beta) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \delta_{ij} \left\{ E^{(k)}(\log z_i) + x_{ij}\beta - \log \left(\sum_{(q,w) \in R(y_{ij})} E^{(k)}(z_q) \exp(x_{qw}\beta) \right) \right\}. \tag{2.12}$$

The new estimates $\beta^{(k)}$ can be obtained from the $l_{part,1}(\beta)$. A new estimate $\theta^{(k)}$ can be obtained immediately

by maximization of $l_{full,2}(\theta; u)$, replacing z_i's and $\log z_i$'s in (2.11) by the current expected values at iteration step k. The Breslow estimator is applied to estimate the baseline hazard function, which is required in the expectation step. In the initialization E step, $\theta^{(0)}$ is set to one and an ordinary Cox model is fitted leading to estimates $\beta^{(0)}$. Next, we iterate between the expectation and maximization steps until convergence. The marginal loglikelihood can be used for assessing the convergence of the algorithm

The `frailtyEM` package was written by Theodor et al. [6]. It provides maximum likelihood estimation of semiparametric shared frailty models using the expectation-maximization algorithm. The main model fitting function in frailtyEM is emfrail, and the user has to define the main arguments formula, data set, distribution and control. This formulation is common to most survival analysis packages, allowing for several scenarios, including possibly left truncated clustered failures and recurrent events in both calendar time and gap time formulation. The distribution argument determines the frailty distribution; the gamma, stable and power variance function family distributions are supported. The control argument can be provided by the emfrail's control() function, and it controls parameters for emfrail. The package can access predicted survival and cumulative hazard curves, both for an individual and on a population level. The results from `frailtyEM` package are very close to the survival package.

Maximum Marginal Likelihood (MML) Algorithm (R package: `parfm`)

The maximum marginal likelihood (MML) approach was proposed for estimating the parameters for shared frailty models [32, 33]. The frailties are integrated out by averaging the conditional likelihood with respect to the frailty distribution. This method can be applied to any frailty distribution with explicit Laplace transform.

For the right-censored clustered survival data, the observation for the jth individual in the ith group are the triple $(y_{ij}, \delta_{ij}, x_{ij})$. Further, if left-truncation is also present, truncation times τ_{ij} are gathered in the vector τ, i.e., $\tau = (\tau_{11} \cdots, \tau_{gn_g})$. Let ψ represent a vector of parameters for the baseline hazard function. The marginal log-likelihood can be written as

$$\begin{aligned} l_{marg}(\psi, \beta, \theta; y, \delta, \tau, x) = \sum_{i=1}^{g} \Bigg\{ \Bigg[\sum_{j=1}^{n_i} \delta_{ij} (\log(h_0(y_{ij} \mid \psi)) + x_{ij}\beta) \Bigg] \\ + \log \left[(-1)^{d_i} \mathcal{L}^{d_i} (\sum_{j=1}^{n_i} H_0(y_{ij} \mid \psi) \exp(x_{ij}\beta)) \right] \\ - \log \left[\mathcal{L} (\sum_{j=1}^{n_i} H_0(\tau_{ij} \mid \psi) \exp(x_{ij}\beta)) \right] \Bigg\}, \end{aligned} \tag{2.13}$$

where θ is used as the vector of parameters for the frailty distribution function, $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ the number of events in the i-th cluster, and $\mathcal{L}^q(\cdot)$ is the q-th derivative of the Laplace transform of the frailty distribution, which is defined as,

$$\mathcal{L}(s) = \int_0^\infty \exp(-sz) f(z)\, dz, \tag{2.14}$$

where $f(z)$ is the density function of frailty term z. If the higher-order derivatives $\mathcal{L}^q(\cdot)$ of the Laplace transform up to $q = \max\{d_1, \ldots, d_G\}$ are able to compute, the estimates of ψ, β, θ, can be obtained by maximising the marginal log-likelihood (2.13).

The `parfm` package [9] estimates the parameters for parametric frailty models by maximizing the marginal

log-likelihood. The baseline hazard distributions can be exponential, Weibull, inverse Weibull (Frechet), Gompertz, lognormal, log-skewNormal, and loglogistic. The frailty distribution can be gamma, positive stable, inverse Gaussian, and lognormal distribution.

Hierarchical Likelihood (HL) Algorithm (R package: `frailtyHL`)

Lee & Nelder [41] proposed the use of hierarchical likelihood for fitting the model with random effects. The hierarchical likelihood consists of data, parameters and unobserved random effects. This method can avoid the integration over the random-effect distributions. The method is the statistically efficient estimation in frailty models by using the Laplace approximation. Thus, the h-likelihood can be used directly for inference on random effects.

For the observe y_{ij} and the censoring indicator is δ_{ij}, the h-likelihood for a frailty model is defined by

$$hl(\beta, \theta, u; y, \delta, \hat{h}_0) = l_0(\beta; y, \delta, z, \hat{h}_0) + l_1(\theta; u), \tag{2.15}$$

where l_0 is the sum of conditional log densities for (y, δ) given the random effect $u = (\log z_1, \cdots, \log z_g)$; then it follows:

$$\begin{aligned} l_0(\beta; y, \delta, u, \hat{h}_0) &= \sum_{ij} \log f(y_{ij}, \delta_{ij} \mid \beta, u_i, \hat{h}_0) \\ &= \sum_{ij} \delta_{ij} \Big\{ \log \hat{h}_0(y_{ij}) + (x_{ij}\beta + u_i) \Big\} - \sum_{ij} \Big\{ \hat{H}_0(y_{ij}) \exp(x_{ij}\beta + u_i) \Big\} \end{aligned} \tag{2.16}$$

l_1 is the sum of log densities for random effects u with parameter θ, which is defined by

$$l_1(\theta; u) = \sum_i \log f_U(u_i \mid \theta). \tag{2.17}$$

The Breslow estimator is employed to estimate the baseline hazard function $\hat{h}_0$. From the equation (2.6), the penalized partial likelihood is defined as,

$$l_{ppl}(\beta, u, \theta) = \sum_{ij} \delta_{ij} \Big\{ (x_{ij}\beta + u_i) - \log \Big[\sum_{(q,w) \in R(y_{ij})} \exp(x_{qw}\beta + u_q) \Big] \Big\} + \sum_i \log f_U(u_i \mid \theta). \tag{2.18}$$

The papers [31, 42] showed that $hl(\theta, \beta)$ is equal to the $l_{ppl}(\beta, u, \theta)$ plus a constant,

$$hl(\beta, u, \theta) = l_{ppl}(\beta, u, \theta) + \sum_{(q,w) \in R(y_{ij})} d_{qw} \Big\{ \log \hat{h}_0(y_{qw}) - 1 \Big\}, \tag{2.19}$$

where $\sum_{(q,w) \in R(y_{ij})} d_{qw} \Big\{ \log \hat{h}_0(y_{qw}) - 1 \Big\}$ is a constant and d_{qw} is the number of element in the risk set $R(y_{qw})$. Accordingly, given the frailty parameter θ, the hierarchical likelihood methods for estimating the parameter estimator β can be obtained by maximizing the profile marginal likelihood after eliminating $H_0(t)$. The Laplace approximation can be used when the marginal likelihood is hard to obtain. Given $\hat{\beta}$ and $\hat{u}$, the maximum adjusted profile hierarchical likelihood for estimating the frailty variance θ can be obtained. We iterate these steps until convergence. The estimates of the standard errors can be computed.

The `frailtyHL` package created by Ha et al. [8] implements the hierarchical-likelihood procedures for

fitting semi-parametric frailty models with non-parametric baseline hazards. The package fits shared or multilevel frailty models for correlated survival data. The lognormal or gamma distributions can be adopted as the frailty distribution, corresponding to the normal or log-gamma distributions for the log frailties. The results of estimates of fixed effects, random effects, and variance components as well as their standard errors are provided. In addition, it provides a statistical test for the variance components of frailties and three AIC criteria for the model selection. However, the package does not provide the interval estimation of frailty.

Pseudo Full Likelihood (PFL) Algorithm (R package: `frailtySurv`)

Pseudo full likelihood [29, 30] is a new method that can handle any parametric frailty distribution with finite moments. A simple univariate numerical integration can deal with non-conjugate frailty distributions. The cumulative hazard function is estimated via a noniterative procedure. Other properties follow the consistency and asymptotic normality of the parameter estimates and a direct, consistent covariance estimator. It is easy to compute and implement. From the study of Gorfine et al. [30], the estimation results for fitting the shared frailty model are very similar to the EM-based method.

In the shared frailty model, we assume further that the observed data consisting of y, δ, x are independent. The proposed approach can estimate the regression coefficient column vector β, the frailty distribution's parameter θ, and the non-parametric cumulative baseline hazard H_0. Let τ be the end of the observation period. The full likelihood can be defined as

$$L(\beta, \theta, H_0) = \prod_{i=1}^{g} \prod_{j=1}^{n_i} \left\{ h_0(y_{ij}) \exp(x_{ij}\beta) \right\}^{\delta_{ij}} \prod_{i=1}^{g} (-1)^{N_{i\cdot}(\tau)} \mathcal{L}^{(N_{i\cdot})}\{H_{i\cdot}(\tau)\}, \tag{2.20}$$

where $N_{ij}(t) = \delta_{ij} I(y_{ij} \leq t)$, $N_{i\cdot}(t) = \sum_{j=1}^{n_i} N_{ij}(t)$, $H_{ij}(t) = H_0(\min\{y_{ij}, t\}) \exp(x_{ij}\beta)$, $H_{i\cdot}(t) = \sum_{j=1}^{n_i} H_{ij}(t)$, $\mathcal{L}$ is the Laplace transform of the frailty distribution and $\mathcal{L}^{(m)}, m = 1, 2, \cdots$ are the mth derivatives of $\mathcal{L}$. Note that the mth derivatives of the Laplace transform evaluated at $H_{i\cdot}(\tau)$ equals to $(-1)^{N_{i\cdot}(\tau)} \int z^{N_{i\cdot}(\tau)} \exp\{-zH_{i\cdot}(\tau)\} f(z)\, dz$. The log-likelihood equals to

$$l(h_0, \theta, \beta) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left\{ \delta_{ij} \log\{h_0(y_{ij}) \exp(x_{ij}\beta)\} \right\} + \sum_{i=1}^{g} \log \mathcal{L}^{(N_{i\cdot})}\{H_{i\cdot}(\tau)\}. \tag{2.21}$$

Obviously, an estimator of H_0 is required in the log-likelihood function to obtain estimators of β and θ. In the initialization step, θ should be set as a value and a standard Cox model is fitted to obtain initial estimates of β. For given these two initial values, H_0 is estimated via the Breslow estimator with jumps at the ordered observed failure times τ_v, $v = 1, \cdots, r$. The detailed step of the baseline hazard estimation is referred to by Gorfine et al. [30]. Then, $\hat{H}_0$ is substituted into the log-likelihood function. The estimators of $\hat{\beta}$ and $\hat{\theta}$ can be obtained by maximizing the log-likelihood function. Iterate these steps until convergence.

The R package `frailtySurv` [7] can be used for simulating and fitting semi-parametric shared frailty models. It can be applied for a variety of frailty distributions, including gamma, log-normal, inverse Gaussian and power variance functions v a pseudo full likelihood approach. The parameters' estimators are consistent and asymptotically normally distributed. The results of this package can be performed using the normal distribution, such as hypothesis testing and confidence intervals. Only right-censoring with clustered failures dataset is supported by `frailtySurv`.

Maximization Penalized Likelihood (MPL) Algorithm (R package: `frailtypack`)

The maximum penalized likelihood estimation [34, 35] can be applied to the nonparametric estimation of a continuous hazard function in a shared frailty model. This approach is based on the penalized full likelihood, which is opposed to the penalized partial likelihood. We assume that the frailty effects are distributed from a gamma distribution with mean 1 and variance θ. For the observe y, δ, and the truncation times τ, the full marginal loglikelihood for the shared gamma frailty model has an analytical formulation [28]

$$\begin{aligned} l(\beta,\theta,h_0) = \sum_{i=1}^{g}\Bigg\{\Bigg[\sum_{j=1}^{n_i}\delta_{ij}\log h_0(y_{ij})\Bigg] - (\frac{1}{\theta}+m_i)\log\Bigg[1+\theta\sum_{j=1}^{n_i}H_0(y_{ij})\Bigg] \\ +\frac{1}{\theta}\log\Bigg[1+\theta\sum_{j=1}^{n_i}H_0(\tau_{ij})\Bigg] + I(m_i\neq 0)\sum_{k=1}^{m_i}\log\Big(1+\theta(m_i-k)\Big)\Bigg\} \end{aligned} \tag{2.22}$$

where the number of recurrent events is $m_i = \sum_{j=1}^{n_i}\delta_{ij}$.

The penalized loglikelihood function for the shared gamma frailty model follows

$$pl(\beta,\theta,h_0) = l(\beta,\theta,h_0) - k\int_0^\infty {h_0}''(t)^2\,dt \tag{2.23}$$

where k is a positive smoothing parameter that controls the trade-off between the data fit and the smoothness of the functions. The smoothing parameter needs to be a fixed value, and the estimators of β and θ can be obtained via the maximization of the penalized likelihood. The robust Marquardt algorithm [43] is used to estimate parameters, which is a combination between a Newton Raphson algorithm and the steepest descent algorithm. The estimator of the baseline hazard function $h_0(\cdot)$ can be approximated on the basis of Cubic M-splines with Q knots. The splines, the regression coefficients, and the variance of the frailty term are initialized to 0.1 in the shared frailty model. The model can be fit firstly, then adjusted Cox model to give new initial values for the splines and the regression coefficients.

The `frailtypack` package [10] allows fitting Cox models and four types of frailty models (shared, nested, joint, additive). The function `frailtyPenal` fits the shared frailty model by using the MPL method with the splines to estimate the baseline hazard. Right-censored or left-truncated data are considered in this package. The arguments are the terms including the fixed effect, the cluster variable, and the data set. In addition, there are three arguments in the formula that need to be specified: n.knots (4 up to 20), kappa1 (smoothing parameter), and frailty (a logical value).

Table 2.1 provides a summary of the above-mentioned six R packages for fitting shared frailty models in terms of the frailty distribution, algorithm, censoring type and data type.

Table 2.1: R packages for fitting shared frailty models in terms of the primary R function, frailty distribution, fitting algorithm, censoring type and data type. PPL = penalized partial likelihood, MML = maximum marginal likelihood, EM = expectation maximization, PFL = pseudo full likelihood, HL = h-likelihood, MPL = maximization penalized loglikelihood.

Package	Version	Function	Frailty distribution	Algorithm	Censoring	Data
survival	3.2-13	coxph	Gamma, Log-normal, t	PPL	Right, interval, Left	Clustered failures, Recurrent events
parfm	2.7.6	parfm	Gamma, Log-normal, Positive Stable, Inverse Gaussian	MML	Right	Clustered failures, Left truncation
frailtyEM	1.0.1	emfrail	Gamma, Positive Stable, Inverse Gaussian, Compound Poisson, Power Variance Function	EM	Right	Clustered failures, Recurrent events, Left truncation
frailtySurv	1.3.7	fitfrail	Gamma, Log-normal, Inverse Gaussian, Power Variance Function	PFL	Right	Clustered failures
frailtyHL	2.3	frailtyHL	Gamma, Log-normal	HL	Right	Clustered failures
frailtypack	3.4.0	frailtyPenal	Gamma, Log-normal	MPL	Right	Clustered failures, Recurrent events, Left truncation, Correlated structure

2.4 Simulations and Results

We conducted simulation studies to investigate the performances of R packages for fitting the shared frailty models. We generated the true failure time from a Weibull regression model with shape parameter ($\alpha = 3$) and scale parameter ($\lambda = 0.007$). More specifically $t_{ij} = \{-\log(u_{ij})/[\lambda \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)} + z_i\Big)]\}^{(1/\alpha)}$, where $i = \{1,\cdots, \text{g}\}$, $j = \{1,\cdots, n_i\}$ and u_{ij} was simulated from Uniform[0, 1]. The censoring time C_i was simulated from an exponential distribution, $\exp(\theta)$, where θ was set to obtain three different censoring rates (c): $20\%, 50\%$, and 80%, respectively. To investigate if the performances of R packages depend on sample size, we simulated datasets with varying sample sizes n ranging from 100 to 800. For a sample of size 100, the observations were grouped into 10 clusters of size 10. For a sample of size 400, the observations were grouped into 10 clusters of size 40 or 40 clusters of size 10. For a sample of size 800, the observations were grouped into 10 clusters of size 80 or 80 clusters of size 10. Three covariates were generated including $x_{ij}^{(1)}$ from a Uniform[0, 1], $x_{ij}^{(2)}$ from a Normal(0, 1), and $x_{ij}^{(3)}$ from a Bern(0.25). We set true regression parameters for the three covariates as $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 0.5$, respectively. The frailty term was generated from a gamma distribution with a variance of 0.5. We considered fitting a shared frailty gamma model assuming $h_{ij}(t_{ij}) = z_i \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)}\Big) h_0(t_{ij})$ as a true model. All current considered R packages were applied to the same simulated dataset in each scenario. Using 1000 datasets generated under each scenario, we examined the precision of the parameter estimates in terms of bias and standard errors of the estimated parameters, as well as the coverage probability (CP) of the estimated parameters. We also investigated the performance of the packages in terms of convergence rate and average computing time under each simulation scenario.

2.4.1 Estimated parameters

Figure 2.1 presents the estimated regression coefficients over 1000 repeated samples when the sample size $n = 400$. The results indicate that current available packages performed similarly. Only `frailtypack` slightly overestimated β_2. Not surprisingly, as the censoring rate increases, the estimated regression coefficients are subject to more variability.

As displayed in the top panels of Figure 2.2, the estimated variance parameter of the frailty term was underestimated. This phenomenon was more pronounced with 10 clusters of size 40 compared to 40 clusters of size 10, suggesting a lower number of clusters leads to higher variability of the estimated variance parameter of the frailty term. The distribution of the variance parameter of the random effect term is known to be positively skewed [44]. To better visualize the distribution of the estimated frailty variance parameter, the log-transformed frailty variance parameter was displayed in the bottom panels of Figure 2.2. As shown in the Figure, the `frailtypack` package had a number of extremely small estimates of the variance parameter of the frailty. In particular, this happens in scenarios with larger censoring rates with a smaller number of clusters.

Tables 2.4, 2.5, and 2.6 in the Appendix present detailed information about parameter estimates including the bias, mean and median of the standard error, empirical standard error and mean square error (MSE) of the estimated model parameters when the total sample size is 400 and the percentage of censoring rate is 20%, 50%, and 80%, respectively. The `survival` package does not provide the estimated standard error of the frailty variance. The empirical standard errors for the regression coefficients and variance parameter are defined based on their point estimates over simulated samples, which are calculated as: Empirical SE=$\sqrt{\frac{\sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2}{n_{sim}-1}}$, where n_{sim} is the number of successful fittings to the 1000 datasets, and θ donates the

true regression coefficient or variance parameter of the frailty. The results indicate that when the censoring rate increases, the frailty variance estimate has a smaller bias but larger variability. This finding is in line with previous research [36, 45]. The underestimation was even observed in the settings without censoring. The mean and median of the standard errors provided by all packages are very close to the empirical standard errors. Figure 2.3 displays the MSEs of all the parameters in the scenario of 40 clusters of size 10 (left panels) and 10 clusters of size 40 (right panels). The results of MSEs for all the estimated regression coefficients indicate that as the percentage of censoring increases, the MSEs of the estimated regression coefficients increased for the current R packages. However, the `frailtypack` had slightly larger MSEs for β_2 compared to other packages. The fourth row of Figure 2.3 shows the results of MSEs for frailty variance. The MSE of frailty variance increases for most R packages as the percentage of censoring increases, but the MSE of 20% censoring is larger than that 50 % censoring for `frailtySurv` package. Moreover, the MSE of frailty variance decreases as the percentage of censoring increases for `frailtypack` package in the 10 clusters of size 40.

2.4.2 Coverage Probability (CP)

For the currently available R packages considered in this paper, the 95% confidence intervals (CI) of the regression coefficients are calculated based on normal approximation, i.e., $\hat{\beta} \pm 1.96 * SE(\hat{\beta})$. Figure 2.4 presents the CP of the 95% CIs of the regression coefficients in the scenario of 40 clusters of size 10 (left panels) and 10 clusters of size 40 (right panels). The CPs of the 95% CIs for most of the R packages are very close to 95%, while `frailtypack` yielded slightly lower CP for β_2 and `frailtySurv` had slightly lower CP for β_3 in the case of 10 clusters of size 40. The detailed results are displayed in Table 2.7 in the Appendix.

For the frailty variance, most of the R packages calculate the CI based on normal approximation as: $\hat{\theta} \pm 1.96 * \text{SE}(\hat{\theta})$. We name this type of interval as $\text{CI}^{(1)}$. The left panels in Figure 2.5 clearly showed the CPs for $\text{CI}^{(1)}$ failed to attain the 95% nominal level. This is not surprising, since the distribution of the frailty variance is widely known for being skewed as shown by Figure 2.2; this was previously reported by [44]. Better CI may be constructed with the sampling distribution of the logarithm of the frailty variance, which is more symmetric. Then, the 95% CI for $\log \hat{\theta}$ can be constructed as

$$[\log \hat{\theta} - 1.96 \times \text{SE}(\log \hat{\theta}), \log \hat{\theta} + 1.96 \times \text{SE}(\log \hat{\theta})]. \tag{2.24}$$

The 95% CI for $\hat{\theta}$ can be then calculated by exponentiating the lower and upper boundaries of the 95% CI for $\log \hat{\theta}$ as,

$$\exp\Big[\log \hat{\theta} - 1.96 \times \text{SE}(\log \hat{\theta}), \log \hat{\theta} + 1.96 \times \text{SE}(\log \hat{\theta})\Big]. \tag{2.25}$$

We call this type of interval $\text{CI}^{(2)}$. Most R packages do not provide the value of $\text{SE}(\log \hat{\theta})$ directly. However, we can calculate it from the $\text{SE}(\hat{\theta})$ using the relationship of the Fisher information between θ and its log transformation $\phi = \log(\theta)$, which is derived briefly in general terms as follows. Let X be a random vector (data) with the probability density function $f(x|\theta)$. Let $I_1(\theta)$ denote the Fisher information of θ and $l_1(\theta; x)$ denote the log-likelihood of θ given x. Suppose we re-parameterize $\theta = \Theta(\phi)$, where $\Theta(\cdot)$ is a differentiable function. The log-likelihood function for ϕ, $l_2(\phi; x)$, is given by:

$$l_2(\phi; x) = l_1(\Theta(\phi); x) = \log f(x|\Theta(\phi)). \tag{2.26}$$

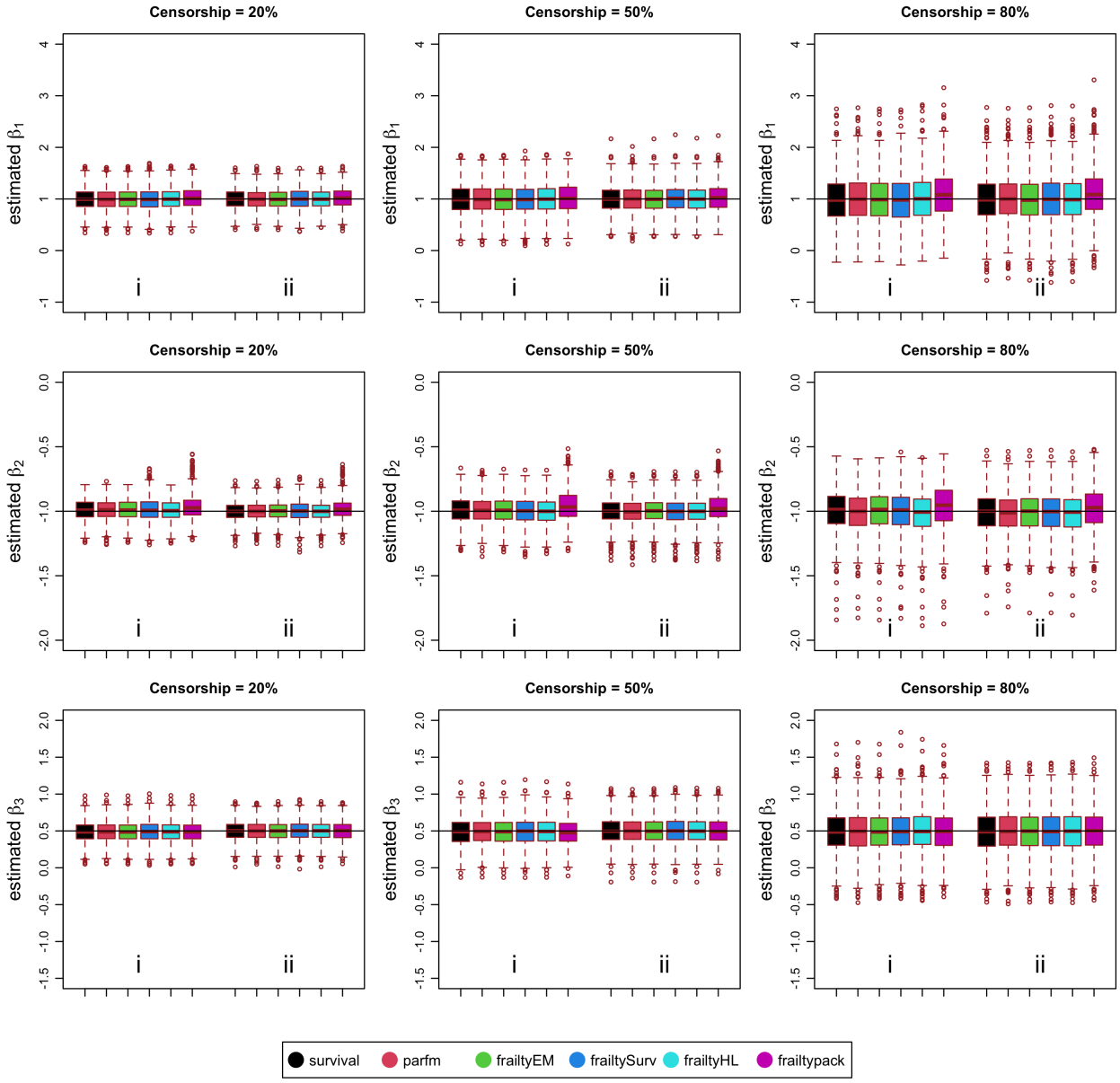


Figure 2.1: The estimated regression coefficients over 1000 samples simulated from the true model. True values of the regression coefficients are indicated as horizontal lines. The first, second and third columns correspond to 20%, 50% and 80% censoring rates, respectively. In each panel, the left half corresponds to scenario (i) with 40 clusters of size 10; the right half corresponds to scenario (ii) with 10 clusters of size 40.

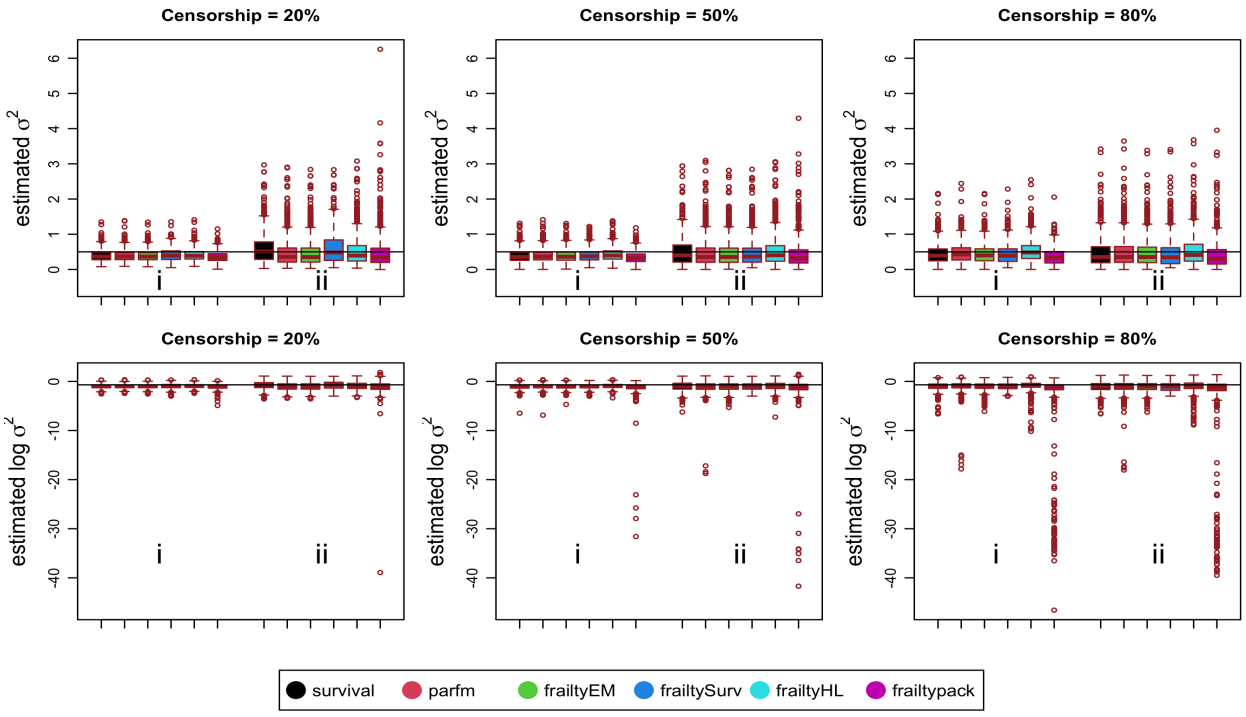


Figure 2.2: The estimated variance parameter of the random effect term over 1000 samples simulated from the true model. True values of the variance parameters are indicated as horizontal lines. The first, second and third columns correspond to 20%, 50% and 80% censoring rates, respectively. In each panel, the left half corresponds to scenario (i) with 40 clusters of size 10; the right half corresponds to scenario (ii) with 10 clusters of size 40. Note that some extreme values for the frailtyHL package have been removed.

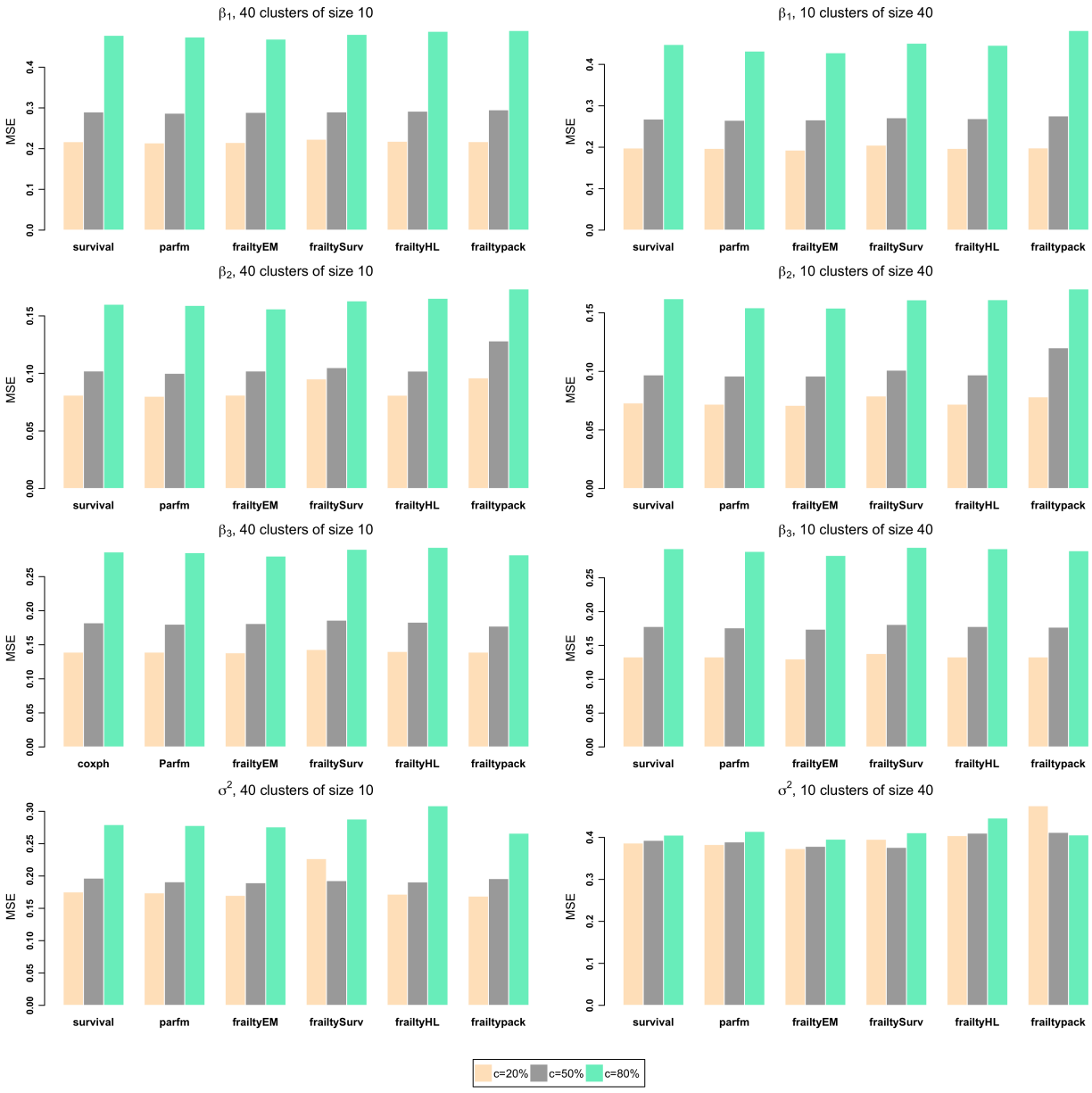


Figure 2.3: The MSE for each estimated regression coefficient and frailty variance. The first, second, third and fourth rows correspond to the results for β_1, β_2, β_3 and σ^2, respectively. The left panels correspond to the scenario with 40 clusters of size 10, and the right panels correspond to the scenario with 10 clusters of size 40. In each panel, the yellow, gray and green bars correspond to 20%, 50% and 80% censoring rates, respectively.

Then the derivative of l_2 is given by:

$$\frac{\partial l_2(\phi; x)}{\partial \phi} = \frac{\frac{\partial f(x|\theta)}{\partial \theta}\frac{\partial \Theta(\phi)}{\partial \phi}}{f(x|\Theta(\phi))}. \tag{2.27}$$

It follows that the fish information of ϕ, $I_2(\phi)$, is obtained as follows:

$$I_2(\phi) = E_X\left\{\left(\frac{\frac{\partial f(X|\theta)}{\partial \theta}}{f(X|\theta(\phi))}\right)^2\right\}\left(\frac{\partial \Theta(\phi)}{\partial \phi}\right)^2 = I_1(\theta)(\Theta'(\phi))^2, \tag{2.28}$$

where Θ' denotes the derivative function of Θ. Applying the above general rule to $\Theta(\phi) = \exp(\phi)$ (ie, $\phi = \log(\theta)$), we arrive at the following equation:

$$I_2(\phi) = I_1(\theta)\theta^2. \tag{2.29}$$

We know that $\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{I_1(\theta)}}$, where $\hat{\theta}$ is the maximum likelihood estimation (MLE) of θ. Finally, we arrive at the following relationship:

$$\mathrm{SE}(\log\left(\hat{\theta}\right)) = \frac{1}{\sqrt{I_2(\phi)}} = \frac{1}{\theta}\frac{1}{\sqrt{I_1(\theta)}} = \frac{1}{\theta}\mathrm{SE}(\hat{\theta}). \tag{2.30}$$

As shown in the second column of Figure 2.5, $\mathrm{CI}^{(2)}$ had consistently higher CP than $\mathrm{CI}^{(1)}$. Interestingly, as the censoring rate increases, CPs of both $\mathrm{CI}^{(1)}$ and $\mathrm{CI}^{(2)}$ for $\hat{\theta}$ became closer to 95%. This is partly due to the larger variability of the estimated parameters as a result of the higher censoring rate. This finding is consistent with the results of Balan et al. [46]. In addition, the currently available packages had lower CPs in the scenario with 10 clusters compared to the scenario with 40 clusters; this is presumably caused by that the shape of the sampling distribution of the frailty variance (or its log) being closer to normal when the number of clusters is larger.

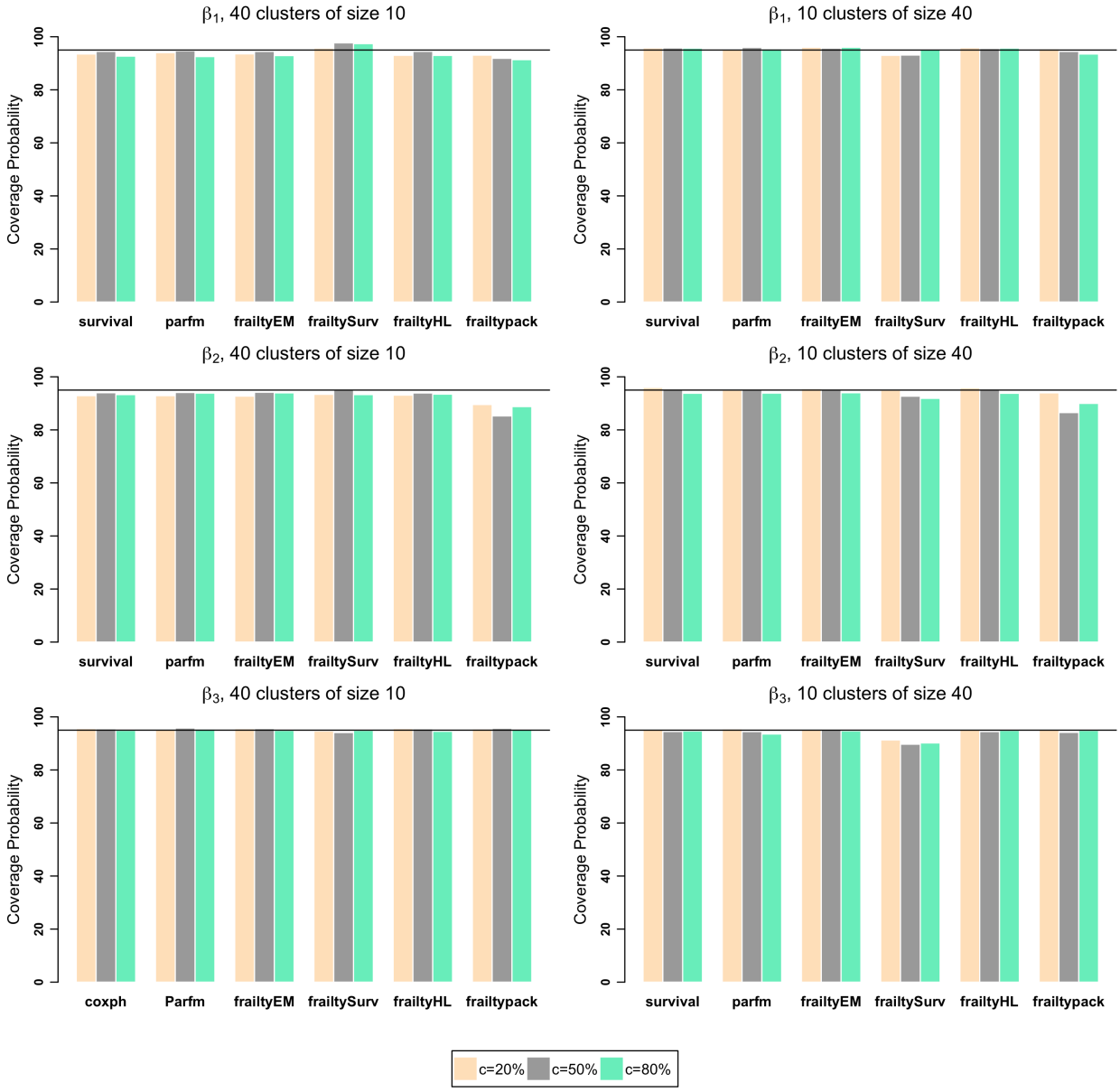


Figure 2.4: The coverage probability of the 95% confidence interval for each estimated regression coefficient. The black horizontal line indicates the 95% nominal level. The first, second and third rows correspond to the results for β_1, β_2 and β_3, respectively. The left panels correspond to the scenario with 40 clusters of size 10, and the right panels correspond to the scenario with 10 clusters of size 40. In each panel, the yellow, gray and green bars correspond to 20%, 50% and 80% censoring rates, respectively.

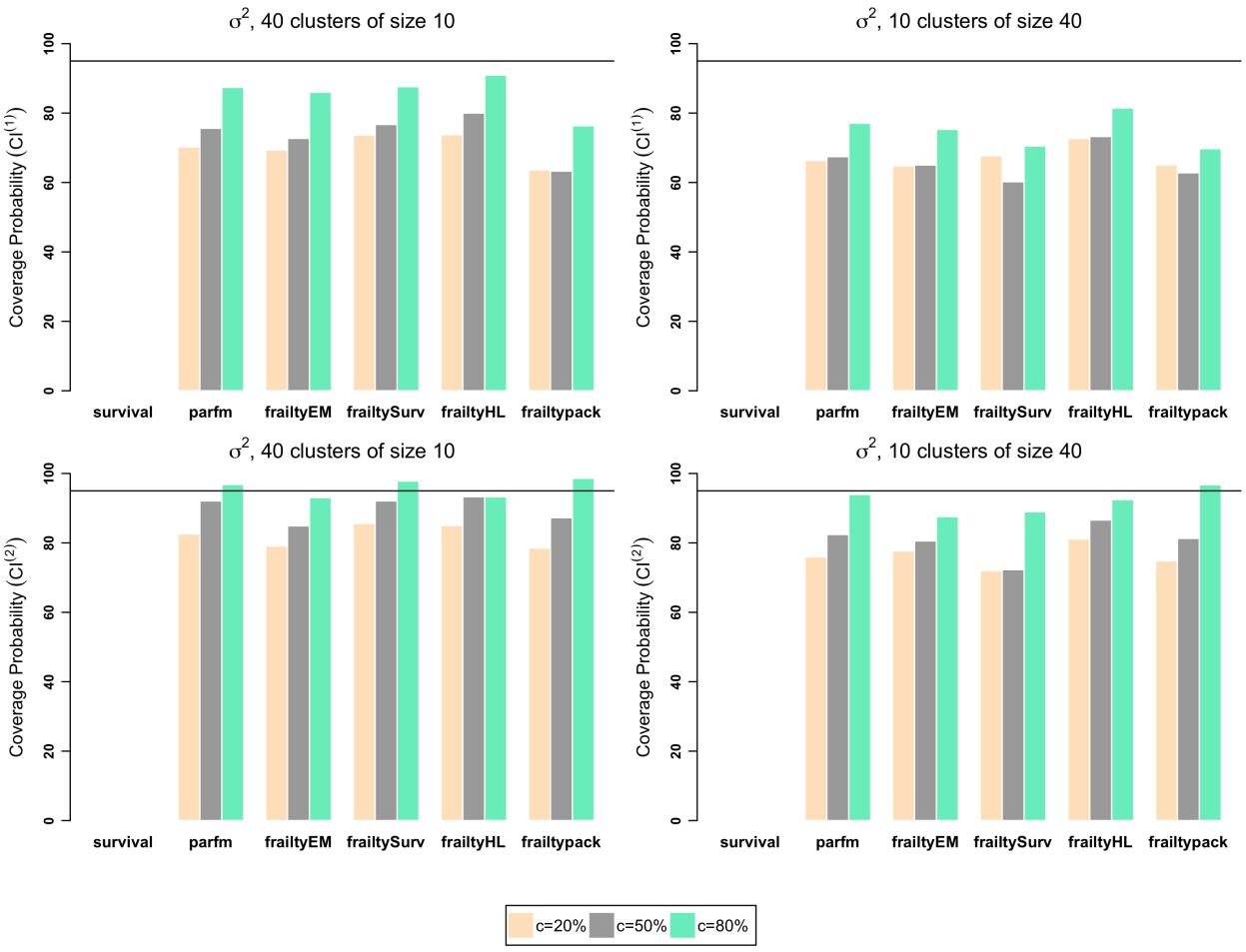


Figure 2.5: The coverage probability of the 95% confidence interval for the estimated frailty variance. The black horizontal line indicates the 95% nominal level. The first and second rows correspond to $CI^{(1)}$ under normal approximation and $CI^{(2)}$ under log transformation as described in section 4.2. The left panels correspond to scenarios with 40 clusters of size 10, the right panels correspond to scenarios with 10 clusters of size 40. In each panel, the yellow, gray and green bars correspond to 20%, 50% and 80% censoring rates, respectively.

2.4.3 Convergence rate

Table 2.2 presents the results of the convergence rate of each package. The `survival`, `frailtySurv`, and `frailtypack` packages had convergence rates over 97% in all scenarios. When the sample size is small with a large censoring rate, `frailtyHL`, `parfm` and `frailtyEM` packages had relatively lower convergence rates. In the scenario with an extremely low sample size of 100 at 80% censorship, the `parfm` and `frailtyEM` packages had the lowest convergence rate compared with other packages at about 66.3% and 57.2%, respectively.

Table 2.2: Convergence rate of the R packages over 1000 simulated datasets. Note that some very poorly fitted models are considered as not convergence.

n	Clusters	Obs	$100c$	survival	parfm	frailtyEM	frailtySurv	frailtyHL	frailtypack
100	10	10	20	100	96.3	93.3	99.8	97.9	99.4
400	40	10	20	100	100	99.1	100	100	99.6
400	10	40	20	100	99.9	96.2	100	99.2	99.3
800	80	10	20	100	100	99.4	100	100	99.8
800	10	80	20	100	100	94.7	100	98.7	99.5
100	10	10	50	99.9	**88.5**	**83.1**	99.2	97.1	100
400	40	10	50	100	100	99.6	100	100	99.7
400	10	40	50	100	99.5	97.5	100	99.2	99.4
800	80	10	50	100	100	99.5	100	100	100
800	10	80	50	100	100	95.8	100	98.7	98.5
100	10	10	80	97.2	**66.3**	**57.2**	98.1	**91.4**	99.1
400	40	10	80	100	97.7	95.8	100	99.5	99.7
400	10	40	80	100	96	93.2	99.7	98.1	99.8
800	80	10	80	100	99.9	99.3	100	99.9	99.8
800	10	80	80	100	99.5	97.8	100	99.2	99.8

2.4.4 Computing time

Table 2.3 reports the average computing time for fitting the shared frailty model using the R packages under each simulation scenario. To evaluate these 6 packages, we generated 1000 datasets under each scenario, which was a time-consuming task. Therefore, we submitted the job to Compute Canada Cedar to obtain results and calculate the average computing time. Cedar has a theoretical peak double precision performance of 14 petaflops with 74 islands, each with different configurations of nodes including CPUs, GPUs, and large memory nodes. Cluster Cedar was used in this study. Cedar is a heterogeneous cluster suitable for a variety of workloads; it is located at Simon Fraser University. It has a total of 94,528 CPU cores (Intel Xeon microprocessor) and 1352 GPU(NVIDIA P100 data Center accelerator) devices for computation, which provides a theoretical peak double precision performance of 14 petaflops. The package `survival` is the fastest one, followed by `frailtyEM` and `frailtypack`, and `parfm`, `frailtySurv` and `frailtyHL`. In general, the larger the number of clusters and cluster size requires more computing time for most packages, except for `frailtyEM` package.

Table 2.3: Average computing time (in minutes) of the R packages under each simulation scenario.

n	Clusters	Obs	$100c$	survival	parfm	frailtyEM	frailtySurv	frailtyHL	frailtypack
100	10	10	20	0.00022	0.05918	0.00810	0.00413	0.00771	0.00752
400	40	10	20	0.00050	0.17822	0.02198	0.24299	0.32823	0.01891
400	10	40	20	0.00033	0.10968	0.03585	0.16468	0.08406	0.01829
800	80	10	20	0.00098	0.29757	0.03849	2.74434	2.05797	0.03703
800	10	80	20	0.00053	0.26513	0.10853	2.07928	0.74127	0.03731
100	10	10	50	0.00023	0.06431	0.00647	0.00388	0.00922	0.00682
400	40	10	50	0.00033	0.12135	0.01271	0.23100	0.19286	0.01704
400	10	40	50	0.00033	0.10938	0.02084	0.16424	0.07566	0.01461
800	80	10	50	0.00099	0.41530	0.03447	3.84304	2.68267	0.04083
800	10	80	50	0.00039	0.16821	0.05750	2.09318	0.32363	0.02760
100	10	10	80	0.00019	0.05300	0.00463	0.00307	0.01072	0.00444
400	40	10	80	0.00035	0.15959	0.00797	0.22592	0.37690	0.01434
400	10	40	80	0.00029	0.10919	0.01019	0.16098	0.12193	0.01503
800	80	10	80	0.00054	0.27038	0.01256	2.70735	2.35211	0.02324
800	10	80	80	0.00034	0.15868	0.02353	2.07499	0.30798	0.02059

2.5 Discussions and Conclusions

In this paper, the currently available R packages with the default parameter settings considered for fitting the shared frailty models gave similar and unbiased parameter estimates for the fixed-effect regression coefficients, regardless of the cluster sizes and censoring rates. However, there were differences between the packages with respect to the estimation of the variance parameter for the frailty term. In general, the variance parameter of the frailty term was consistently underestimated for the currently available R packages considered in this paper. However, as the censoring rate increases, the bias is less pronounced but subject to more variability, which leads to higher MSE. This finding is consistent with the finding in other studies [36, 45]. Our results also showed that a larger number of clusters can lead to a higher precision of the estimated variance parameter of the frailty term. The CP of the 95% CIs of the regression coefficients for most of the R packages are very close to 95%, and the currently available packages of the frailty variance had lower CP in the scenario with a smaller number of clusters compared to the scenario with a larger number of clusters. Most packages had convergence rates over 97% in all scenarios, except for the `parfm` and `frailtyEM` packages in the scenario with a small sample size (n=100) and large censorship (80%). The computing time for all scenarios of `survival`, `frailtyEM` and `frailtypack` packages are within 0.1 minutes; the `parfm` takes no more than 0.5 minutes. However, the computing time for `frailtySurv` and `frailtyHL` packages need two to three minutes under the sample size n=800.

The best package to estimate the parameters of a frailty model is the `survival` package, which is computationally fast with a high convergence rate in almost all simulation scenarios. However, the `survival` package does not provide the estimate of standard error for the variance component of the frailty. Since the EM and PPL algorithms lead to the same estimates in a frailty model, `frailtyEM` can be used to substitute survival if the standard error of the frailty variance is required in a real application. However, we do not suggest using `frailtyEM` package when the sample size is small with a large censoring rate due to its lower convergence rate. The `parfm` has a lower convergence rate as well in the scenario with a small sample size at a large censoring rate. The parametric estimation is more powerful if the baseline hazard distribution is known, then the `parfm` is a good choice in the large sample size study. The `frailtySurv` fits the frailty model with a wide range of frailty distributions, and the `frailtyHL` allows multilevel frailties in the frailty model. However, the `parfm`, `frailtySurv` and `frailtyHL` packages require more computing time. The `frailtypack` package performs similarly to the other packages in terms of the estimation of regression coefficients and variance component of the random effect. However, `frailtypack` allows for more complex structures of the frailty terms, such as the nested and joined frailties and the frailty interaction.

In this study, a new type of confidence interval for the frailty variance θ, using the standard error of $\log\left(\hat{\theta}\right)$ was implemented. The coverage probability of the proposed confidence interval is much higher than the confidence interval based on the standard error of the frailty variance. Most packages do not provide the standard error of $\log\hat{\theta}$. Our proposed approach provides a solution by using the Fisher information approach. We recommend adding this approach to the R packages for calculating a more reliable 95% confidence interval for the frailty variance in frailty models.

Supplementary Materials

Table 2.4: Performance of the parameter estimation of different R packages. We only considered the converged fitted models for 1000 simulated datasets for each package. The total sample size is 400 and the censorship is 20%.

Parameter	True	Mean	Bias	Mean.se	Emp. se	Median	Median.se	MSE
40 clusters of size 10								
survival								
β_1	1.000	0.992	-0.008	0.217	0.217	0.993	0.216	0.217
β_2	-1.000	-0.990	0.010	0.076	0.081	-0.991	0.076	0.081
β_3	0.500	0.488	-0.012	0.142	0.139	0.486	0.142	0.139
σ^2	0.500	0.389	-0.111	-	0.163	0.366	-	0.175
parfm								
β_1	1.000	0.991	-0.009	0.215	0.214	0.990	0.214	0.214
β_2	-1.000	-0.989	0.011	0.076	0.080	-0.988	0.076	0.080
β_3	0.500	0.488	-0.012	0.141	0.139	0.487	0.141	0.139
σ^2	0.500	0.391	-0.109	0.127	0.162	0.367	0.121	0.174
frailtyEM								
β_1	1.000	0.991	-0.009	0.217	0.215	0.993	0.216	0.215
β_2	-1.000	-0.989	0.011	0.076	0.081	-0.990	0.076	0.081
β_3	0.500	0.488	-0.012	0.142	0.138	0.486	0.142	0.138
σ^2	0.500	0.386	-0.114	0.126	0.157	0.365	0.121	0.170
frailtySurv								
β_1	1.000	0.991	-0.009	0.249	0.223	0.990	0.244	0.223
β_2	-1.000	-0.986	0.014	0.098	0.095	-0.993	0.090	0.095
β_3	0.500	0.492	-0.008	0.144	0.143	0.490	0.141	0.143
σ^2	0.500	0.438	-0.062	0.408	0.223	0.392	0.133	0.227
frailtyHL								
β_1	1.000	0.996	-0.004	0.217	0.218	0.996	0.216	0.218
β_2	-1.000	-0.994	0.006	0.076	0.081	-0.995	0.076	0.081
β_3	0.500	0.490	-0.010	0.142	0.140	0.488	0.142	0.140
σ^2	0.500	0.411	-0.089	0.129	0.164	0.388	0.124	0.172
frailtypack								
β_1	1.000	1.014	0.014	0.215	0.217	1.013	0.216	0.217
β_2	-1.000	-0.968	0.032	0.076	0.095	-0.974	0.076	0.096
β_3	0.500	0.488	-0.012	0.141	0.139	0.486	0.141	0.139
σ^2	0.500	0.359	-0.141	0.119	0.149	0.343	0.116	0.169
10 clusters of size 40								
survival								
β_1	1.000	0.996	-0.004	0.208	0.198	0.995	0.208	0.198
β_2	-1.000	-1.004	-0.004	0.074	0.073	-1.001	0.074	0.073
β_3	0.500	0.501	0.001	0.136	0.133	0.503	0.136	0.133
σ^2	0.500	0.568	0.068	-	0.382	0.510	-	0.387
parfm								
β_1	1.000	0.993	-0.007	0.205	0.197	0.995	0.204	0.197
β_2	-1.000	-1.002	-0.002	0.072	0.072	-1.001	0.072	0.072
β_3	0.500	0.500	0.000	0.134	0.133	0.499	0.134	0.133
σ^2	0.500	0.468	-0.032	0.211	0.382	0.357	0.169	0.383
frailtyEM								
β_1	1.000	0.991	-0.009	0.207	0.193	0.990	0.207	0.193
β_2	-1.000	-0.999	0.001	0.074	0.071	-0.997	0.074	0.071
β_3	0.500	0.500	0.000	0.135	0.130	0.502	0.135	0.130
σ^2	0.500	0.462	-0.038	0.206	0.372	0.346	0.164	0.373
frailtySurv								
β_1	1.000	1.000	0.000	0.209	0.205	0.998	0.207	0.205
β_2	-1.000	-0.999	0.001	0.086	0.079	-0.998	0.084	0.079
β_3	0.500	0.503	0.003	0.129	0.138	0.502	0.127	0.138
σ^2	0.500	0.567	0.067	0.308	0.391	0.481	0.171	0.395
frailtyHL								
β_1	1.000	0.995	-0.005	0.207	0.197	0.994	0.207	0.197
β_2	-1.000	-1.002	-0.002	0.074	0.072	-1.000	0.074	0.072
β_3	0.500	0.501	0.001	0.135	0.133	0.502	0.135	0.133
σ^2	0.500	0.516	0.016	0.237	0.404	0.394	0.192	0.404
frailtypack								
β_1	1.000	1.016	0.016	0.206	0.198	1.017	0.206	0.198
β_2	-1.000	-0.984	0.016	0.073	0.078	-0.985	0.073	0.078
β_3	0.500	0.500	0.000	0.135	0.133	0.503	0.134	0.133
σ^2	0.500	0.480	-0.020	0.227	0.475	0.343	0.167	0.475

Table 2.5: Performance of the parameter estimation of different R packages. We only considered the converged fitted models for 1000 simulated datasets for each package. The total sample size is 400 and the censorship is 50%.

Parameter	True	Mean	Bias	Mean.se	Emp. se	Median	Median.se	MSE
40 clusters of size 10								
survival								
β_1	1	0.993	-0.007	0.283	0.29	0.984	0.283	0.290
β_2	-1	-0.99	0.01	0.097	0.102	-0.992	0.096	0.102
β_3	0.5	0.486	-0.014	0.183	0.182	0.489	0.182	0.182
σ^2	0.5	0.383	-0.117	-	0.183	0.356	-	0.197
parfm								
β_1	1	0.997	-0.003	0.28	0.287	0.989	0.279	0.287
β_2	-1	-0.993	0.007	0.097	0.1	-0.994	0.097	0.100
β_3	0.5	0.487	-0.013	0.182	0.18	0.492	0.181	0.180
σ^2	0.5	0.395	-0.105	0.147	0.18	0.364	0.141	0.191
frailtyEM								
β_1	1	0.994	-0.006	0.283	0.289	0.985	0.282	0.289
β_2	-1	-0.991	0.009	0.097	0.102	-0.992	0.096	0.102
β_3	0.5	0.487	-0.013	0.183	0.181	0.49	0.182	0.181
σ^2	0.5	0.388	-0.112	0.147	0.177	0.361	0.141	0.190
frailtySurv								
β_1	1	0.990	-0.01	0.337	0.29	0.988	0.336	0.290
β_2	-1	-0.997	0.003	0.11	0.105	-0.997	0.108	0.105
β_3	0.5	0.493	-0.007	0.184	0.186	0.497	0.181	0.186
σ^2	0.5	0.396	-0.104	0.164	0.182	0.374	0.149	0.193
frailtyHL								
β_1	1	1.002	0.002	0.284	0.292	0.996	0.284	0.292
β_2	-1	-0.999	0.001	0.097	0.102	-1	0.097	0.102
β_3	0.5	0.49	-0.01	0.184	0.183	0.493	0.183	0.183
σ^2	0.5	0.424	-0.076	0.152	0.185	0.395	0.145	0.191
frailtypack								
β_1	1	1.018	0.018	0.274	0.295	1.01	0.279	0.295
β_2	-1	-0.953	0.047	0.096	0.126	-0.967	0.097	0.128
β_3	0.5	0.481	-0.019	0.181	0.177	0.476	0.181	0.177
σ^2	0.5	0.342	-0.158	0.133	0.171	0.322	0.131	0.195
10 clusters of size 40								
survival								
β_1	1.000	0.999	-0.001	0.280	0.268	0.995	0.278	0.268
β_2	-1.000	-1.001	-0.001	0.097	0.097	-0.999	0.097	0.097
β_3	0.500	0.503	0.003	0.180	0.178	0.496	0.179	0.178
σ^2	0.500	0.499	-0.001	-	0.393	0.400	-	0.393
parfm								
β_1	1.000	1.001	0.001	0.275	0.265	1.001	0.274	0.265
β_2	-1.000	-1.004	-0.004	0.095	0.096	-1.005	0.094	0.096
β_3	0.500	0.505	0.005	0.177	0.176	0.497	0.177	0.176
σ^2	0.500	0.465	-0.035	0.221	0.388	0.353	0.180	0.389
frailtyEM								
β_1	1.000	0.999	-0.001	0.279	0.266	0.993	0.278	0.266
β_2	-1.000	-0.999	0.001	0.097	0.096	-0.997	0.097	0.096
β_3	0.500	0.503	0.003	0.179	0.174	0.496	0.178	0.174
σ^2	0.500	0.458	-0.042	0.218	0.377	0.350	0.179	0.379
frailtySurv								
β_1	1.000	1.009	0.009	0.290	0.271	1.010	0.286	0.271
β_2	-1.000	-1.005	-0.005	0.103	0.101	-1.001	0.101	0.101
β_3	0.500	0.506	0.006	0.166	0.181	0.500	0.164	0.181
σ^2	0.500	0.465	-0.035	0.226	0.375	0.365	0.151	0.376
frailtyHL								
β_1	1.000	1.001	0.001	0.279	0.269	0.999	0.278	0.269
β_2	-1.000	-1.003	-0.003	0.097	0.097	-1.000	0.097	0.097
β_3	0.500	0.504	0.004	0.179	0.178	0.497	0.178	0.178
σ^2	0.500	0.513	0.013	0.248	0.410	0.402	0.207	0.410
frailtypack								
β_1	1.000	1.023	0.023	0.272	0.275	1.023	0.274	0.276
β_2	-1.000	-0.966	0.034	0.095	0.119	-0.980	0.095	0.120
β_3	0.500	0.499	-0.001	0.178	0.177	0.491	0.177	0.177
σ^2	0.500	0.437	-0.063	0.218	0.408	0.320	0.170	0.412

Table 2.6: Performance of the parameter estimation of different R packages. We only considered the converged fitted models for 1000 simulated datasets for each package. The total sample size is 400 and the censorship is 80%.

Parameter	True	Mean	Bias	Mean.se	Emp. se	Median	Median.se	MSE
40 clusters of size 10								
survival								
β_1	1	0.985	-0.015	0.446	0.478	0.977	0.442	0.478
β_2	-1	-0.993	0.007	0.146	0.16	-0.986	0.145	0.160
β_3	0.5	0.484	-0.016	0.283	0.286	0.485	0.281	0.286
σ^2	0.5	0.441	-0.059	-	0.276	0.391	-	0.279
parfm								
β_1	1	1.006	0.006	0.442	0.474	0.995	0.439	0.474
β_2	-1	-1.006	-0.006	0.15	0.159	-1.001	0.148	0.159
β_3	0.5	0.487	-0.013	0.284	0.285	0.494	0.282	0.285
σ^2	0.5	0.467	-0.033	0.243	0.277	0.428	0.231	0.278
frailtyEM								
β_1	1	0.994	-0.006	0.447	0.469	0.989	0.443	0.469
β_2	-1	-0.995	0.005	0.147	0.156	-0.986	0.146	0.156
β_3	0.5	0.484	-0.016	0.283	0.280	0.487	0.281	0.280
σ^2	0.5	0.446	-0.054	0.244	0.273	0.399	0.232	0.276
frailtySurv								
β_1	1	0.976	-0.024	0.572	0.48	0.98	0.557	0.481
β_2	-1	-0.998	0.002	0.161	0.163	-0.99	0.156	0.163
β_3	0.5	0.488	-0.012	0.292	0.29	0.49	0.285	0.290
σ^2	0.5	0.436	-0.064	0.284	0.284	0.391	0.26	0.288
frailtyHL								
β_1	1	1.008	0.008	0.452	0.488	1.006	0.448	0.488
β_2	-1	-1.015	-0.015	0.149	0.165	-1.007	0.148	0.165
β_3	0.5	0.495	-0.005	0.286	0.293	0.492	0.284	0.293
σ^2	0.5	0.525	0.025	0.261	0.308	0.481	0.251	0.309
frailtypack								
β_1	1	1.084	0.084	0.425	0.483	1.081	0.435	0.490
β_2	-1	-0.963	0.037	0.148	0.172	-0.953	0.147	0.173
β_3	0.5	0.492	-0.008	0.279	0.282	0.493	0.278	0.282
σ^2	0.5	0.369	-0.131	0.21	0.249	0.339	0.21	0.266
10 clusters of size 40								
survival								
β_1	1.000	0.997	-0.003	0.457	0.448	0.981	0.452	0.448
β_2	-1.000	-1.008	-0.008	0.152	0.162	-0.999	0.150	0.162
β_3	0.500	0.491	-0.009	0.288	0.293	0.488	0.285	0.293
σ^2	0.500	0.479	-0.021	-	0.405	0.353	-	0.405
parfm								
β_1	1.000	1.008	0.008	0.446	0.432	1.000	0.443	0.432
β_2	-1.000	-1.018	-0.018	0.149	0.154	-1.012	0.147	0.154
β_3	0.500	0.498	-0.002	0.284	0.289	0.492	0.281	0.289
σ^2	0.500	0.483	-0.017	0.271	0.414	0.353	0.227	0.414
frailtyEM								
β_1	1.000	0.998	-0.002	0.455	0.428	0.980	0.450	0.428
β_2	-1.000	-1.008	-0.008	0.151	0.154	-1.000	0.149	0.154
β_3	0.500	0.493	-0.007	0.286	0.283	0.497	0.284	0.283
σ^2	0.500	0.472	-0.028	0.269	0.395	0.351	0.228	0.396
frailtySurv								
β_1	1.000	1.008	0.008	0.554	0.451	0.992	0.526	0.451
β_2	-1.000	-1.010	-0.010	0.158	0.161	-1.001	0.151	0.161
β_3	0.500	0.497	-0.003	0.275	0.295	0.494	0.267	0.295
σ^2	0.500	0.454	-0.046	0.272	0.409	0.345	0.206	0.411
frailtyHL								
β_1	1.000	1.006	0.006	0.457	0.446	0.985	0.453	0.446
β_2	-1.000	-1.015	-0.015	0.152	0.161	-1.007	0.150	0.161
β_3	0.500	0.495	-0.005	0.288	0.293	0.497	0.285	0.293
σ^2	0.500	0.534	0.034	0.299	0.445	0.407	0.257	0.446
frailtypack								
β_1	1.000	1.108	0.108	0.438	0.470	1.084	0.440	0.481
β_2	-1.000	-0.980	0.020	0.150	0.170	-0.972	0.149	0.170
β_3	0.500	0.502	0.002	0.284	0.290	0.503	0.281	0.290
σ^2	0.500	0.411	-0.089	0.244	0.398	0.304	0.207	0.406

Table 2.7: Coverage probability of the 95% CI for the estimated regression coefficients and frailty variance.

Parameter	survival	parfm	frailtyEM	frailtySurv	frailtyHL	frailtypack
c= 20, cluster size = 10, 40 clusters						
β_1	93.5	94	93.54	95.7	93	93.72
β_2	92.9	92.9	92.73	93.4	93.1	89.56
β_3	95.6	95.3	95.56	94.7	95.6	95.48
$CI^{(1)}(\sigma^2)$	-	70.2	69.32	73.7	73.8	63.65
$CI^{(2)}(\sigma^2)$	-	82.6	79.1	85.6	85	78.51
c = 20, cluster size = 40, 10 clusters						
β_1	95.8	95.30	96.05	93	95.87	95.47
β_2	96	94.99	95.53	95	95.87	93.96
β_3	95.3	95.49	95.32	91.3	95.26	95.17
$CI^{(1)}(\sigma^2)$	-	66.37	64.76	67.7	72.68	65.06
$CI^{(2)}(\sigma^2)$	-	75.98	77.68	72	81.15	74.82
c= 50, cluster size = 10, 40 clusters						
β_1	94.5	94.7	94.48	97.7	94.5	91.88
β_2	94	94.1	94.18	95.3	93.9	85.26
β_3	95.5	95.8	95.58	94.1	95.5	95.69
$CI^{(1)}(\sigma^2)$	-	75.6	72.69	76.7	80	63.29
$CI^{(2)}(\sigma^2)$	-	92.1	84.9	92.1	93.3	87.26
c = 50, cluster size = 40, 10 clusters						
β_1	95.8	95.98	95.96	93.1	95.57	94.47
β_2	95.3	95.07	95.18	92.7	95.37	86.52
β_3	94.5	94.47	95.18	89.7	94.47	94.16
$CI^{(1)}(\sigma^2)$	-	67.44	65.03	60.2	73.24	62.78
$CI^{(2)}(\sigma^2)$	-	82.41	80.65	72.3	86.59	81.29
c= 80, cluster size = 10, 40 clusters						
β_1	92.7	92.53	92.90	97.4	92.96	91.37
β_2	93.3	93.86	93.95	93.3	93.47	88.77
β_3	94.9	95.39	94.89	94.9	94.57	95.29
$CI^{(1)}(\sigma^2)$	-	87.41	86.01	87.6	90.95	76.33
$CI^{(2)}(\sigma^2)$	-	96.83	93.04	97.8	93.25	98.60
c = 80, cluster size = 40, 10 clusters						
β_1	95.7	95.42	96.03	95.19	95.72	93.49
β_2	93.8	93.85	93.99	91.88	93.78	89.98
β_3	94.7	93.54	94.74	90.17	94.90	94.89
$CI^{(1)}(\sigma^2)$	-	77.08	75.32	70.51	81.45	68.74
$CI^{(2)}(\sigma^2)$	-	93.96	87.58	89	92.47	96.79

3 Z-residual Diagnostics for Detecting Misspecification of the Functional Form of Covariates for Shared Frailty Models[1].

Abstract: In survival analysis, the hazard function often depends on a set of covariates. Martingale and deviance residual are most widely used for examining the validity of the function form of covariates by checking whether there is a discernible trend in their scatterplot against continuous covariates. However, visual inspection of martingale and deviance residuals is often subjective. In addition, these residuals lack a reference distribution due to censoring. It is therefore challenging to derive numerical statistical tests based on martingale or deviance residuals. In this paper, we extend the idea of randomized survival probability (Li et al. 2021) and develop a residual diagnostic tool that can provide both graphical and numerical tests for checking the covariate functional form in semi-parametric shared frailty models. We develop a general function that calculates Z-residuals for semi-parametric shared frailty models based on the output from the `coxph` function in the `survival` package in R. Our extensive simulation studies indicate that the derived numerical test based on Z-residuals has great power for checking the functional form of covariates. In a real data application on modelling the survival time of acute myeloid leukemia patients, the Z-residual diagnosis results show that a model with log-transformation is inappropriate for modelling the survival time, which could not be detected by other diagnostic methods.

3.1 Introduction

Survival data with a multilevel structure occur frequently in many applications. For example, patients are often clustered within hospitals. The hazard of events differs from one cluster to another cluster induced by unobserved cluster-level factors. In survival analysis, conventional Cox proportional hazard models [22] and accelerated failure time models [23] assume that subjects are independent. Random effects can be incorporated into conventional survival models to account for cluster-level heterogeneity. Such heterogeneity is often called frailty in the context of survival analysis. A shared frailty model extends the classic survival models by incorporating random effects (frailties) acting multiplicatively on the baseline hazard function [24], where the frailties are common or shared among individuals within a cluster or group [1–4]. Despite the increasing popularity of shared frailty models for modelling clustered survival data, examining model assumptions is often overlooked partly due to the limited model diagnostic tools.

Residual diagnostics are often used to assess the overall goodness of fit (GOF) and to identify specific model misspecification (e.g., functional form of covariate effects). However, in the presence of censored observations, residual diagnostics is not as straightforward as a normal linear regression model. Cox-Snell (CS) residual [12] is the most widely used tool for diagnosing survival models, which are defined as the negative

[1]This chapter has been deposited as an arXiv paper:2302.09106

logarithm of estimated survival probability. In the absence of censored observations, the survival probability is uniformly distributed when the model is true; therefore, the CS residual is exponentially distributed. However, in the presence of censored observations, CS residuals are no longer exponentially distributed since the survival probability is not uniformly distributed. To account for censored observations, diagnostics based on CS residuals compares the agreement of cumulative hazard plot of CS residuals estimated with Kaplan-Meier method [18] and the 45° straight line, which is the cumulative hazard of the standard exponential distribution.

Although the overall GOF checking such as the cumulative hazard plot of CS residuals is widely used for diagnosing survival models, the overall GOF test reveals little information about the nature of the model inadequacies. Tailored graphical and numeric diagnostic tools are therefore needed. A number of residuals diagnostics tools have been proposed [11] for checking the functional form of covariates, of which martingale [13] and deviance [14, 15] residuals are most widely used. Martingale residuals can be viewed as the difference between the observed value of a subject's failure indicator and its expected value, integrated over the time for which that patient was at risk, which can be used to assess the functional form covariates and identify outliers in the survival data. Deviance residuals are a normalized transform of the martingale residuals. They also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Although these two types of residuals are widely used and available in the `survival` package in R software, each of these traditional types of residuals has limitations. Martingale residuals are asymmetric, with the upper bound of martingale residuals being one and no lower bound, making it difficult for visual inspection. Deviance residuals are less skewed and more normally distributed. The locally weighted scatterplot smoothing (LOWESS) lines on the scatterplots of the residuals against the continuous covariates are useful for revealing patterns in the residuals that would not otherwise be perceived. However, visual inspection of LOWESS lines can be still subjective. It is desirable to have a numerical measure of the statistical significance of the observed trend. However, martingale and deviance residuals lack a reference distribution due to censoring. It is therefore challenging to derive a numerical test to measure the statistical significance of the observed pattern in the residual plots.

Li et al. [19] proposed to use randomized survival probabilities (RSPs) to define residuals for checking the model assumptions of accelerated failure time (AFT) models without random effects. The key idea of RSP is to replace the survival probability of a censored failure time with a uniform random number between 0 and the survival probability of the censored time. The RSPs are uniformly distributed under the true model, hence, can then be transformed into normally distributed residuals with the normal quantile function. The new residual was called the normally-transformed RSP (NRSP) residual. Provided with the normally distributed reference distribution for the NRSP residual, statistical tests can be derived based on NRSP residuals for checking model assumptions, such as distributional assumption, functional form of covariates, etc. However, NRSP residuals have not been extended to diagnose Cox proportional hazard models or semi-parametric shared frailty models.

In this study, we extend the idea of NRSP residuals to develop residual diagnostics tools for checking the functional form of the covariates in semi-parametric shared frailty models. We rename NRSP residuals as Z-residuals for simplicity, as Z is often used to denote a standard normal random variable. For calculating the Z-residuals, we treat the random effects as fixed effects; that is, our Z-residual is conditional on the group identities. We developed a general function for calculating such conditional Z-residuals given the output of `coxph` in the `survival` package in R and proposed a non-homogeneity test for testing whether there is a trend in Z-residuals. We conducted extensive simulation studies to investigate the performance of the Z-residuals

diagnostics tool in detecting misspecification of the functional form of covariates. Our results showed that the non-homogeneity test based on Z-residuals has greater power and satisfactory type I error compared to the overall GOF tests in detecting misspecification of the covariate functional form. We also demonstrated the effectiveness of Z-residuals in diagnosing the functional form of covariates in real data analysis of mortality risk of acute myeloid leukemia patients [47, 48]. Our proposed Z-residual diagnostic tool discovered that a model with log transformation of a continuous covariate is inappropriate in this real data application, which however can not be captured by other diagnostic methods.

The rest of this paper is organized as follows. Section 3.2 gives a brief review of semi-parametric shared frailty models. In Section 3.3 we review the conventional residuals and model diagnostics methods for shared frailty models. In Section 3.4 we present the definition of Z-residuals and the non-homogeneity test based on Z-residuals. In Section 3.5, we conduct simulation studies to investigate the performances of the Z-residual diagnostics tool. Section 3.6 presents the results of applying the Z-residual diagnostics tool for diagnosing the functional form of covariates in a real data application. The article is concluded in Section3.7.

3.2 Shared Frailty Model and Statistical Inference

3.2.1 Notation and Shared Frailty Model

A shared frailty model is a frailty model where the frailties are common or shared among individuals within groups. The formulation of a frailty model for clustered failure survival data is defined as follows. Suppose there are g groups of individuals with n_i individuals in the ith group, $i = 1, 2, \ldots, g$. If the number of subjects n_i is 1 for all groups, then the univariate frailty model is obtained [3]. Otherwise, the model is called the shared frailty model [2, 20, 21] because all subjects in the same cluster share the same frailty value z_i. Suppose t_{ij} is the true failure time for the jth individual from the ith group, which we assume to be a continuous random variable in this article, where $j = 1, 2, \cdots, n_i$. Let t^*_{ij} denote the realization of t_{ij}. In the scenario of right censoring, we can observe that t_{ij} is greater than a value c_{ij}, where c_{ij} is the corresponding censoring time. The observed failure times are denoted by the pair (y_{ij}, δ_{ij}), where $y_{ij} = \min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} < c_{ij})$. The observed data can be written as $y = (y_{11}, \cdots, y_{gn_g})$ and $\delta = (\delta_{11}, \cdots, \delta_{gn_g})$. Since we will consider only the right-censoring in this article, we will use "censoring" as a short for "right-censoring". The survival function of t_{ij} based on a postulated model is defined as $S_{ij}(t^*_{ij}) = P(t_{ij} > t^*_{ij})$, where the subscript ij indicates that the probability depends on covariate x_{ij} for the jth individual of the ith group.

For a shared frailty model, the hazard of an event at time t for the jth individual, $j = 1, 2, \cdots, n_i$, in the ith group, is then

$$h_{ij}(t) = z_i \exp(x_{ij}\beta) h_0(t); \tag{3.1}$$

and the survival function for the jth individual of the ith group at time t follows:

$$S_{ij}(t) = \exp\left\{ -\int_0^t h_{ij}(t)\,\mathrm{d}t \right\} = \exp\left\{ - z_i \exp(x_{ij}\beta) H_0(t) \right\}, \tag{3.2}$$

where x_{ij} is a row vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \cdots, x_{gn_g})$; β is the column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, $H_0(t)$ is the baseline cumulative hazard function (CHF), and z_i is the frailty term that is common for all n_i individuals within the ith group. Let $z = (z_1, \cdots, z_g)$. The hazard and survival functions with frailty can

also be written as,

$$h_{ij}(t) = \exp(x_{ij}\beta + u_i)h_0(t), \tag{3.3}$$

and

$$S_{ij}(t) = \exp\Big\{ - \exp(x_{ij}\beta + u_i)H_0(t)\Big\}, \tag{3.4}$$

where u_i= $\log(z_i)$ is a random effect in the linear component of the proportional hazards model. Note that z_i cannot be negative, but u_i can be any value. If u_i is zero, corresponding to z_i is one, the model does not have frailty. The form of the baseline hazard function may be assumed to be unspecified as a semi-parametric model or fully specified to follow a parametric distribution.

In our study, we focus mainly on the shared gamma frailty model, since gamma distribution is one of the most common distributions for modelling the frailty effect [11]. It is easy to obtain a closed-form representation of the observable survival, cumulative density, and hazard functions due to the simplicity of the Laplace transform [49]. The gamma distribution is a two-parameter distribution with a shape parameter k and scale parameter θ. It takes a variety of shapes as k varies: when $k = 1$, it is identical to the well-known exponential distribution; when k is large, it takes a bell-shaped form reminiscent of a normal distribution; when k is less than one, it takes exponentially shaped and asymptotic to both the vertical and horizontal axes. Under the assumption $k = \frac{1}{\theta}$, the two-parameter gamma distribution turns into a one-parameter distribution. The expected value is one and the variance is equal to θ.

3.2.2 Parameter Estimation and Inference

Arguably the most popular R package for fitting semi-parametric shared frailty models is the `survival` package [5]. The `coxph` function of the `survival` R package can be used to fit semi-parametric shared frailty models via penalized partial likelihood method [25, 26, 40]; and the Breslow (1972) estimator [37] is used for estimating the baseline CHF. The frailty distribution can be specified as Gamma, Gaussian, or t distribution. It accommodates the clustered failures and recurrent events data with the right, left, and interval censoring types. When the `coxph` function fits the shared frailty model with clustered failures data, the cluster size should be above five. Otherwise, the random effects will be treated as fixed effects. The `survival` R package is used for estimating parameters and inference in this study.

In Cox proportional hazards regression, the Breslow estimator [37] is the nonparametric maximum likelihood estimation for the baseline CHF. The baseline CHF is $H_0(t) = \int_0^t h_0(s)\, ds$. Breslow (1972) suggested estimating the baseline CHF via maximizing the likelihood function. After getting the estimators $\hat{\beta}$ and $\hat{u}_i$, nonparametric maximum likelihood estimator of $\hat{H}_0(t)$ can be derived as:

$$\hat{H}_0(t) = \sum_{\{v:y_{(v)}\leq t\}} \left\{ \frac{d_{(v)}}{\sum\limits_{(i,j)\in R_{(y_{(v)})}} \exp\Big(x_{ij}\hat{\beta} + \hat{u}_i\Big)} \right\}, \tag{3.5}$$

where $y_{(1)} < \cdots < y_{(r)}$ are the ordered distinct event time among the y_{ij}'s and $R(y_{(v)}) = \{(i,j) : y_{ij} \geq y_{(v)}\}$ is the risk set at $y_{(v)}$, i.e., $d_{(v)}$ is the number of events at $y_{(v)}$. The Breslow approximation is the first option to estimate the baseline hazard function in nearly all the R packages for fitting Cox regression models with or without frailties.

The penalized partial likelihood (PPL) approach can be used to estimate parameters in a shared frailty model [2, 40]. The full data log-likelihood contains the frailty terms z, which are assumed to be observed

random variables first. The full data log-likelihood follows the joint density of (y, δ) and z, which can be split into two parts. The first part is the conditional likelihood of the data given the frailties, which takes the random effects $u = \log(z)$ as another set of the parameter in the first part of the likelihood. The second part is the log-likelihood of the random effects. Since the full likelihood is only used to estimate the p components of β and the g components of u, the terms involving θ alone can be omitted to give the penalized. The second part corresponds to the frailties distribution in which the likelihood is considered a penalty term. The estimation is based on maximizing the penalized partial log-likelihood (PPL) for the frailty model, which is given by

$$l_{ppl}(\beta, u, \theta; y, \delta) = l_{part}(\beta, u; y, \delta) + l_{pen}(\theta; u), \tag{3.6}$$

over both β and u. Here $l_{part}(\beta, u)$ is the partial log-likelihood for the Cox model that includes the random effects.

$$l_{part}(\beta, u; y, \delta) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \delta_{ij} \left\{ \eta_{ij} - \log \left[\sum_{(q,w) \in R(y_{ij})} \exp(\eta_{qw}) \right] \right\}, \tag{3.7}$$

where $\eta_{ij} = x_{ij}\beta + u_i$ and $\eta = (\eta_{11}, \ldots, \eta_{gn_g})$. In the penalty function $l_{pen}(\theta; u)$, θ is the parameter for the frailty. The random effect u is equal to $\log(z)$, where z is usually assumed to have a gamma distribution. The penalty function can be written as,

$$l_{pen}(\theta; u) = \sum_{i=1}^{g} \log f_U(u_i|\theta), \tag{3.8}$$

where $f_U(u_i)$ denotes the density function of the random effect u_i.

The maximization of the PPL consists of an inner and an outer loop [2]. For the gamma frailty effects with unit mean and variance θ, the penalized likelihood can be maximized with the Newton-Raphson algorithm in the inner loop. The estimates of β's and the u's are first taken to be values that maximize $l_{ppl}(\beta, u, \theta)$ for a given value of the θ. The outer loop is based on the maximization of a profiled version of the marginal likelihood for θ given estimates $\hat{\beta}$ and $\hat{u}$. The process is iterated until convergence.

3.3 Review of Existing Residuals and Test Methods

In this section, we review some existing residuals used in survival analysis. A central concept in these residuals is formulated based on the survival probability (SP). The widely used CS residual is defined as $r_{ij}^c(t_{ij}) = -\log(S_{ij}(t_{ij}))$, where t_{ij} is the true failure time. In the absence of censored observations, the survival probability is uniformly distributed when the model is true; therefore, the CS residual is exponentially distributed. A plot of the CHF against the true failure time will give a straight line through the origin with a unit slope when the residuals have a unit exponential distribution, which is expected when the survival model is correctly specified. In addition to the graphical checking, we can apply numerical GOF testing methods such as Kolmogorov-Smirnov (KS) test to CS residuals. When there are censored failure times, the distribution of $S_{ij}(y_{ij})$ is no longer uniformly distributed under the true model, which means the CS residuals are no longer exponentially distributed. The CS residuals r_{ij}^c can be regarded as a dataset with censoring. The Kaplan-Meier (KM) estimate of the survivor function can still be computed for CS residuals. Hence, the most widely used diagnostics tool is to apply the KM method to get an estimate of the CHF of CS residuals and compare the CHF against the 45^o straight line.

Transforming SPs into exponentially-distributed CS residuals is only one option among many others.

For example, one can also transform SPs using the quantile of standard normal distribution [50], defined as $r_{ij}^n(y_{ij}) = -\Phi^{-1}(S_{ij}(y_{ij}))$, where y_{ij} is the observed failure time or censoring time. We will call it **censored Z-residuals** in this paper. The diagnosis of the GOF of $S_{ij}(y_{ij})$ can be converted to the diagnosis of the normality of $r_{ij}^n(y_{ij})$. The function `gofTestCensored` in R package `EnvStats` [51, 52] provides an SF test for testing the normality of multiply censored data. Hence, `gofTestCensored` can be applied to check the normality of censored Z-residuals for checking the overall GOF of survival models. We will refer to this test using the **CZ-CSF** method in this paper.

Although the aforementioned overall GOF checking methods can be used to determine how closely the residuals are distributed corresponding to their reference distributions when the model assumptions are met, they cannot be used to test the plausibility of specific model assumptions, in particular, the functional form of covariates. For checking whether a functional form of individual covariates may be misspecified, tailored graphical and quantitative diagnostics tools are needed. Martingale and deviance residuals have been proposed to check the functional form in survival analysis. The martingale residuals [13] provide a measure of the discrepancy between the number of predicted death by the model and the number of observed failures in the interval $(0, t_{ij})$, which is either 1 or 0. The martingale residuals are defined as $r_{ij}^M = \delta_{ij} - r_{ij}^c$, where δ_{ij} is the event indicator for the jth individual of the ith group observation, δ_{ij} is equal to 1 if that observation is an event; otherwise zero if censored, and r_{ij}^c is the Cox-Snell residual. The martingale residuals sum to zero, but are not symmetrically distributed about zero [11]. The deviance residuals [14, 15] can be regarded as an attempt to make the martingale residuals symmetrically distributed about zero, and are defined as $r_{ij}^D = sgn(r_{ij}^M)[-2(r_{ij}^M + \delta_{ij}\log(\delta_{ij} - r_{ij}^M))]^{\frac{1}{2}}$, where r_{ij}^M is the martingale residual, the function $sgn(.)$ is the sign function [11]. Other residual-based diagnostics tools have also been proposed for censored survival models; see Grambsch and Therneau [17], Peng and Taylor [53], Keleş and Segal [54], Farrington [55], Davison and Gigli [56], Lin et al. [57], Law and Jackson [58], Shepherd et al. [59], Hillis [60] and the references therein. A common drawback for these residuals is that their distributions under the true model are very complicated due to the censoring, hence, they cannot be characterized by a known distribution or probability table, posing challenges for devising numerical tests based on these conventional residuals for diagnosing survival models.

3.4 Z Residual

3.4.1 Definition of Z Residual

In this paper, we extended Z-residual [19], to diagnose shared frailty models in a Cox proportional hazard setting with a baseline function unspecified. The normalized randomized survival probabilities (RSPs) for y_{ij} in the shared frailty model is defined as:

$$S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij}) = \begin{cases} S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is uncensored, i.e., } \delta_{ij} = 1, \\ U_{ij}\, S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is censored, i.e., } \delta_{ij} = 0, \end{cases} \tag{3.9}$$

where U_{ij} is a uniform random number on $(0, 1)$, and $S_{ij}(\cdot)$ is the postulated survival function for t_{ij} given x_{ij}. $S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij})$ is a random number between 0 and $S_{ij}(y_{ij})$ when y_{ij} is censored. It is proved that the RSPs are uniformly distributed on $(0, 1)$ given x_i under the true model [19]. Therefore, the RSPs can be transformed into residuals with any desired distribution. We prefer to transform them with the normal

quantile:

$$r_{ij}^{Z}(y_{ij},\delta_{ij},U_{ij}) = -\Phi^{-1}(S_{ij}^{R}(y_{ij},\delta_{ij},U_{ij})), \tag{3.10}$$

which is normally distributed under the true model, so that we can conduct model diagnostics with Z-residuals for censored data in the same way as conducting model diagnostics for a normal linear regression model. There are a few advantages of transforming RSPs into Z-residuals. First, the diagnostics methods for checking normal linear regression are rich in the literature. Second, transforming RSPs into normal deviates facilitates the identification of extremely small and large RSPs. The frequency of such small RSPs may be too small to be highlighted by the plots of RSPs. However, the presence of such extreme SPs, even very few, is indicative of model misspecification. Normal transformation can highlight such extreme RSPs.

3.4.2 Diagnosis of the Functional Form of Covariates using Z-residuals

A QQ plot based on Z-residuals can be used to graphically assess the model's overall GOF, and Shapiro-Wilk (SW) or Shapiro-Francia (SF) test applied to Z-residuals can be used to numerically test the overall GOF of the model. The conditional distribution of Z-residual given x_i is approximately a standard normal and is homogeneous at varying levels of covariates when a model is correctly specified. For checking the functional form of the covariate, we can plot Z-residuals against covariates and/or linear predictors. When the functional form is correctly specified, we expect that there is no trend in these scatterplots. However, such a graphical examination is difficult to determine whether the observed trend in Z-residuals is caused by chance or by the misspecification in the covariate function. Therefore, we desire a formal test to quantify the statistical significance of the difference between the observed trend and the expected horizontal line at 0. In this paper, we propose the following diagnostics procedure. The Z-residuals can be divided into k groups by cutting the covariates or linear predictors into equally-spaced intervals. Figure 3.1 demonstrates two scatterplots about Z-residuals by cutting the covariate X into equally-spaced intervals. The left panel shows that the Z-residuals are randomly scattered without showing differential group means or variances. The right panel clearly shows that the Z-residuals are not homogeneous; particularly their group means differ substantially. A quantitative method to assess the homogeneity of such grouped Z-residuals is to test the equality of group means of the Z-residuals. We apply the F-test in ANOVA to test the equality of the means of grouped Z-residuals as shown in Figure 3.1.

3.4.3 A P-value Upper Bound for Assessing Replicated Z-residuals GOF Test p-values

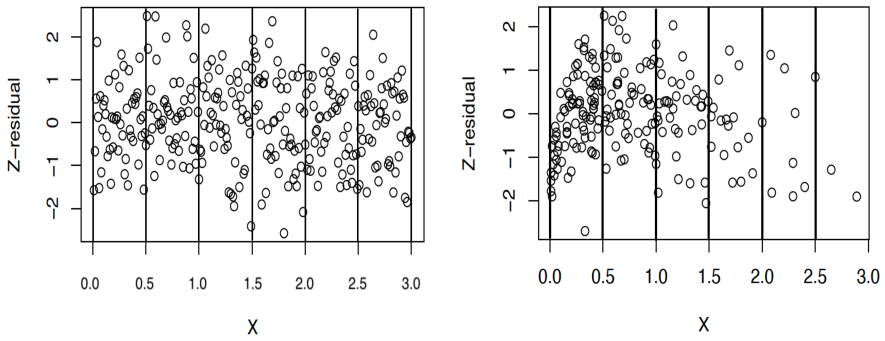
A difficulty in conducting statistical tests with Z-residuals is the randomness in the test p-values. Given a fitted model, we can generate many sets of Z-residuals and obtain replicated test p-values. According to the distribution of order statistics of correlated random variables [61, 62], we can obtain the following inequality for the rth order statistics $p_{(r)}$:

$$P(p_{(r)} < t) \leq \min\left(1, t\frac{J}{r}\right). \tag{3.11}$$

Based on (3.11), a p-value upper bound for observed (simulated) rth statistics $p_{(r)}^{\text{obs}}$ is given by $\min\left(1, p_{(r)}^{\text{obs}}\frac{J}{r}\right)$. To avoid the selection of r, we report the minimal upper bound for $r = 1, \ldots, J$, denoted by $p_{\min}$:

$$p_{\min} = \min_{r=1,\ldots,J} \min\left(1, p_{(r)}^{\text{obs}}\frac{J}{r}\right). \tag{3.12}$$

Figure 3.1: An illustrative plot showing how to construct the non-homogeneity test with Z-residuals: dividing Z-residuals by a covariate or linear predictor (LP) with equally-spaced interval, then testing the equality of the means of grouped residuals. This figure shows two scatterplots of Z-residuals of two models: a linear effect model (the left plot) and a nonlinear effect model (the right plot). The covariate X_{ij} is from positive Normal(0, 1), we generate the failure times t_{ij} from a shared frailty model with Weibull baseline with the following hazard function: $h_{ij}(t_{ij}) = z_i \exp(\beta \log(X_{ij}))h_0(t_{ij})$, where h_0 is the hazard function of Weibull with shape α=3 and scale λ=0.007. In addition to fitting the nonlinear model with $\log(X)$ as a covariate to these datasets, we also consider fitting the shared frailty gamma model assuming linear effect for X as a linear model. Then we can check whether the Z-residuals of the k groups are homogeneously distributed.



The $p_{\min}$ is rather conservative for assessing model fit because of its generality. When a model has a small $p_{\min}$, it is highly suspected that the model can be improved for better fitting the dataset. Considering the conservatism of $p_{\min}$, a rule of thumb for declaring model failure in practice should be much larger, say 0.25 as suggested by Yuan and Johnson [63], than the conventional 0.05 for exact p-values.

3.5 Simulation Studies

In this section, we present simulation studies to demonstrate the effectiveness of the Z-residuals in checking the adequacy of the functional form of covariates. Three covariates are generated as follows: $x_{ij}^{(1)}$ is from Uniform[0, 1], $x_{ij}^{(2)}$ is from positive Normal(0, 1), and $x_{ij}^{(3)}$ is from Bern(0.25). We generate the failure times t_{ij} from a shared frailty model with Weibull baseline with the following hazard function:

$$h_{ij}(t_{ij}) = z_i \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_2 \log\Big(x_{ij}^{(2)}\Big) + \beta_3 x_{ij}^{(3)}\Big) h_0(t_{ij}), \tag{3.13}$$

where h_0 is the hazard function of Weibull with shape α=3 and scale λ=0.007. The data generator is given by $t_{ij} = \left\{ \frac{-\log(u_{ij})}{\lambda z_i \exp\left(x_{ij}^{(1)} - 2\log\left(x_{ij}^{(2)}\right) + 0.5x_{ij}^{(3)}\right)} \right\}^{(1/\alpha)}$, where $i = \{1, \cdots, \text{g}\}$ and $j = \{1, \cdots, n_i\}$ and u_{ij} is simulated from Uniform ([0, 1]); the frailty term z_i is generated from a gamma distribution with a variance of 0.5. The censoring times C_{ij} are simulated from exponential distributions. The rates γ were set to four different values to obtain four different censoring rates: 0%, 20%, 50%, and 80%. We fixed the number of clusters $g = 20$ and set the cluster size (n_i, the sample size in each cluster) to be 10 values: 10, 20, ..., 100. For each combination of cluster size and censoring rate, we generated 1000 datasets for estimating model rejection rates of different diagnostics methods. In addition to fitting the true model with $\log(x_2)$ as a covariate to these datasets, we also consider fitting the shared frailty gamma model assuming linear effect for x_2 as a wrong model to investigate the performance of different diagnostics methods.

We first show the performance of graphical methods for assessing the overall GOF for a single simulated dataset with 20 clusters of 40 observations in each cluster and the percentage of censoring $c \approx 50\%$. As shown in the panels of the first row of Figure 3.2, the CHFs of the CS residuals of both of the true and wrong models align well along the 45° straight line, suggesting that the CS residuals cannot effectively detect the model misspecification of the wrong model with linear covariate effects. The normality of the Z-residual under the true and the wrong models is examined via QQ plots, as shown in the panels of the second row of Figure 3.2. The points in the two QQ plots for Z-residuals align very well along with straight lines, indicating that the distributions of the Z-residuals under the true and the wrong models are very close to a normal distribution. Therefore, the QQ plots of Z-residuals cannot detect the misspecification in the wrong model either.

The panels in the third and fourth rows of Figure 3.2 demonstrate the advantage of examining the scatterplots of Z-residuals against the linear predictor for diagnosing the misspecification of the functional form of covariates. Under the true model, the residuals are mostly bounded between -3 and 3 as the standard normal variates without a visible trend. We can see the LOWESS curve in the scatterplot under the true model is very close to the horizontal line at 0. For the wrong model, a non-linear trend in the Z-residuals is clearly observed. In the fourth row, we first divide Z-residuals into $k = 10$ groups by cutting the linear predictors into equally spaced intervals. The scatterplot and the boxplot indicate that the Z-residuals are homogeneous across groups under the true model, but exhibit differential group means under the wrong model. We further checked the scatterplots and grouped boxplots of Z-residuals against $\log(x_2)$ under the true and wrong models, as shown in the fifth and sixth rows of Figure 3.2. The Z-residuals of the true models are fairly homogeneous against $\log(x_2)$. By contrast, for the wrong model, we see a clear non-linear pattern in the scatterplots and differential group means in the boxplots against $\log(x_2)$; these plots suggest that the model with linear covariate effects does not fit well to the dataset.

As a comparison, we also show the performance of martingale and deviance residuals for assessing the functional form of x_2 in Figure 3.3 by displaying the martingale and deviance residuals against the covariate $\log(x_2)$ under the true and wrong models, respectively. Under the true model, the martingale residuals are mostly within the interval (-4, 1); the deviance residuals are more symmetrically distributed than martingale residuals and they are mostly within the interval (-3, 3). The LOWESS curves in the scatterplots of martingale and deviance residuals under the true model are very close to horizontal lines. Note that the LOWESS curve is slightly tilted downward on the right because the censoring occurs more frequently for cases with large $\log(x_2)$. Under the wrong model, the LOWESS curves show more pronounced non-horizontal trends in the scatterplots of martingale and deviance residuals. From this comparison, we see that the scatterplots of martingale and deviance residuals can distinguish the true and wrong models and confirm that the true model is a better model for the dataset. However, due to the lack of numerical measures, we cannot tell whether the observed non-horizontal trend is caused by chance or due to a misspecified functional form for the covariate. The decision based on visual inspection is often subjective.

In addition to the graphical assessment, numerical tests with Z-residuals can be constructed as Z-residuals are approximately distributed as the standard normal under the true model. We compare a set of residual-based testing methods for detecting the inadequacy of fitted models. The overall GOF test methods are denoted by "R-T" with "R" denoting the residual name and "T" denoting the test method. For example, Z-SW is the test method that the normality of Z-residuals is tested with the SW test. In particular, CZ-CSF is the method that the normality of censored Z-residuals (shortened by CZ) is tested by an extended SF method for censored observations, which is implemented with `gofTestCensored` in the R package `EnvStats`. For detecting the misspecification in the covariate functional form, we can divide Z-residuals into groups by

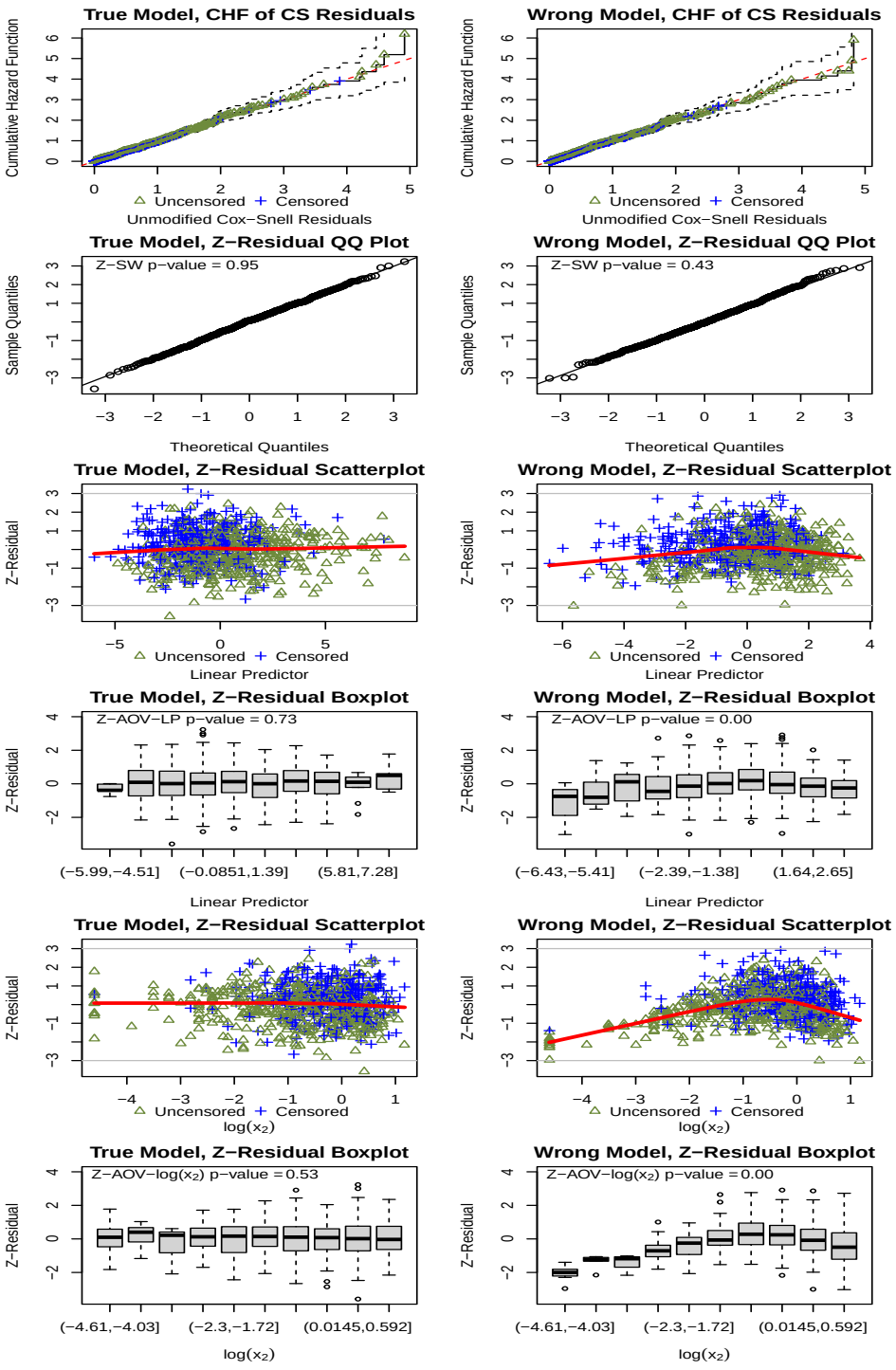


Figure 3.2: Performance of the Z-residuals and CS residuals as graphical tools for detecting the misspecification of the functional form of covariates. The dataset was generated with 20 clusters of 40 observations in each cluster and a censoring rate $c \approx 50\%$.

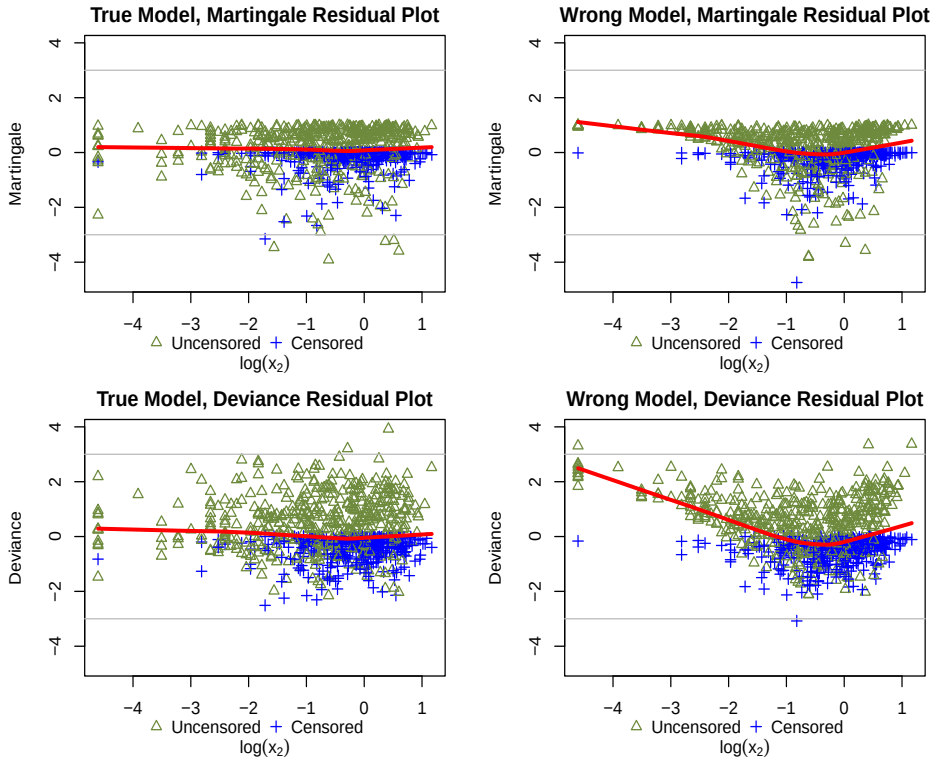


Figure 3.3: Performance of the martingale and deviance residuals as a graphical tool for checking the functional form of covariates. The dataset has a sample size n = 800 and a censoring rate $c \approx 50\%$.

cutting the linear predictor or a covariate into equally-spaced intervals as shown by the boxplots of Figure 3.2. We can then test the homogeneity of Z-residuals across the groups. Z-AOV-LP is the method of applying ANOVA to test the equality of the means of Z-residuals against the groups formed with the linear predictor (LP) and Z-AOV-log(x_2) is the method of testing the equality of the means of Z-residuals against the groups formed with the covariate log(x_2).

We simulated 1000 datasets for each combination of cluster size and censoring rate as described at the beginning of this section. Using the 1000 datasets generated from the true model under each scenario, the model rejection rate of each test method was estimated by the proportion of the test p-values less than 0.05. The model rejection rates of all the considered test methods are shown in Figures 3.4 and 3.7. The non-homogeneity test methods, Z-AOV-LP and Z-AOV-log(x_2), can detect the non-linear covariate effects with very high true-positive rates (model rejection rates under the wrong models) and low false-positive rates (model rejection rates under the true models). Of all the compared test methods, Z-AOV-log(x_2) performs the best for detecting the nonlinear covariate effects with the highest powers, which are nearly 100%, and the powers stay high even for the scenario with a cluster size as small as 10. The Z-SW, Z-SF, and CZ-CSF tests have false-positive rates close to the nominal level of 5% for all scenarios and have certain powers when the censoring rate is less than 80%. We also note that their powers increase as the cluster size increases. However, the powers of these overall GOF tests are significantly smaller than the corresponding powers of the Z-AOV-LP and Z-AOV-log(x_2) methods. The comparison demonstrates the advantage of testing the homogeneity of Z-residuals for checking the assumption of covariate functional form in addition to the overall GOF tests, which do not inspect the relationship between residuals and covariates.

In appended Figure 3.7, we show the performances of the Z-KS and Dev-SW tests, which were separated from Figure 3.4 for better visualization. Z-KS test has low false-positive rates but also very low powers, which shows the conservatism of the KS test for testing the normality of Z-residuals. When the censorship is 0, the performance of Dev-SW is satisfactory. However, when there are censored observations, the Dev-SW method has very high (nearly 100%) model rejection rates when the model is correctly specified. Hence, the high powers of Dev-SW do not indicate that it is a good test method.

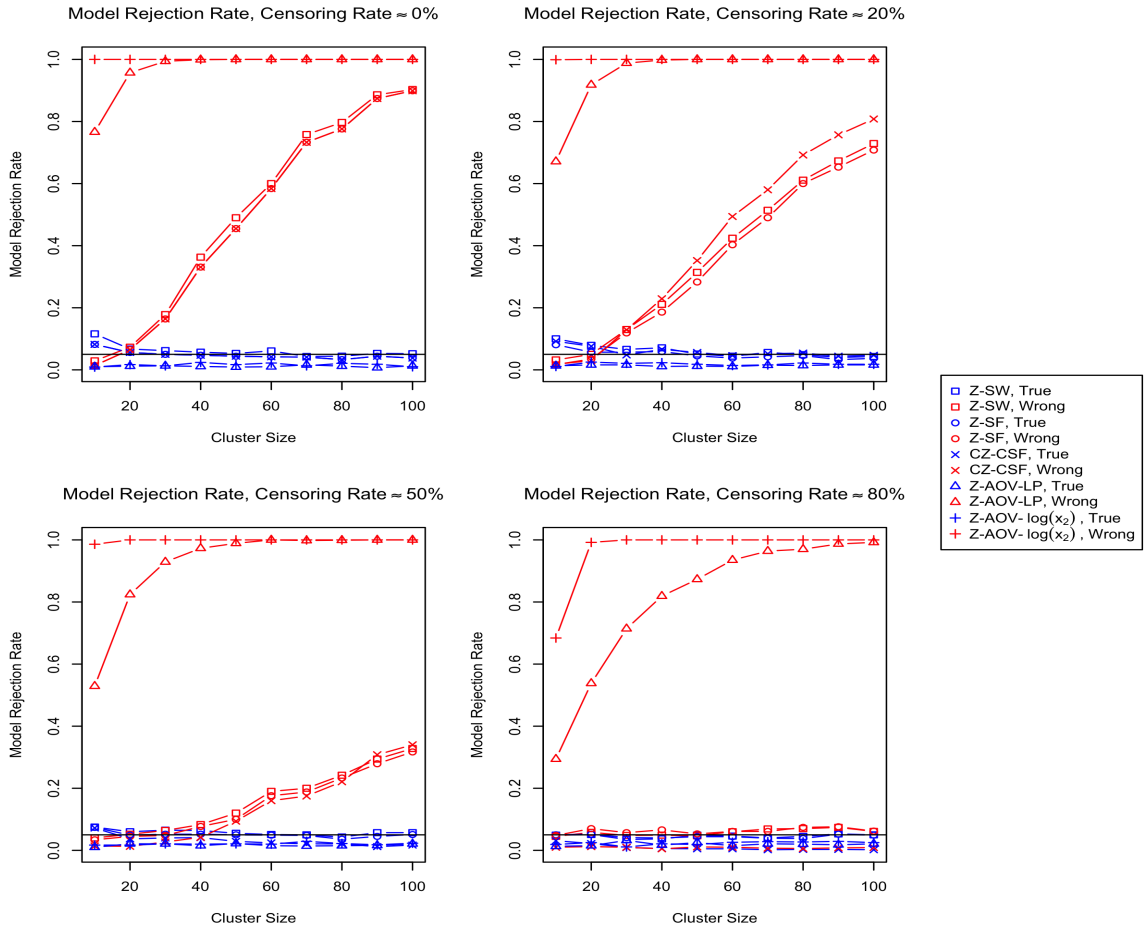


Figure 3.4: Model rejection rates of various statistical tests based on Z-residual. A model is rejected when the test p-value is smaller than 5%. Note that we use a random Z-residual test p-value rather than the $p_{\min}$.

3.6 A Real Data Example

In this section, we apply the proposed residual diagnostics tools based on Z-residuals to diagnose the functional form of covariates in a real application for modelling the survival times of acute myeloid leukemia patients. The dataset contains 1498 patients recorded at the M. D. Anderson Cancer Center between 1980 and 1996 [47]. The dataset used in our analysis contains 411 patients who are aged below 60 from 24 administrative districts recorded at the M.D Anderson Cancer Center between 1980 and 1996. The data collected information on the survival time for acute myeloid leukemia and prognostic factors, including age, sex, white blood cell

count (wbc) at diagnosis, and the townsend score (tpi) for which higher values indicate less affluent areas. The censoring rate is 29.2%. The response variable of interest is the survival time in days, which is the time from entry to the study or death. The preliminary study showed that the wbc is highly right-skewed. Natural logarithm transformation is often used to reduce the impact of extremely large values of the covariate on the response variable, such as the wbc variable in this application. However, a natural logarithm transformation may mask the impact of extremely large values of the covariate on the outcome variable.

This study was deemed exempt from ethics approval since the data utilized were publicly available in the journal [47], and no personally identifiable information was collected or used. This research adhered to the principles outlined in TCPS 2-2nd Edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans [64].

We fitted two shared frailty models, one with covariates wbc, age, sex and tpi, which is labelled as the wbc model, and the other with natural log(wbc) replacing wbc, which is labelled as the lwbc model. Table 3.1 shows the estimated regression coefficients, the corresponding standard errors and p-values for the covariate effects from fitting the two shared frailty models. The results indicate that the estimated effect of wbc is statistically significant (p-value < 0.001) but the effect of log(wbc) is not significant (p-value=0.135). The difference in the p-values for wbc and log(wbc) highlights that the statistical inference of the covariate effect may depend on the assumption of the functional form of the covariates.

Table 3.1: Parameter estimates of the shared gamma frailty model in the real data application.

(a) The wbc model

Covariates	Estimate	SE	P-value
Age	0.021	0.005	0.000
$SexMale$	0.215	0.118	0.068
wbc	0.005	0.001	0.000
tpi	0.023	0.016	0.140
$Frailty$			0.906

(b) The lwbc model

Covariates	Estimate	SE	P-value
Age	0.021	0.005	0.000
$SexMale$	0.216	0.118	0.069
$log(wbc)$	0.035	0.024	0.135
tpi	0.024	0.016	0.128
$Frailty$			0.906

The overall GOF tests and graphical checking with CS residuals and Z-residuals show that both the wbc and lwbc models provide adequate fits to the dataset. The first row of Figure 3.5 shows that the estimated CHFs of the CS residuals of both of the wbc and lwbc models align closely along the 45° diagonal line. Similarly, the QQ plots (the second row of Figure 3.5) of Z-residuals of these two models align well with the 45° diagonal line. The scatterplots of Z-residuals against the linear predictor don't exhibit visible trends; their LOWESS lines are very close to the horizontal line at 0; the boxplots of Z-residuals grouped by cutting linear predictors into equal-spaced intervals (the fourth row of Figure 3.5) appear to have equal means and variance across groups. The Z-AOV-LP test also gives large p-values for the wbc and lwbc models (0.63 and 0.76 respectively).

The above diagnostics results reveal no serious misspecification in these two models. However, the inspection of the Z-residuals against the covariate log(wbc) reveals that the functional form of the lwbc model is likely misspecified. The scatterplots and comparative boxplots of the Z-residuals against log(wbc) are shown in the fifth and sixth rows of Figure 3.5. The LOWESS curve of the wbc model appears to align well with the horizontal line at 0 and the grouped Z-residuals of the wbc model appear to have equal means and variances across groups. However, the diagnosis results for the lwbc model are very different. It appears that there is a non-linear trend in the LOWESS curve of the lwbc model and the grouped Z-residuals appear to have different means across groups. To measure the statistical significance of the observed trends, we

apply Z-AOV-log(wbc) to test the equality of the means of the grouped Z-residuals of these two models. The p-values are 0.16 and < 0.001 respectively for the wbc and lwbc models as shown in the boxplots. The very small p-value of the Z-AOV-log(wbc) test for the lwbc models strongly suggests that the log transformation of wbc is likely inappropriate for modelling the survival time.

The Z-residual test p-values quoted above contain randomness because of the randomization in generating Z-residuals. To ensure the robustness of the model diagnostics results, we generated 1000 replicated test p-values with 1000 sets of regenerated Z-residuals for each test method. Figure 3.6 displays the histograms of 1000 replicated Z-residual test p-values for the wbc and lwbc models. The red vertical lines in these histograms show the upper bound summaries of these replicated p-values, $p_{\min}$ (see Sec. 3.4.3 for details). These histograms show that the Z-SW, Z-SF, and Z-AOV-LP tests for both models give a large proportion of p-values greater than 0.05, and the large p-values result in large $p_{\min}$ values. In contrast, the replicated Z-AOV-log(wbc) p-values for the lwbc model are almost all smaller than 0.001. The consistently small Z-AOV-log(wbc) p-values further confirm that the log transformation of wbc is inappropriate for modelling the survival time.

Table 3.2 tabulates all the $p_{\min}$ values (shown with red lines in Figure 3.6) for diagnosing the two models with Z-residual-based tests. In addition, we also report the non-random CZ-CSF test p-values for the two models and the AIC values for comparing these two models. The CZ-CSF p-values of both models are larger than 5% (Table 3.2). Therefore, the CZ-CSF test does not identify the inadequacy of the lwbc model either. The AIC value, 3132.105, of the lwbc model, is much larger than the AIC value 3111.669 of the wbc model, which indicates that the wbc model provides a better model fit compared to the lwbc model. This conclusion is consistent with the model diagnostics results as given by the Z-AOV-log(wbc) test, which reveals that the lwbc model is inappropriate for modelling the survival time of this dataset by checking the homogeneity of Z-residuals against log(wbc). Although the AIC of the wbc model is smaller than that of the lwbc model, we also see that a large proportion of Z-AOV-log(wbc) p-values for the wbc model are tiny; the $p_{\min}$ value is 0.074. We think that the wbc model could be improved to provide a better fit for the survival time of this dataset.

Besides the log transformation, we have also fit models with three more transformations of the wbc variable, including squaring, taking the square root, and the cubic root of wbc, for finding a better model. The residual diagnostic results with Z-residuals and AIC values for the three non-linear models are shown in the appended Table 3.3. The overall GOF and non-homogeneity tests with Z-residuals, including the CZ-CSF test, give p-values or $p_{\min}$ values that are greater than 0.05. Therefore, they do not identify serious inadequacies in these three models. However, the $p_{\min}$ of the non-homogeneity test with Z-residuals for the cbrt(wbc) model is as small as 0.091, suggesting that the cubic root transformation is not ideal for modelling the survival time of acute myeloid leukemia patients. In addition, the AIC values of all the models are all moderately larger than the AIC value of the wbc model. In summary, these transformations may not have caused serious model inadequacies as shown by Z-residual diagnostic results but do not provide a significantly better model fit.

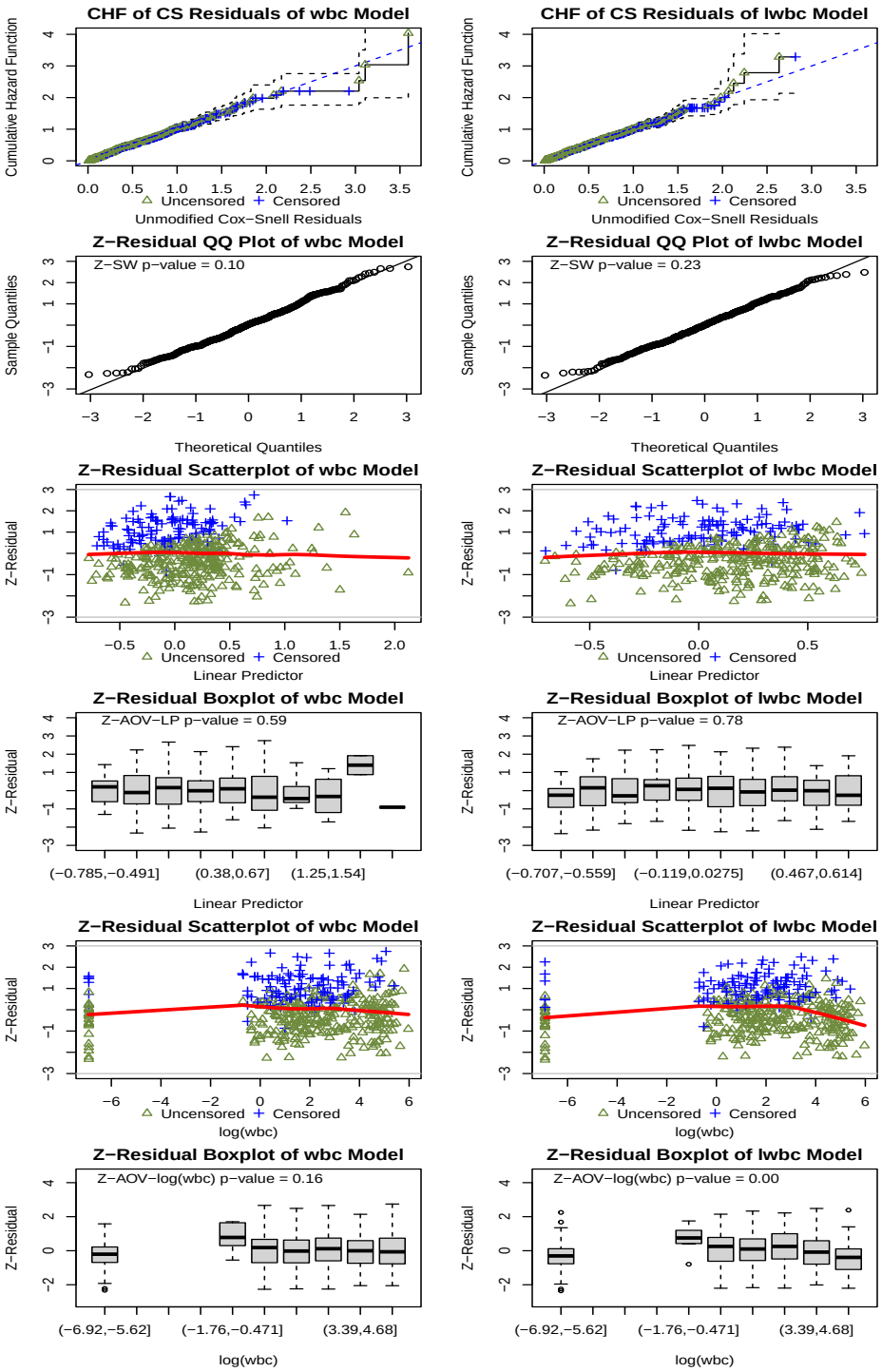


Figure 3.5: Diagnostics results for the wbc (left panels) and lwbc (right panels) models fitted to the survival data of acute myeloid leukemia patients.

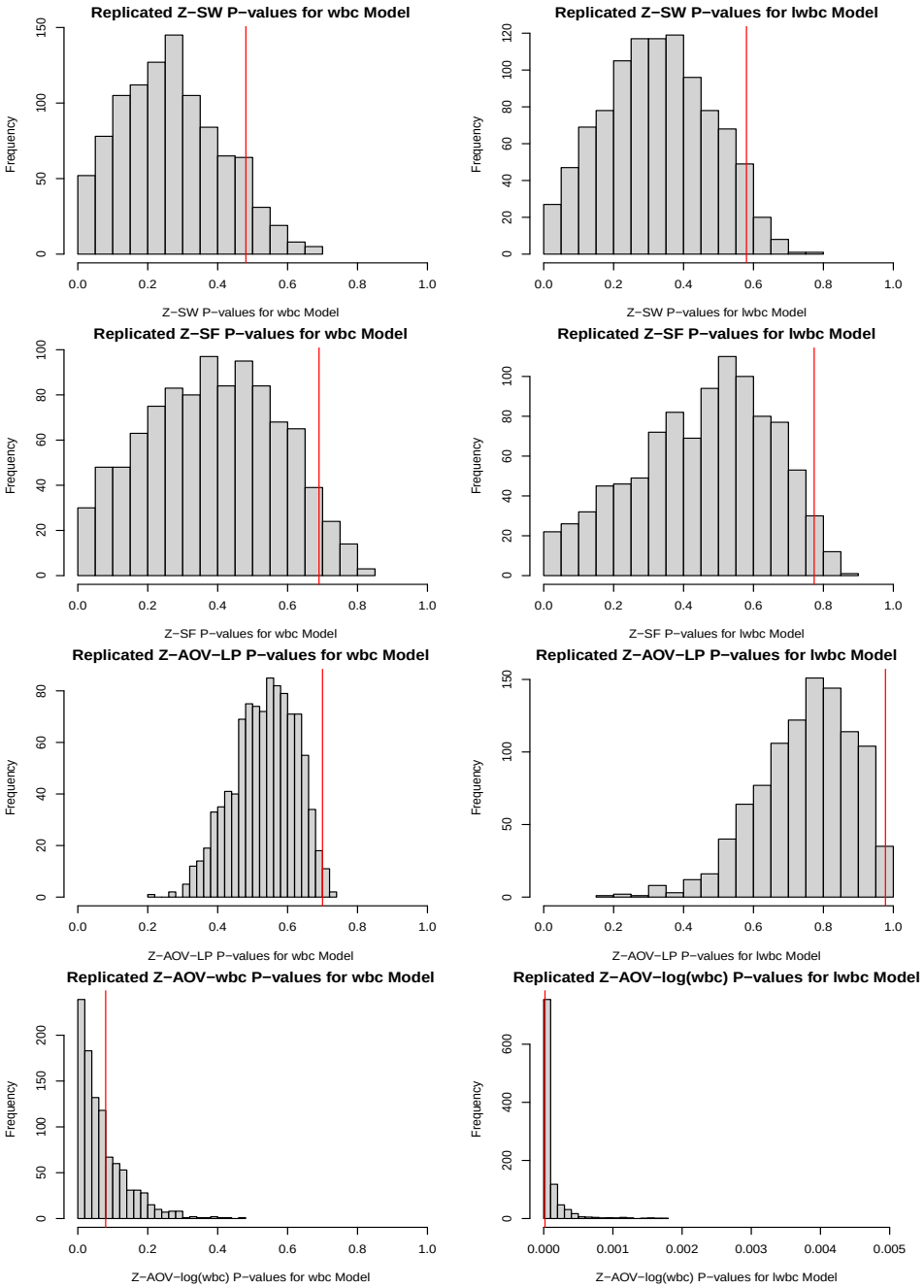


Figure 3.6: The histograms of 1000 replicated Z-SW, Z-SF, Z-AOV-LP and Z-AOV-log(wbc) p-values for the wbc model (left panels) and the lwbc model (right panels) fitted with the survival times of acute myeloid leukemia patients. The vertical red lines indicate $p_{\min}$ for 1000 replicated p-values. Note that the upper limit of the x-axis for Z-AOV-log(wbc) p-values for the lwbc model is 0.005, not 1 for others.

Table 3.2: AIC, p-values or $p_{\min}$ values for the CZ-CSF test, $p_{\min}$ for Z-SW, Z-SF, Z-AOV-LP and Z-AOV-log(wbc) test for the wbc and lwbc models, respectively, for the acute myeloid leukemia data.

Model	AIC	CZ-CSF $p-$value	Z-SW $p_{\min}$	Z-SF $p_{\min}$	Z-AOV-LP $p_{\min}$	Z-AOV-log(wbc) $p_{\min}$
wbc model	3111.669	0.255	0.495	0.693	0.703	0.074
lwbc model	3132.105	0.305	0.579	0.781	0.978	**<0.00001**

3.7 Discussions and Conclusions

In this paper, we extended the idea of randomized survival probability [19] to develop a residual diagnostic tool that can provide both graphical and numerical results for checking the covariate functional form in semi-parametric shared frailty models. We proposed a non-homogeneity test for testing whether there is a trend in Z-residuals for checking the covariate functional form. Our extensive simulation studies showed that the overall GOF tests (including CS-CSF, Z-SW, and Z-SF) may not be powerful enough for detecting the misspecification in covariate functional form and that the proposed non-homogeneity tests based on the Z-residuals are significantly more powerful than the aforementioned overall GOF tests. Applied to a real dataset, the Z-residual diagnostics discovers that a model with log-transformation is inappropriate for modelling the survival time of acute myeloid leukemia patients, which is not captured by other diagnostics methods.

The Z-residuals-based diagnostics methods can be extended in several directions. When the full dataset is used to estimate the model parameters and used to calculate residuals for model checking, there might be a conservatism problem (bias) due to the double use of the dataset. The double use of the data may reduce the power of detecting model misspecification, especially in the case of a small sample size or high censoring rate. Cross-validation could be a good method to solve this problem. The cross-validatory Z-residual may be a more powerful tool for identifying the model inadequacy in the survival data.

In this paper, we considered semiparametric shared frailty models assuming proportional hazards. However, if the model includes time-varying coefficients or time-dependent explanatory variables, the proportional hazards assumption is violated. A number of residuals have been proposed for evaluating the assumption of proportional hazards. Traditionally, the Schoenfeld [11, 16] and Scaled Schoenfeld [17] residuals are often used in testing the assumption of proportional hazard. Lin et al. [57] proposed the cumulative sums of martingale residuals to check the validity of the PH assumption. Extending the Z-residual for diagnosing the proportional hazard assumption and comparing it with existing residual diagnostics tools warrants a research topic in the future.

Additional Figures and Tables

Supplementary Figures for Section 3.5

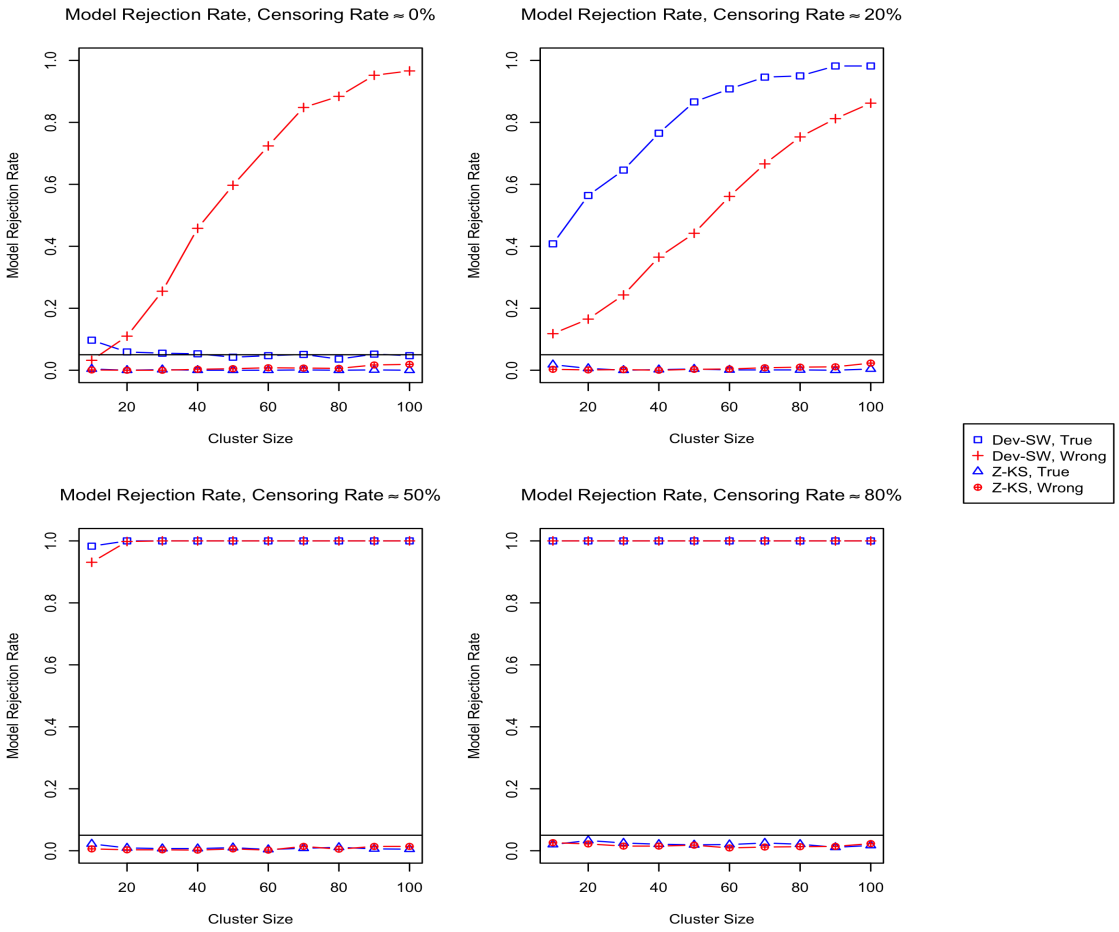


Figure 3.7: Model rejection rate of the KS test applied to Z-residuals (Z-KS) and the SW test applied to deviance residuals (Dev-SW) for the simulation study in Sec. 3.5. A model is rejected when the test p-value is smaller than 5%. The model rejection rates of Dev-SW tests are nearly 1 under the true and wrong models when the censoring rate is 50% and 80%, hence, they are almost overlapped in the plots.

Supplementary Tables for Section 3.6

Table 3.3: AIC, p-values or $p_{\min}$ values for the CZ-CSF test, $p_{\min}$ for Z-SW, Z-SF, Z-AOV-LP and Z-AOV-wbc test for the square(wbc), square root(wbc) and cubic root(wbc) models, respectively, for the acute myeloid leukemia data.

Model	AIC	CZ-CSF $p-$value	Z-SW $p_{\min}$	Z-SF $p_{\min}$	Z-AOV-LP $p_{\min}$	Z-AOV-wbc $p_{\min}$
sq(wbc) model	3118.478	0.210	0.498	0.696	0.732	0.686
sqrt(wbc) model	3114.666	0.271	0.327	0.318	0.918	0.809
cbrt(wbc) model	3118.848	0.370	0.627	0.811	0.996	0.091

4 Cross-validatory Z-Residual for Diagnosing Shared Frailty Models[1]

Abstract: Residual diagnostic methods play a critical role in assessing model assumptions and detecting outliers in statistical modelling. In the context of survival models with censored observations, The paper [19] introduced the Z-residual, which follows an approximately normal distribution under the true model. This property makes it possible to use Z-residuals for diagnosing survival models in a way similar to how Pearson residuals are used in normal linear regression. However, computing residuals based on the full dataset can result in a conservative bias that reduces the power of detecting model misspecification, as the same dataset is used for both model fitting and validation. Although cross-validation is a potential solution to this problem, it has not been commonly used in residual diagnostics due to computational challenges. In this paper, we propose a cross-validation approach for computing Z-residuals in the context of shared frailty models. Specifically, we develop a general function that calculates cross-validatory Z-residuals using the output from the `coxph` function in the `survival` package in R. Our simulation studies demonstrate that, for goodness-of-fit tests and outlier detection, cross-validatory Z-residuals are significantly more powerful and more discriminative than Z-residuals without cross-validation. We also compare the performance of Z-residuals with and without cross-validation in identifying outliers in a real application that models the recurrence time of kidney infection patients. Our findings suggest that cross-validatory Z-residuals can identify outliers that are missed by Z-residuals without cross-validation.

4.1 Introduction

Residual diagnosis is a critical step in statistical modelling for checking the validity of model assumptions. Several residual diagnostic tools have been commonly used for checking the survival models [11], including Cox-Snell (CS) [12], martingale [13], deviance [14, 15], Schoenfeld [11, 16] and scaled Schoenfeld [17] residuals. For example, the plot based on Cox-Snell (CS) residuals can be used as a graphical assessment tool for checking the overall goodness-of-fit (GOF) of a fitted model. The functional form of covariate is often examined using the plots of the martingale and deviance residuals against the covariates. The Schoenfeld and scaled Schoenfeld residuals are often used in testing the assumption of proportional hazards in the Cox proportional hazard model. Other residuals have also been proposed for diagnosing survival models [53–60]. However, there is a lack of residuals with a characterized reference distribution for censored regression. The paper [19] recently proposed the Z-residual diagnosis tool for diagnosing survival models with censored observations. The Z-residual is approximately normally distributed under the true model and has greater statistical power and is more informative than some traditional residual diagnostic tools for diagnosing model misspecifications, such as incorrect choice of distribution family and/or functional form of covariates of survival models.

[1]This chapter has been deposited as an arXiv paper:2303.09616

The residuals considered in practical survival analysis are typically calculated based on the full dataset. When the same dataset is used to estimate the model parameters and calculate residuals for checking the fitted model, the power of detecting model misspecification may be reduced (bias) due to the double use of the dataset. The bias of the double use of the dataset has received much attention in the context of checking and comparing Bayesian models; see [65–73] and the references therein. For example, The paper [72] introduced integrated importance sampling methods for approximating leave-one-out cross-validatory predictive evaluations for models with unit-specific and possibly correlated latent variables. Cross-validatory predictive p-values can also be used to identify outliers by examining the tail probability of the predictive distribution [65, 66]. However, cross-validation is not commonly used in practical residual diagnosis in the frequentist paradigm. This is probably due to the computational challenges in cross-validation and the lack of awareness of the severity of the bias caused by the double use of the dataset.

In this paper, we consider developing cross-validation methods for calculating the Z-residual for diagnosing survival models and comparing them to the Z-residual without cross-validation. We will focus on investigating the performance of cross-validatory Z-residuals in diagnosing shared frailty models. A shared frailty model is a survival model by incorporating random effects (frailties) to account for unobserved heterogeneity [24], where the frailties are shared among individuals within a cluster or group [1–4]. We develop a general R function for calculating the cross-validatory Z-residuals based on the outputs from fitting a survival model using the `coxph` function in `survival` package. We build an R function for splitting data into K-fold to ensure adequate representations of groups and other covariates in each fold. In our study design, the Z-residuals are calculated using three methods: the full dataset (No-CV), 10-fold cross-validation (10-fold) and leave-one-out cross-validation (LOOCV). We conduct simulation studies to investigate the performances of the three types of Z-residuals in detecting nonlinear covariate effects and identifying outliers through graphical visualization and SW tests. Our simulation results show that the SW tests based on 10-fold Z-residual and LOOCV Z-residual are significantly more powerful and more discriminative for detecting non-linear covariate effects. Moreover, cross-validatory Z-residuals are more powerful and more discriminative for identifying outliers than No-CV Z-residuals. In spite of these improved performances, our simulation studies also show that the cross-validation results in elevated type-I error rates when cross-validatory Z-residuals are used in conducting GOF tests. Future research can be conducted to remedy this problem for cross-validatory Z-residuals. We also compared the performance of the No-CV Z-residual and LOOCV Z-residual in identifying outliers for a kidney infection dataset [74]. The results show that the methods with LOOCV Z-residuals can identify some outliers that are missed by the No-CV method.

The rest of this paper is organized as follows. Section 4.2 gives a brief review of shared frailty models. In Section 4.3 we present the definition of the cross-validatory Z residual with a discussion of the algorithm for computing cross-validatory Z residuals. Section 4.4 presents the results of simulation studies for investigating the performances of 10-fold and LOOCV Z residuals. In section 4.5, we present the results of applying the LOOCV Z residual to identify outliers for a kidney infection dataset. The article is concluded in Section 4.6.

4.2 Shared frailty models

A shared frailty model is a frailty model where the frailties are common or shared among individuals within groups. The formulation of a frailty model for clustered failure survival data is defined as follows. Suppose there are g groups of individuals with n_i individuals in the ith group, $i = 1, 2, \ldots, g$. If the number of subjects n_i is 1 for all groups, then the univariate frailty model is obtained [3]. Otherwise, the model is called

the shared frailty model [2, 20, 21] because all subjects in the same cluster share the same frailty value z_i. Suppose t_{ij} is the true failure time for the jth individual of the ith group, which we assume to be a continuous random variable in this article, where $j = 1, 2, \ldots, n_i$. Let t^*_{ij} denote the realization of t_{ij}. In many practical problems, we may not be able to observe t^*_{ij} exactly, but we can observe that t_{ij} is greater than a value c_{ij}, where c_{ij} be the corresponding censoring time. The observed failure times are denoted by the pair (y_{ij}, δ_{ij}), where $y_{ij} = \min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} < c_{ij})$. The observed data can be written as $y = (y_{11}, \ldots, y_{gn_g})$ and $\delta = (\delta_{11}, \ldots, \delta_{gn_g})$. This is called right censoring. Since we will consider only the right censoring in this article, we will use "censoring" as a short for "right censoring". Suppose the survival function of t_{ij} based on a postulated model is defined as $S_{ij}(t^*_{ij}) = P(t_{ij} > t^*_{ij})$, where the subscript ij indicates that the probability depends covariate x_{ij} for the jth individual of the ith group.

For a shared frailty model, the hazard of an event at time t for the jth individual, $j = 1, 2, \ldots, n_i$, in the ith group, is then

$$h_{ij}(t) = z_i \exp(x_{ij}\beta) h_0(t); \tag{4.1}$$

and the survival function for the jth individual of the ith group at time t follows:

$$S_{ij}(t) = \exp\left\{ -\int_0^t h_{ij}(t)\,\mathrm{d}t \right\} = \exp\left\{ - z_i \exp(x_{ij}\beta) H_0(t) \right\}, \tag{4.2}$$

where x_{ij} is a row vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \ldots, x_{gn_g})$; β is the column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, $H_0(t)$ is the baseline cumulative hazard function, and z_i is the frailty term that is common for all n_i individuals within the ith group. Let $z = (z_1, \ldots, z_g)$. The hazard and survival functions with frailty effect can also be written as,

$$h_{ij}(t) = \exp(x_{ij}\beta + u_i) h_0(t), \tag{4.3}$$

and

$$S_{ij}(t) = \exp\left\{ - \exp(x_{ij}\beta + u_i) H_0(t) \right\}, \tag{4.4}$$

where $u_i = \log(z_i)$ is a random effect in the linear component of the proportional hazards model. Note that z_i cannot be negative, but u_i can be any value. If u_i is zero, correspondingly z_i being one, which means the model does not have frailty. The form of the baseline hazard function may be assumed to be unspecified as a semi-parametric model or fully specified to follow a parametric distribution.

In our study, we focus mainly on the shared gamma frailty model, since gamma distribution is the most common distribution for modelling the frailty effect [11]. It is easy to obtain a closed-form representation of the observable survival, cumulative density, and hazard functions due to the simplicity of the Laplace transform [49]. The gamma distribution is a two-parameter distribution with a shape parameter k and scale parameter θ. It takes a variety of shapes as k varies: when $k = 1$, it is identical to the well-known exponential distribution; when k is large, it takes a bell-shaped form reminiscent of a normal distribution; when k is less than one, it takes exponentially shaped and asymptotic to both the vertical and horizontal axes. Under the assumption $k = \frac{1}{\theta}$, the two-parameter gamma distribution turns into a one-parameter distribution. The expected value is one and the variance is equal to θ.

4.3 Cross-validatory Z-residual

The Z-residual is transformed from the randomized survival probability (RSP) introduced in [19]. The key idea of RSP is to replace the survival probability (SP) of a censored failure time with a uniform random number between 0 and the SP of the censored time. RSPs were proved to have a uniform distribution on (0, 1) under the true model with the true generating parameters. The RSP for y_{ij} in a shared frailty model can be then defined as:

$$S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij}) = \begin{cases} S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is uncensored, i.e., } \delta_{ij} = 1, \\ U_{ij}\, S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is censored, i.e., } \delta_{ij} = 0, \end{cases} \tag{4.5}$$

where U_{ij} is a uniform random number on $(0, 1)$, and $S_{ij}(\cdot)$ is the postulated survival function for y_{ij} given x_{ij}. $S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij})$ is a random number between 0 and $S_{ij}(y_{ij})$ when y_{ij} is censored. Li et al.[19] illustrated and proved that the RSP is uniformly distributed on $(0, 1)$ given x_{ij} under the true model. Therefore, they can be transformed into residuals with any desired distribution. It is preferred to transform them with the normal quantile:

$$r_{ij}^Z(y_{ij}, \delta_{ij}, U_{ij}) = -\Phi^{-1}(S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij})), \tag{4.6}$$

where $\Phi()$ is the cumulative distribution function (CDF) of a standard normal distribution. We refer to the residuals as defined in (5.11) as Z-residuals.

Cross-validation (CV) is a re-sampling method for assessing the predictive value. Leave-one-out cross-validation (LOOCV) method is the simplest approach in which each observation is left out as a test case. The outcome from the test data is predicted from a model fitted to the remaining data by using the remaining observations. Since LOOCV is time-consuming, k-fold cross-validation (k-fold CV) is widely used. The observations are randomly divided into k folds of approximately equal size, and observations in one fold are predicted from a model fitted with the observations in the other folds (called training data). In our study, LOOCV and 10-fold CV methods will be used to calculate the cross-validatory Z-residuals.

More specifically, for the LOOCV Z-residual, each observation t_{ij}^{test} is left out from the full dataset with n observations. This dataset with each case is considered as test data and the datasets with the remaining cases are considered as the training dataset, which is used for estimating the parameters. Once the shared frailty model has been fitted to the training dataset, a row vector of the estimated regression coefficients, $\hat{\beta}'$, and the estimated frailty effects, $\hat{z}_i$, can be obtained. In addition, the Breslow (1972) estimator [37] is employed for estimating the cumulative baseline hazard to get $\hat{H}_0$ based on the training dataset. The predictive survival function $S_{ij}(y_{ij})$ for the observation y_{ij}^{test} of the test case is given by:

$$\hat{S}_{ij}(y_{ij}^{test}) = \exp\{-\hat{z}_i \exp\Big(x_{ij}\hat{\beta}'\Big)\hat{H}_0(y_{ij}^{test})\}. \tag{4.7}$$

Then, the RSP for the actually observed t_{ij} of the test case is defined as:

$$\hat{S}_{ij}^R(t_{ij}^{test}, \delta_{ij}, U_{ij}) = \begin{cases} \hat{S}_{ij}(t_{ij}^{test}), & \text{if } t_{ij}^{test} \text{ is uncensored, i.e., } \delta_{ij} = 1, \\ U_{ij}\, \hat{S}_{ij}(t_{ij}^{test}), & \text{if } t_{ij}^{test} \text{ is censored, i.e., } \delta_{ij} = 0. \end{cases} \tag{4.8}$$

The Z-residual for t_{ij}^{test} is given by:

$$\hat{z}_{ij}(t_{ij}^{test}, \delta_{ij}, U_{ij}) = -\Phi^{-1}(\hat{S}_{ij}^R(t_{ij}^{test}, \delta_{ij}, U_{ij})). \tag{4.9}$$

Repeating these steps n times for each observation, we have n different pairs of training and test datasets and a LOOCV predictive Z-residual is computed for each observation.

For implementing cross-validation, each cluster and each value of a categorical covariate should appear at least once in both the training and test datasets. If a cluster has only one observation and it is left out as a test case, the cluster cannot appear in the training dataset. We do not calculate the cross-validatory Z-residual for such observations. This is because the information of $\hat{z}_i$ for such a cluster with only one observation cannot be obtained from the training dataset. Similar requirements are enforced for all categorical covariates. The cross-validatory Z-residuals for such observations are set to be NA in our implementation.

The k-fold CV method splits the full dataset into k groups of observations and the groups are of approximately equal size. One group is left out to form the test dataset, and the dataset of the remaining $k-1$ groups is used as the training dataset for fitting the shared gamma frailty model. The estimates, $\hat{\beta}'$, $\hat{z}_i$, and $\hat{H}_0$, can be obtained from the fitted model, and the predictive Z-residuals are calculated for the observations in the test dataset. These steps are the same as the LOOCV method described above. In splitting the dataset into k groups, we try to make each group contain a similar number of observations of each cluster and of each category of categorical covariates. In creating the cross-validation folds, we ensure that the set of cluster identities and the values of each categorical covariate in the test dataset is a subset of the corresponding values in the training dataset. In addition, we also avoid the situation that there is no event (observed) failure time for a certain cluster or a certain category of categorical covariates, which is not allowed in fitting non-parametric Cox proportional hazard models with the `survival` package.

4.4 Simulation Studies and results

4.4.1 Detection of Non-linear Covariate Effect

In this section, we compare the performance of Z-residuals with and without cross-validation in detecting non-linear covariate effects via simulation studies. We generate failure times t_{ij} from a Weibull regression model with shape parameter α=3 and scale parameter λ=0.007, as follows:

$$t_{ij} = \left(\frac{-\log(v_{ij})}{\lambda z_i \exp\left(x_{ij}^{(1)} + \beta_2 \log\left(x_{ij}^{(2)}\right) + 0.5 x_{ij}^{(3)}\right)} \right)^{1/\alpha}, \tag{4.10}$$

where $i = \{1,\ldots, 10\}$ and $j = \{1, \ldots, m\}$ and v_{ij} is simulated from Uniform(0, 1). The censoring times C_i is simulated from an exponential distribution, exp(θ), where θ is set to have censoring rates (c) approximately equal to 50%. The three covariates are generated as follows: $x_{ij}^{(1)}$ from Uniform(0, 1), $x_{ij}^{(2)}$ from positive-Normal(0, 1), and $x_{ij}^{(3)}$ from Bern(0.25). The frailty term is generated from the gamma distribution with a mean of 1 and a variance of 0.5.

We consider fitting a shared frailty gamma model assuming $h_{ij}(t) = z_i \exp\left(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)}\right) h_0(t)$ as a wrong model, and fitting a shared frailty gamma model assuming $h_{ij}(t) = z_i \exp\left(\beta_1 x_{ij}^{(1)} + \beta_2 \log\left(x_{ij}^{(2)}\right) + \beta_3 x_{ij}^{(3)}\right) h_0(t)$ as the true model. We first visualize the difference of the Z-residuals with and without cross-validation on a single dataset generated with a strong non-linearity covariate effect ($\beta_2 = -2$). The dataset has 10 clusters of 50 observations (the total sample size $n = 500$). The scatterplots of Z-residuals against the covariate $x_{ij}^{(2)}$ are shown in Fig. 4.1, in which the two rows show the true and the wrong models and the three columns show three different methods for computing Z-residuals — the No-CV, 10-fold and LOOCV methods respectively. Under the true model, the scatterplots of the three types of Z-residuals are randomly scattered

without exhibiting any pattern and they are mostly within the interval (-3, 3). Note that most Z-residuals are concentrated on the left side of the x-axis because $x_{ij}^{(2)}$ was simulated from the positive Normal(0, 1). Under the wrong model, all the scatterplots of the three types of Z-residuals show a non-linear pattern. However, we see that, for one observation, the Z-residuals computed by cross-validation methods are near the value 6, but the corresponding No-CV Z-residual is near 3. In addition, there are more cross-validatory Z-residuals greater than 3 than No-CV Z-residuals. This is an indicator of the conservatism of the No-CV method. The QQ plot of the three types of Z-residuals under the true model aligns nearly perfectly with the 45° straight line in the appended Fig. 4.9. Under the wrong model, the QQ plot of the No-CV Z-residuals aligns with the diagonal line; however, the 10-fold and LOOCV Z-residuals show more severe deviations from the 45° straight line in the upper tail, demonstrating the increased power of the cross-validatory Z-residuals in detecting non-linear covariate effects compared to the No-CV Z-residuals.

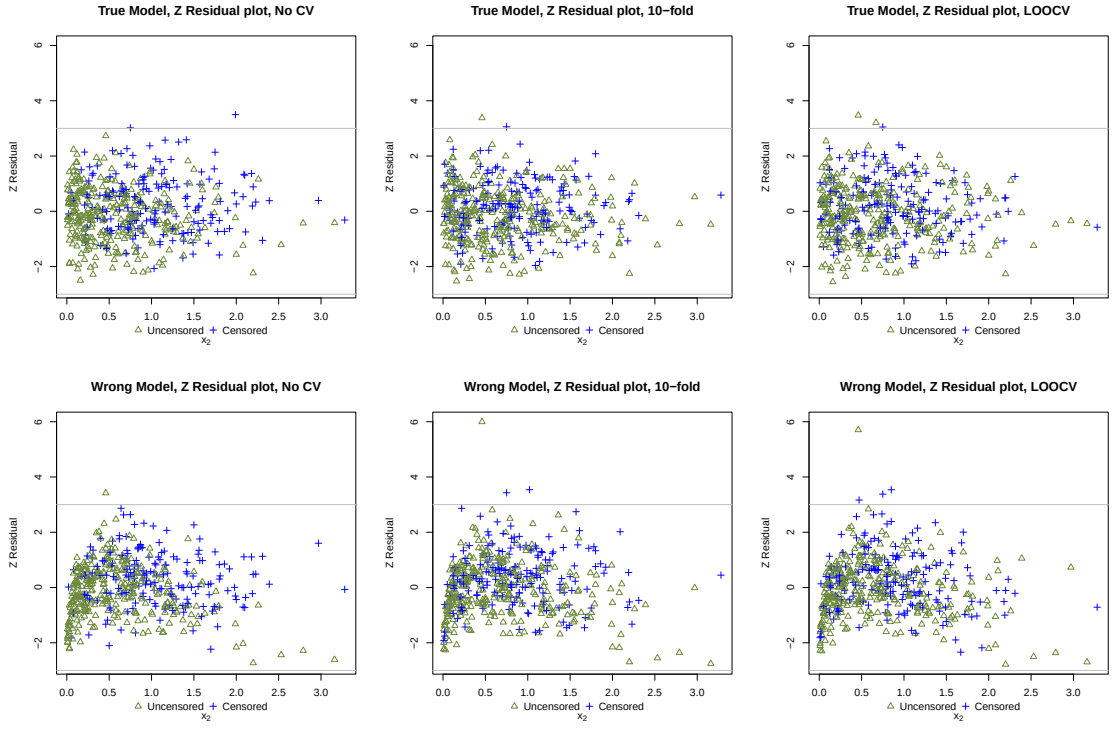


Figure 4.1: The scatterplots of the No-CV, 10-fold and LOOCV Z-residuals for a simulated dataset with non-linear covariate effect, described in Section 4.4.1. The sample size is 500 (10 clusters of 50 observations), the censoring percentage is 50%, and the β_2 for $\log\left(x^{(2)}\right)$ is set to -2. The gray horizontal lines indicate the values 3 and -3. The green points are event times and the blue points are censored times.

We used multiple simulated datasets to investigate the difference between Z-residuals with and without cross-validation when they are used in GOF tests. We apply the Shapiro–Wilk (SW) test to check the normality of the three types of Z-residuals for checking the overall GOF of fitted models. For this investigation, we generated 1000 datasets with 10 clusters of equal size, each having m observations, with m varying in the set of $\{10, 20, \ldots, 100\}$. We also set two different values for β_2 (-2 and -1) for representing strong and moderate non-linear covariate effects. The model rejection rate is estimated by the proportion of SW test p-values less than 0.05 in the 1000 datasets. We also calculated the mean of the SW p-values in the 1000

datasets for comparing the difference with and without cross-validation. Fig. 4.2 presents the results for the scenario with a strong non-linearity effect ($\beta_2 = -2$), where the three columns correspond to the No-CV, 10-fold, and LOOCV Z-residuals respectively. The first row of Fig. 4.2 displays the model rejection rates of the SW test under the true (blue lines) and the wrong (red lines) models. Under the true model, the model rejection rates of the No-CV Z-residuals are close to but slightly lower than the nominal level of 0.05 for all scenarios. By contrast, the SW tests with the 10-fold and LOOCV Z-residuals under the true model have slightly higher type-I error rates than the nominal level of 0.05 (explained below) when the sample size is smaller than 400, but the type-I error rates are close to 0.05 as the sample size increases. More importantly, the powers of the SW tests with the No-CV Z-residuals are very low and significantly lower than the corresponding powers of the SW tests with the 10-fold and LOOCV Z-residuals in all scenarios. Figure 4.2 also shows that the performances of the SW tests with the 10-fold and LOOCV Z-residuals are very similar. This finding is practically important because the computation of LOOCV Z-residuals is much more time-consuming. The panels in the second row of Figure 4.2 present the means of SW test p-values. We see that the means of SW p-values of the 10-fold and LOOCV Z-residuals under the wrong models are remarkably smaller than those of the No-CV Z-residuals. We also observe that the gap between the means of SW p-values under the true and wrong models is significantly larger for cross-validatory Z-residuals than No-CV Z-residuals. In summary, these results suggest that the SW tests with cross-validatory Z-residuals are more powerful in detecting the nonlinear covariate effects than those with No-CV Z-residuals.

We have seen that the SW tests with cross-validatory Z-residuals have slightly larger type-I error rates than the nominal level when the sample size is small (Figure 4.2). We postulate that the elevation is due to the finite-sample error in estimating the model parameters. In particular, there might be large errors in the estimation of frailties, which can receive information only from the observation within each cluster. In theory, the exact normal distribution for Z-residuals holds when they are calculated with the true model *with the true parameters*. Given a dataset with a finite sample size, there are still sampling errors in estimating the model parameters, even though the fitted model has the correctly specified form. As a result, the fitted model is not exactly the true model for the dataset. To illustrate the difference between the fitted and the true models, we calculated the R^2 value between the survival probabilities calculated with the parameters estimated with the three different methods (No-CV, 10-fold, LOOCV) and the true survival probabilities calculated with the true generating parameters. The panels in the third row of Figure 4.2 show the average of the R^2 values in the 1000 datasets and also in different cross-validation folds for each simulation setting. We see that the R^2 for the cases with small sample sizes is substantially smaller than the cases with large sample sizes. Therefore, it is reasonable that the type-I error rates of the SW tests with cross-validatory Z-residuals are larger than the nominal level of 0.05. Figure 4.10 in the appendix provides the results based on the scenario with a moderate non-linear covariate effect ($\beta_2 = -1$). The results are generally consistent with the scenario with a strong non-linear covariate effect, but the model rejection rates and means of the SW p-values are slightly lower when the wrong model is fitted to the datasets.

We also use the area under the ROC curve (AUC) to summarize the discriminative powers of SW test p-values and use them to compare the three types of Z-residuals. The AUC measures the difference between two groups of 1000 SW test p-values, one from fitting the true model and the other from fitting the wrong model. When the AUC is high, the SW p-values are well separated between the two groups, indicating that they are discriminative for discerning adequate and inadequate models. Figure 4.3 shows the AUC values for the scenarios with strong and moderate non-linear covariate effects in the left and right plots respectively. The AUC of all three methods increases as the sample size increases, and the AUC values of the 10-fold and

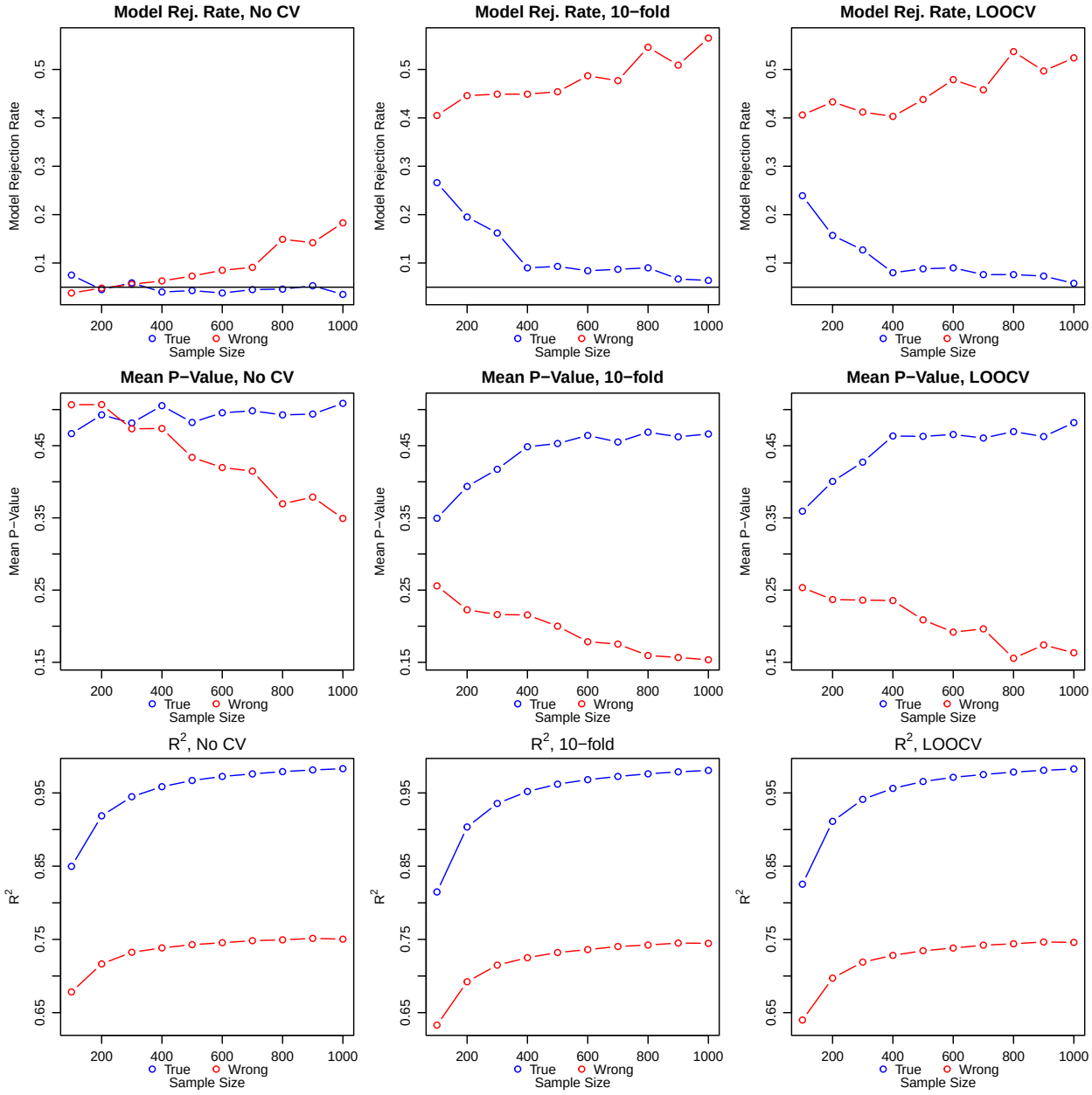


Figure 4.2: Comparison of model rejection rates (proportions of SW test p-values ≤ 0.05) and the means of SW p-values with Z-residuals based on the No-CV, 10-fold and LOOCV methods for detecting the non-linear covariate effect. The percentage of censoring is 50% and the true regression coefficient for the nonlinear covariate, $\log(x_2)$, is -2. The plots in the third row show the values of R^2 for measuring the agreement between the survival probabilities calculated with the fitted models and the survival probabilities calculated with the true generating models.

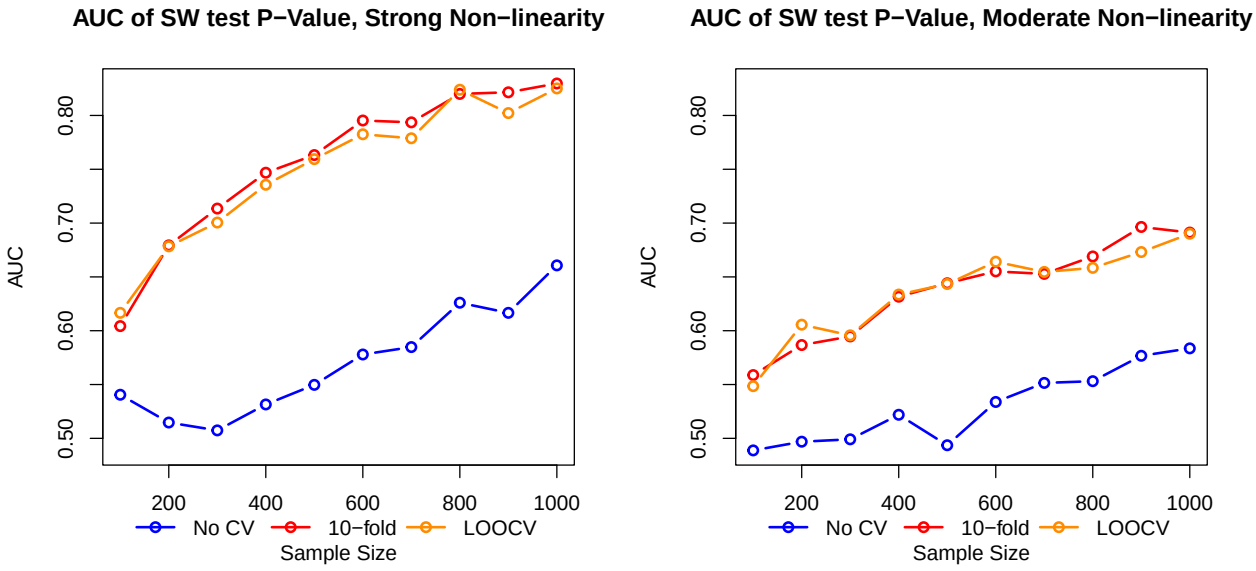


Figure 4.3: Comparison of the AUC values of SW test p-values based on Z-residuals computed with the No-CV, 10-fold and LOOCV methods for simulation datasets with non-linearity effects.

LOOCV Z-residuals are very close to each other. More importantly, the AUC values of 10-fold and LOOCV Z-residuals are consistently much higher than the corresponding values of No-CV Z-residuals. Furthermore, we notice that the superiority of cross-validatory Z-residuals does not diminish when the sample size increase, at least up to 1000. This finding is fairly remarkable as we might think that the bias due to the double use of the dataset may disappear when the sample size is sufficiently large. In summary, the SW p-values calculated with cross-validatory Z-residuals are much more discriminative in separating the proper and improper models than those calculated with No-CV Z-residuals.

4.4.2 Detecting Outliers

In this section, we compare the performance of the Z-residuals with and without cross-validation in identifying outliers via simulation studies. We generate a clean dataset from a Weibull model and then add jitters to create a corresponding contaminated dataset, for which we know the identities of outliers. For the clean datasets, we generate the true failure times from a Weibull regression model with shape parameter α=3 and scale parameter λ=0.007 as follows:

$$t_{ij} = \left(\frac{-\log(v_{ij})}{\lambda z_i \exp\left(x_{ij}^{(1)} - 2x_{ij}^{(2)} + 0.5x_{ij}^{(3)}\right)} \right)^{1/\alpha}, \tag{4.11}$$

where $i = \{1,\ldots, 10\}$ and $j = \{1, \ldots, m\}$ and v_{ij} is simulated from Uniform (0, 1). The censoring times C_{ij} is simulated from an exponential distribution, exp(θ), with θ being set to obtain censoring rates approximately equal to 50%. The three covariates are generated as follows: $x_{ij}^{(1)}$ from Uniform(0, 1), $x_{ij}^{(2)}$ from Normal(0, 1), and $x_{ij}^{(3)}$ from Bern(0.25). The frailties are generated from the gamma distribution with a mean of 1 and a variance of 0.5. The jitters added to outliers are generated from $\max(w, e)$, where e is a random number from exp(1) and the value w is set to 2 or 4 for indicating moderate and strong jitters respectively. The value

w is introduced to ensure that the jitters are at least greater than w. We also consider two different schemes of adding jitters to clean datasets. One is adding to randomly selected 10% event times, and the other is adding to a random selection of 10 event times. Note that the contaminated failure times may not always appear excessively large if the failure time before contamination is small enough. We repeatedly simulate 1000 datasets with 10 clusters of m observations, where the cluster size m is varied in the set of $\{10, 20, \ldots, 100\}$ for investigating how the performance of Z-residuals depends on the cluster size. To these simulated datasets, we fit the shared frailty gamma model assuming $h_{ij}(t) = z_i \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)}\Big) h_0(t)$, which is the true model generating clean datasets.

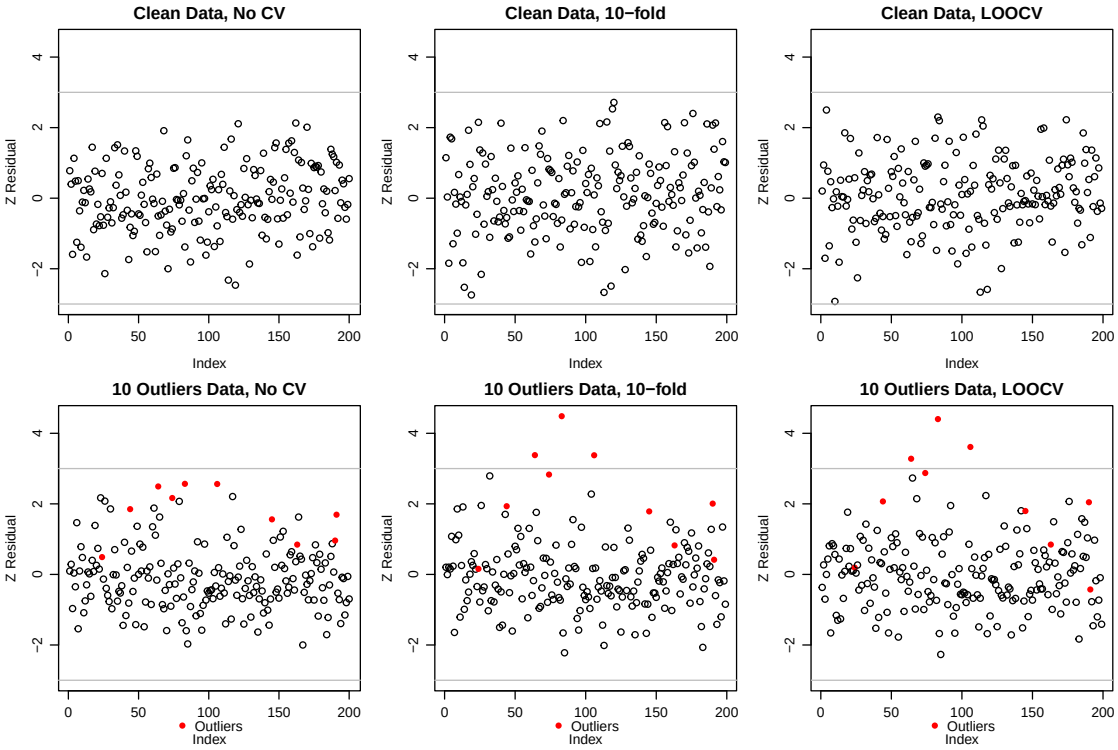


Figure 4.4: Comparison of the performance of the No-CV, 10-fold, and LOOCV Z-residuals in detecting outliers on a pair of clean and contaminated datasets. The datasets have 10 clusters with 20 observations in each.

We first visualize the difference of Z-residuals with and without cross-validation on a pair of clean and contaminated datasets with the cluster size $m = 20$. Strong jitters are added to 10 randomly selected failure times for generating the corresponding contaminated dataset. Fig. 4.4 displays the residual plots for the clean and contaminated datasets in the first and second rows respectively. The red points indicate the outliers with their failure times being added with jitters. The Z-residuals for the clean dataset are mostly bounded between -3 and 3 as standard normal variates without any unusual patterns. For the contaminated dataset, all the No-CV Z-residuals are bounded between -3 and 3, which means that they fail to detect the outliers if we declare outliers when the Z-residual is out of $(-3, 3)$; by contrast, the cross-validatory Z-residuals of three outliers (red points) fall out of the interval (-3, 3). This comparison suggests that the cross-validatory Z-residuals have increased powers in detecting outliers even though not all the outliers could be detected because their failure times are still not excessive to the model after being added with jitters.

We use multiple simulated datasets to investigate the performance of the SW tests based on the three

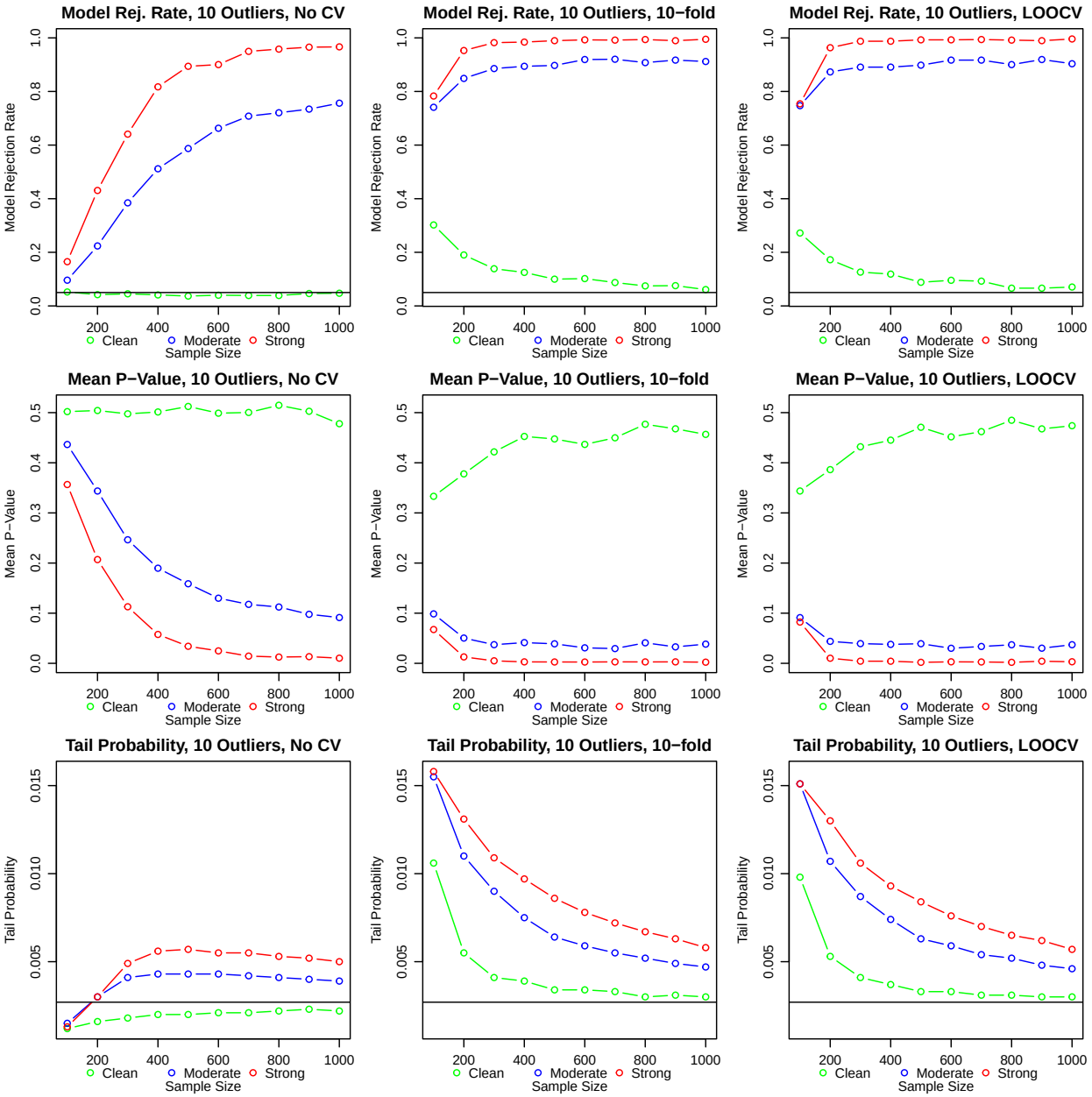


Figure 4.5: Comparison of model rejection rates based on the SW test p-values ≤ 0.05, the mean of SW p-values, and the tail probability of the No-CV, 10-fold, and LOOCV Z-residuals for the datasets with 10 outliers. The horizontal lines for the model rejection rate show the nominal type-I error rate of SW tests under the true model, ie, 0.05. The horizontal lines for the tail probability show the expected value for clean datasets, ie, $P(|Z|> 3) = 0.0027$ where $Z \sim N(0, 1)$.

types of Z-residuals in identifying model inadequacy for contaminated datasets. We fit the true model for clean datasets to both clean and contaminated datasets. This model is adequate for clean datasets but is inadequate for contaminated datasets, which require a more sophisticated model. Therefore, we expect that the SW test should have a high chance (power) of rejecting this model for contaminated datasets. As in Section 4.4.1, we use 1000 simulated datasets for each simulation setting to calculate the proportion of the SW test p-values less than 0.05 and calculate the mean of SW p-values. Figure 4.5 shows the results for the scenario with 10 strong outliers. As displayed in Figure 4.5, the model rejection rates of No-CV Z-residuals for clean datasets (green line) remain at the nominal level of 0.05 for all scenarios. However, for contaminated datasets, the model rejection rates (powers) of No-CV Z-residuals are much lower than the corresponding powers of 10-fold and LOOCV Z-residuals; the reduction in powers is substantial when the sample size is less than 300, for example from about 0.8 to about 0.2 when $m = 10$. We also notice that the type-I error rates (for the clean datasets) of the SW tests with cross-validatory Z-residuals are slightly higher than the nominal level of 0.05 when the sample size is small; nevertheless, they approach 0.05 as the sample size increases. From the second row of Figure 4.5, we also observe that the means of the SW p-values of the No-CV Z-residual are much higher than those of cross-validatory Z-residuals. For investigating the performance in outlier detection, we also calculate a tail probability about Z-residuals, which is the proportion of Z-residuals with absolute values greater than 3, which is often used to identify outliers in practice. The plots in the third row of Figure 4.5 show the means of the tail probabilities in 1000 simulated datasets under different simulation scenarios. The tail probabilities of No-CV Z-residuals for clean datasets are all below the expected value of 0.0027, which is $P(|Z|> 3)$ where $Z \sim N(0, 1)$. More importantly, we see that the tail probabilities of 10-fold and LOOCV Z-residuals for the contaminated datasets are much higher than those of No-CV Z-residuals, and the tail probabilities of 10-fold and LOOCV Z-residuals for clean datasets converge to the expected tail probability — 0.0027. Figure 4.11 in the Appendix displays the results for the scenarios in which contaminated datasets have 10% outliers, and the results are consistent with the scenarios with 10 outliers.

In order to evaluate the discriminative abilities of SW test p-values, we use the AUC to measure the difference of the SW test p-values between two groups - one from clean datasets and the other from contaminated datasets. The results are presented in the four plots (Fig. 4.6) that correspond to four simulation scenarios (combinations of two different levels of jitters and two different schemes for adding jitters). Across all scenarios, we observe that the AUC values of 10-fold and LOOCV Z-residuals are significantly higher than those of No-CV Z-residuals. When the sample size is around 100, the AUC values of No-CV Z-residuals are near 0.5, indicating no discriminative power, whereas the corresponding AUC values of 10-fold and LOOCV Z-residuals are approximately 0.8. Additionally, we notice that the difference in AUC values between Z-residuals with and without cross-validation diminishes to 0 as the sample size increases for three of the four scenarios. However, in the scenario where the number of moderate outliers is fixed at 10, the gap remains visible even when the sample size is 1000.

Finally, we compare the sensitivity and false positive rate (FPR) in detecting outliers using Z-residuals with and without cross-validation. Our rule for identifying an outlier is that the absolute value of its Z-residual is greater than 3. Given this rule, the sensitivity is the proportion of the true outliers that are correctly identified as outliers, and the FPR is the proportion of non-outliers that are falsely identified as outliers. Figure 4.7 shows the sensitivities and FPRs for the four simulation scenarios as we consider for Figure 4.6. Clearly, we see that 10-fold and LOOCV Z-residuals have much higher sensitivities and almost the same FPRs when they are used to detect true outliers compared to the No-CV Z-residuals, for all the considered

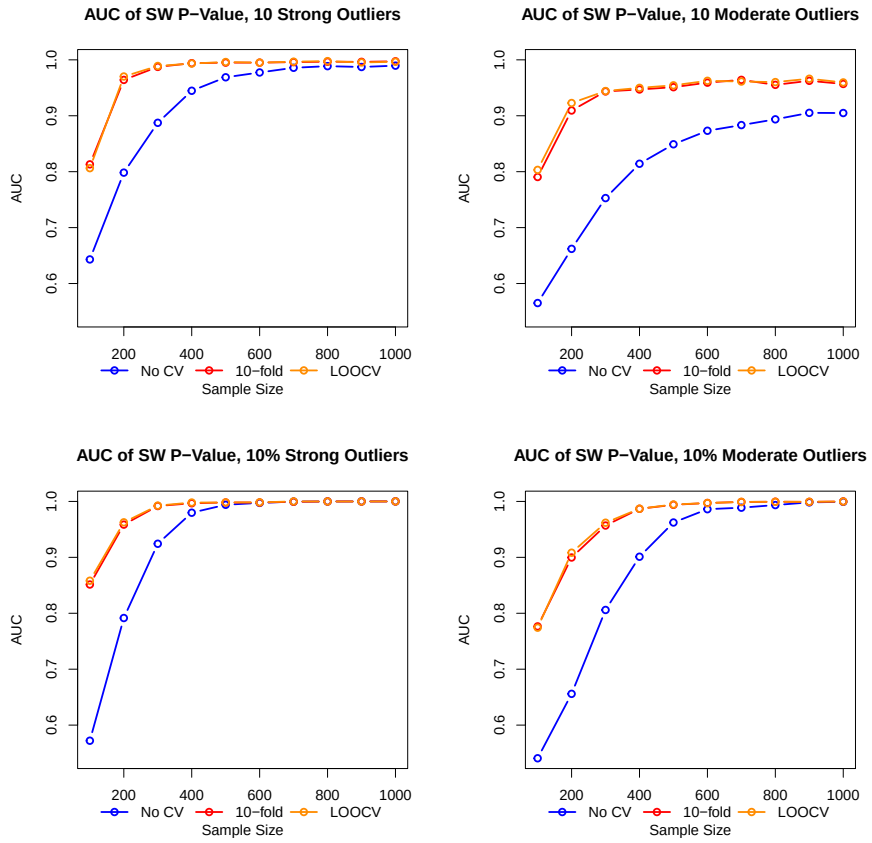
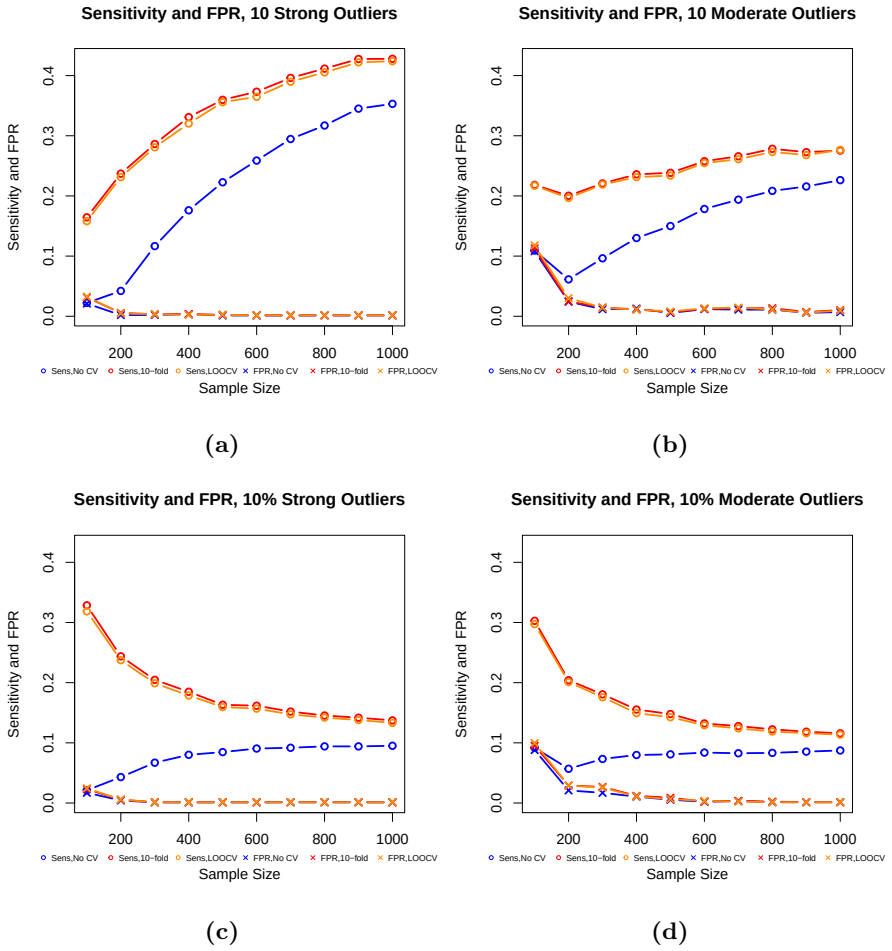


Figure 4.6: Comparison of the AUC values of SW test p-values based on Z-residuals computed with the No-CV, 10-fold and LOOCV methods for simulation datasets with outliers.

four scenarios. This comparison demonstrates clearly the advantage of using cross-validatory Z-residuals for the purpose of identifying outliers, although we previously see that the SW tests based on cross-validatory Z-residuals have a slight elevation of type-I error rates. Interestingly, we see that the sensitivity of 10-fold and LOOCV Z-residuals increases as the sample size increase when the number of outliers is fixed at 10, but it decreases and converges to a value of about 0.1 when the percentage of outliers is fixed at 10%.



(a) **(b)**

(c) **(d)**

Figure 4.7: Comparison of the sensitivities (points with ∘) and the false positive rates (points with ×) in detecting outliers using No-CV, 10-fold, and LOOCV Z-residuals.

4.5 A Real Data Example

In this section, we will demonstrate the effectiveness of the cross-validatory Z-residuals in identifying outliers in a real data application studying kidney infection [74]. The dataset consists of 38 kidney patients using a portable dialysis machine, and the times of the first and second recurrences of the kidney infection are recorded for these patients. Each survival time is defined as the time until infection since the insertion of the catheter. The same patient is considered as a cluster because of shared frailty describing the common patient's effect. If a catheter is removed for reasons other than infection, the observation is considered censored. The censoring percentage is 24%. The dataset contains 38 patients (cluster), and each patient has

exactly two observations, with a total sample size of 76. This data has often been used to illustrate a shared frailty model. More details on this dataset can be found from [74].

This study was deemed exempt from ethics approval since the data utilized were publicly available in the paper [74], and no personally identifiable information was collected or used. This research adhered to the principles outlined in TCPS 2-2nd Edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans [64].

We fit a linear shared gamma frailty model with three covariates — age in year, sex of male or female, and four different disease types (0=GN, 1=AN, 2=PKD, 3=Other) to the recurrence failure times, with details given in Table 4.2 in the Appendix. The fitting is done with the `coxph` function in the `survival` package. Table 4.1a shows the estimated regression coefficients, the corresponding standard errors, and p-values for the covariate effects from fitting the shared gamma frailty model with the full dataset. The results shown in Table 4.1a indicate that the two covariates, sex and disease type of PKD, are significantly associated with the hazard of recurrence of kidney infection.

Table 4.1: Parameter estimates of three shared gamma frailty models fitted with the kidney infection dataset. The tables (4.1b) and (4.1c) show the estimates for two subsets of the original datasets with two and three cases removed as they are identified as outliers with LOOCV Z-residuals.

(a) The original dataset

Covariate	$\hat{\beta}$	SE	p-value
Age	0.003	0.011	0.775
Sex:Male	1.480	0.358	0.000
D:GN	0.088	0.406	0.829
D:AN	0.351	0.400	0.380
D:PKD	-1.430	0.631	0.023
Frailty			0.933

(b) Exclusing two outliers

Covariate	$\hat{\beta}$	SE	p-value
Age	0.007	0.011	0.530
Sex:Male	2.117	0.400	0.000
D:GN	0.359	0.406	0.380
D:AN	0.349	0.407	0.390
D:PKD	-0.797	0.638	0.210
Frailty			0.940

(c) Excluding three outliers

Covariate	$\hat{\beta}$	SE	p-value
Age	0.012	0.011	0.280
Sex:Male	2.120	0.402	0.000
D:GN	0.727	0.415	0.080
D:AN	0.319	0.404	0.430
D:PKD	-0.802	0.636	0.210
Frailty			0.940

We calculated Z-residuals and CS residuals using the No-CV and LOOCV methods for this dataset. We only considered the LOOCV method since the 10-fold CV and LOOCV Z-residual methods perform very similarly as shown in the simulation studies and the computational burden to implement LOOCV is not a concern due to the small sample size. Figure 4.8 shows the residual diagnosis results for the original kidney infection dataset without removing outliers. The first and second columns of Figure 4.8 present the scatterplots versus the index and the QQ plots of the Z-residuals computed with the No-CV and LOOCV methods. The No-CV Z-residuals are mostly between -3 and 3 and the QQ plot of NO-CV Z-residuals aligns well with the 45° straight line. The SW p-value of No-CV Z-residuals is about 0.70 as shown in the QQ plot. Such a large p-value indicates a good fit of the model to the dataset. In summary, the diagnosis results with No-CV Z-residuals suggest that the shared frailty model appears appropriate for the dataset and no outlier is identified. However, the scatterplot of LOOCV Z-residuals shows that the Z-residuals of the two cases labelled with numbers 20 and 42 are greater than 3. We can consider these two cases as outliers for the shared frailty model. The QQ plot of LOOCV Z-residuals shows a large deviation from the 45° straight line, which is clearly caused by the large Z-residuals of the two outliers. The SW p-value of LOOCV Z-residuals is also very small — less than 0.01, as shown in the QQ plot. In summary, the diagnosis results with LOOCV Z-residuals suggest that the fitted shared frailty model is inadequate for this dataset and there are two cases with excessive Z-residuals, which are identified as outliers for this model.

Compared to all the raw infection times as shown in the appended Figure 4.12, the infection time of case 42 is the highest value among all but does not appear very outstanding; the infection time of case 20 is near the median of all infection times, hence, does not appear outlying at all. This observation illustrates the difference between the concepts of outliers relative to raw observations and relative to a fitted model.

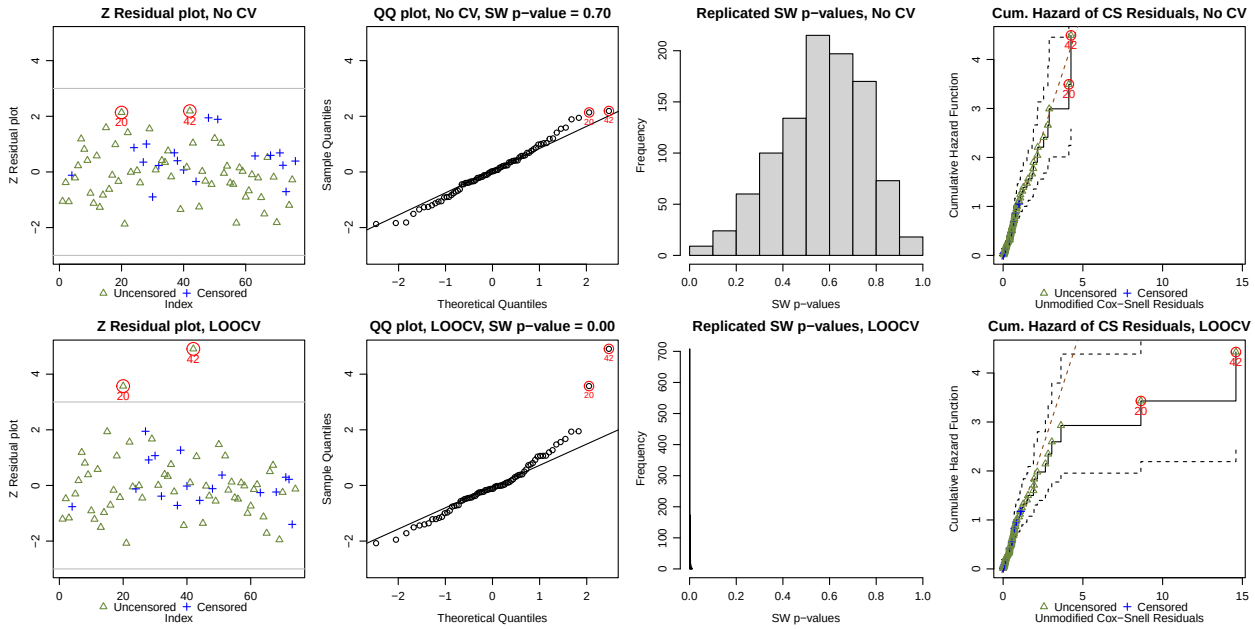


Figure 4.8: Scatterplots and QQ plots of No-CV and LOOCV Z-residuals of the fitted shared frailty models based on the original kidney infection dataset. The third column presents the histograms of 1000 replicated SW p-values of Z-residuals The fourth column shows the CS residuals computed with the No-CV and LOOCV methods.

Z-residual is a monotone transformation of the tail (or survival) probabilities of the conditional distribution of failure time given covariates (see equation (4.6)). Therefore, the identification of outliers based on Z-residuals has considered covariate effects. However, the identification of outliers by examining only raw failure times does not consider covariate effects; in other words, it is based on a model with only the intercept term.

There is randomness in the Z-residuals of the censored observations. For the same dataset, we can produce different sets of Z-residuals with different random numbers. Therefore, we would like to replicate a large number of realizations of Z-residuals to see the robustness of the above diagnosis. The third column of Figure 4.8 displays the histograms of 1000 SW test p-values, each given by a set of No-CV or LOOCV Z-residuals. The histograms show that more than 95% of the SW p-values of No-CV Z-residuals are larger than 0.05; however, 100% of the SW p-values of LOOCV Z-residuals are smaller than 0.05. Therefore, the judgment of the misspecification of the shared frailty model for the dataset is not incidental based on a particular set of LOOCV Z-residuals but a consistent conclusion under large-scale replications of Z-residuals.

To further verify the above diagnosis results and illustrate the effect of cross-validation in residual diagnostics, we also compute CS residuals with both No-CV and LOOCV methods and plot their cumulative hazard functions (CHFs) in the fourth column of Figure 4.8. The CHF of No-CV CS residuals aligns well with the 45° straight line, indicating a good model fit for the dataset. However, the CHF of the LOOCV CS residuals deviates from the 45° straight line in the upper tail, suggesting the inadequacy of the fitted model to the dataset. The conclusion for checking the model adequacy with CS residuals is consistent with the diagnosis results with Z-residuals. Nevertheless, we notice that the diagnosis with Z-residuals provides more information regarding the nature of the discrepancy of the inadequate model — the existence of outliers, as well as a quantitative measure of the statistical significance of the model departure.

Finally, we consider deleting the two outliers (cases 42 and 20) from the original kidney infection dataset

and then re-fit the linear shared gamma frailty model. Table 4.1b shows the covariate DiseasePKD is no longer statistically significant at the level of 5%, and the effect size for the covariate sex becomes larger. The differences in Table 4.1a and 4.1b indicate that parameter estimation and inference may be greatly affected by including outliers, which highlights the importance of model diagnosis and outlier detection in practical data analysis. The appended Figure 4.13 presents the results of residual diagnosis after excluding these two outliers. The LOOCV Z-residual diagnosis results indicate that the refitted model is a fairly good model for the dataset without cases 42 and 20. Nevertheless, we notice that case 15 has a Z-residual marginally greater than 3. Although case 15 may not be of great concern, as most of the SW p-values of LOOCV Z-residuals show are greater than 0.05, we refit the model after further removing case 15. Table 4.1c shows the parameter estimates based on the kidney infection dataset after excluding the three cases, which are similar to those in Table 4.1b. The Z-residual diagnosis, as shown in appended Figure 4.14, neither suggests evidence that the model fitted with the three cases removed is inadequate for the dataset nor identifies an outlier for the model.

4.6 Discussions and Conclusions

Residual diagnosis plays a critical role in the model-building process for validating the correctness of a fitted model. However, residuals are typically calculated based on the model fitted to the full dataset, without using a strategy to split the dataset into different subsets for model fitting and validation. The double use of the dataset for model fitting and validation might lead to conservatism in model diagnosis, leading to the reduced power of detecting inadequate model fit and identifying outliers for the model. To the best of our knowledge, cross-validation is rarely used in residual diagnosis for survival analysis. In this paper, we developed cross-validation methods to compute Z-residuals for detecting model inadequacy and identifying outliers in the context of shared frailty models. We compare the performance of cross-validatory (10-fold and LOOCV) Z-residuals and No-CV Z-residuals for the purpose of the overall GOF test and outlier detection. Our simulation studies and the application to a real dataset demonstrate that the residual diagnosis without cross-validation tends to be conservative for detecting model misspecification due to the double use of the data, and the cross-validation methods can improve the power of the SW-test with Z-residuals in detecting model inadequacy and improve the power of Z-residuals in identifying outliers.

Our simulation studies also reveal that the cross-validation may cause a slight elevation of type-I error rates in SW tests with Z-residuals. As we explained with the R^2 between the survival probabilities calculated with the fitted models and the survival probabilities calculated with the true generating models, the elevation might be caused by inaccuracy in estimating the parameter, in particular, the estimation of the frailties in small cluster size situations. For such situations, the model fitting algorithms for shared frailty models could improve on estimating the frailties, for example, with a stronger penalization for the frailties. Alternatively, another direction is to work on improving the methods for computing the cross-validatory Z-residuals or the methods for conducting SW tests with Z-residuals, with the goal of obtaining Z-residuals that are less aggressive in rejecting models. Marginalizing the frailties when we calculate the randomized survival probability may be a solution. If we marginalize the frailties, the cluster size may have a smaller impact on the computation of Z-residuals; moreover, the restriction that the size of each cluster must be greater than 1 could be resolved. A comparison of the methods for computing Z-residuals with or without marginalizing the frailties is an interesting topic for future work. Lastly, in the present study, we focused on investigating the performance of cross-validatory Z-residuals in diagnosing shared frailty models; however, the proposed

cross-validatory residuals could be more broadly applied in diagnosing other types of regression models.

4.7 Additional Figures and Tables

4.7.1 Supplementary Figures for Section 4.4.1

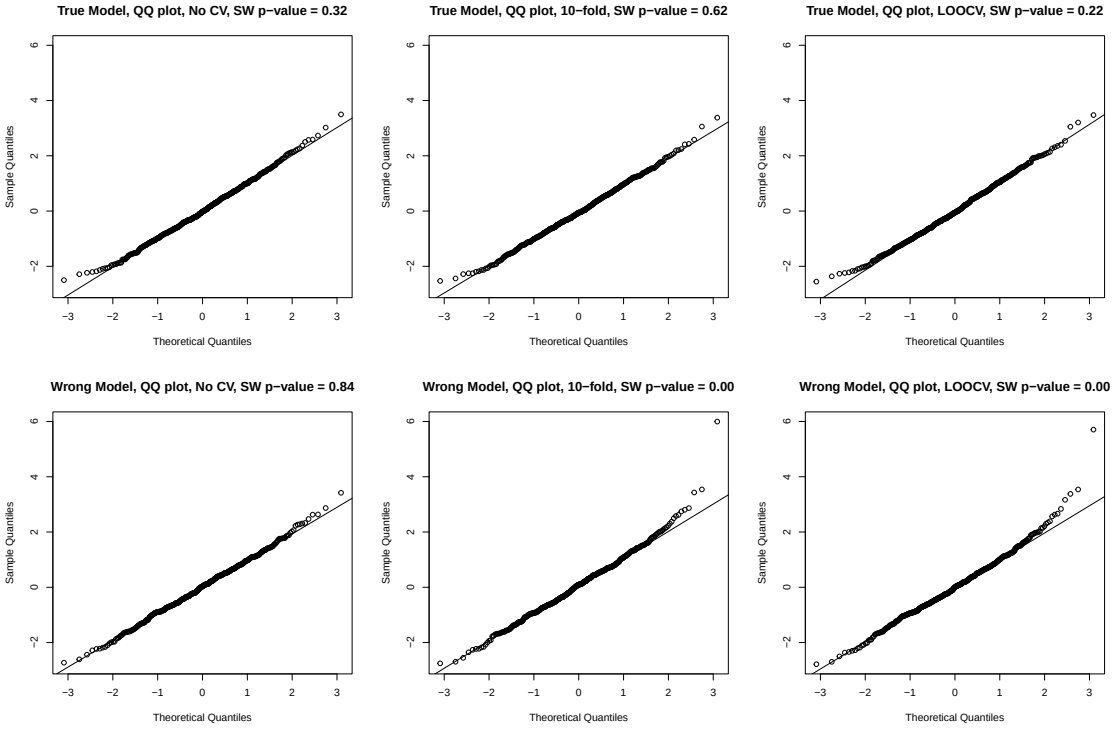


Figure 4.9: The QQ plot of the No-CV, 10-fold and LOOCV Z-residuals as a graphical tool for detecting non-linear effect in covariate with the strong non-linear association. The sample size is 500 (10 clusters of 50 observations), and the censoring percentage is 50%.

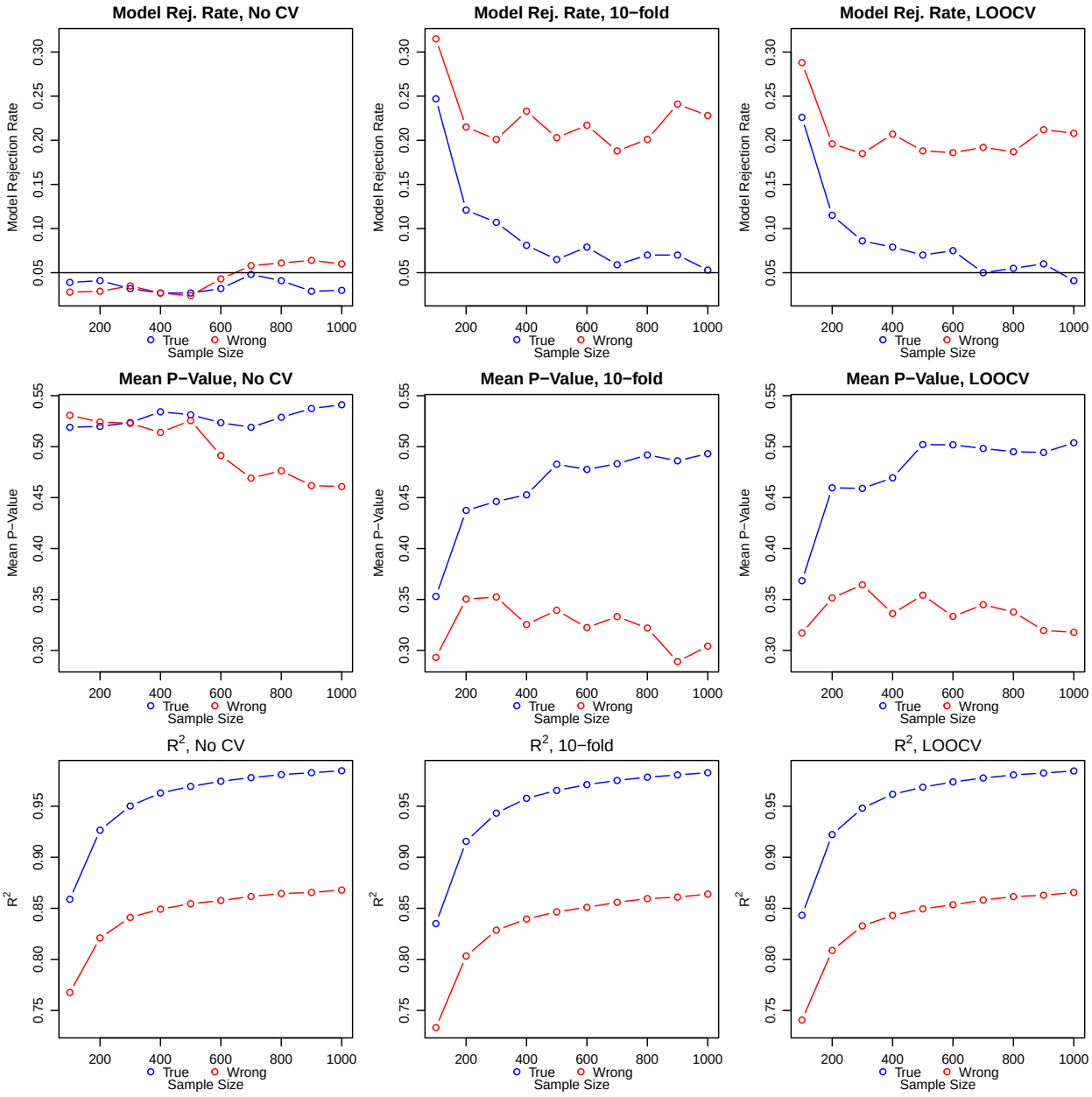


Figure 4.10: Comparison of model rejections based on SW test, the mean of SW p-values and R^2 of the No-CV, 10-fold and LOOCV Z-residuals for detecting the moderate non-linear covariate effect.

4.7.2 Supplementary Figures for Section 4.4.2

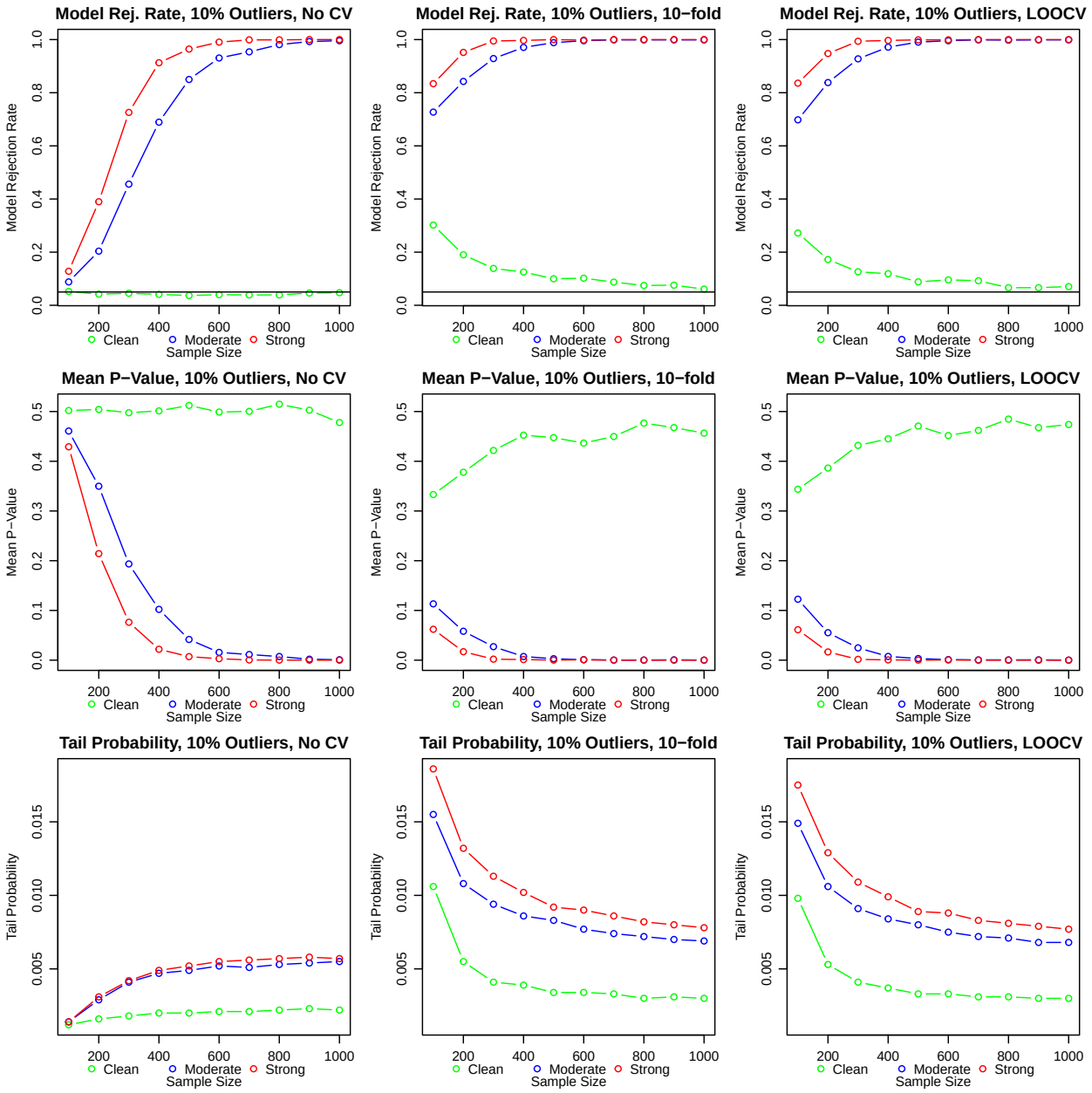


Figure 4.11: Comparison of model rejections rate based on the SW test, the mean of SW p-values and tail probability of the No-CV, 10-fold and LOOCV Z-residuals when the data are contaminated by adding 10% outliers with moderate and strong deviation from the clean data, respectively.

4.7.3 Supplementary Figures and Tables for Section 4.5

Table 4.2: Variable definitions for the kidney infection dataset.

Variable	Definition
ID	Patient number
Time	Recurrence time (days)
Status	Event indicator (1 = infection occurs; 0 = censored)
Age	Patient age (years)
Sex	Sex status (1 = male; 2 = female)
Disease	Disease Type (0 = GN; 1 = AN; 2 = PKD; 3 = other)

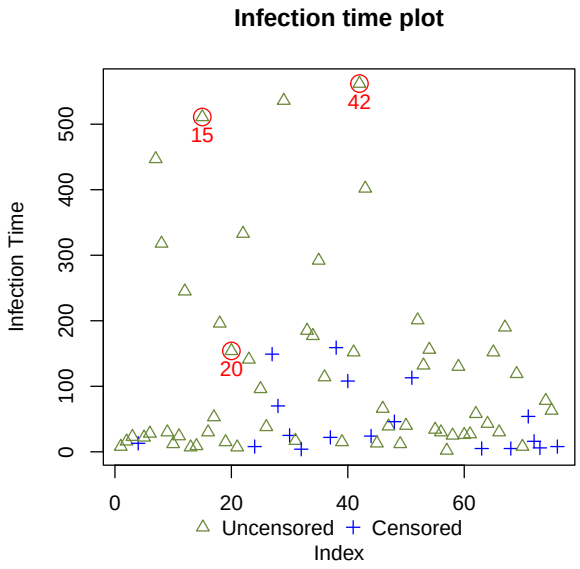


Figure 4.12: The scattered plot of infection times for the kidney infection dataset.

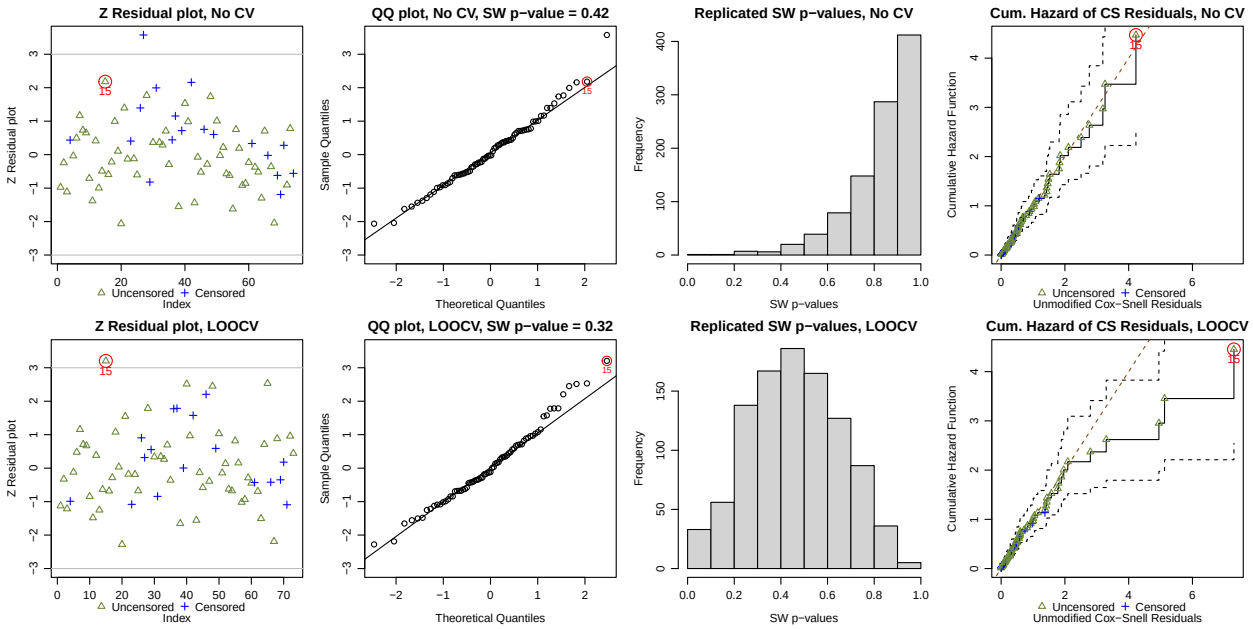


Figure 4.13: Scatterplots and QQ plots of No-CV and LOOCV Z-residuals of the fitted shared frailty models based on the kidney infection dataset with the cases 42 and 20 removed. The third column presents the histograms of 1000 replicated SW p-values of Z-residuals The fourth column shows the CS residuals computed with the No-CV and LOOCV methods.

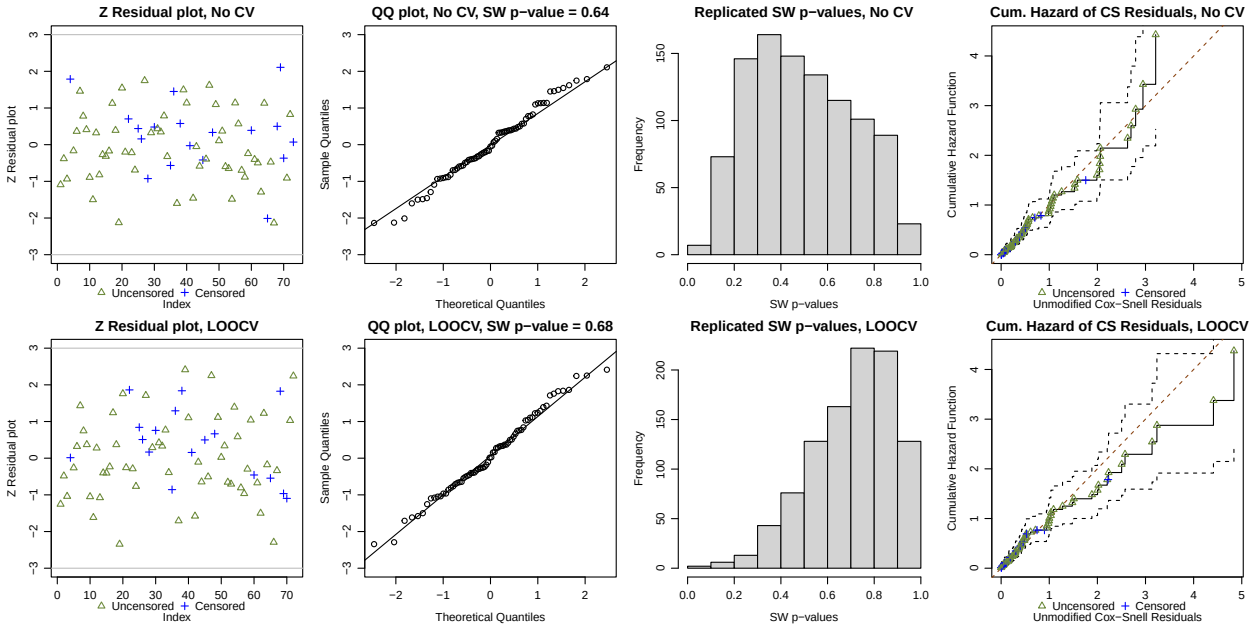


Figure 4.14: Scatterplots and QQ plots of No-CV and LOOCV Z-residuals of the fitted shared frailty models based on the kidney infection dataset with the cases 42, 20, and 15 removed. The third column presents the histograms of 1000 replicated SW p-values of Z-residuals The fourth column shows the CS residuals computed with the No-CV and LOOCV methods.

5 Z-Residual Diagnostics for Detecting Non-Proportional Hazards for Survival Models

Abstract: In Cox proportional hazard (PH) regression models, proportional hazard (PH) is a fundamental assumption, which assumes that the hazard ratio remains constant over time. Schoenfeld residuals and the related tests are the most commonly used diagnostic tools for checking the PH assumption. However, the Schoenfeld residuals based diagnostics only consider a specific case of violations of PH assumption due to the time-varying covariate effect. Therefore, the power to detect violation of PH assumption due to other reasons, such as accelerated failure time, may be low. In this paper, we propose using the Z-residual diagnostics for detecting Non-PH in a survival model. Our simulation studies show that, compared to the score tests related to Schoenfeld residuals, the tests based on Z-residuals have similar powers and type I error rates in time-varying covariate effect scenarios but they have significantly higher powers in accelerated failure time scenarios. In a real data application on identifying prognostic factors associated with time to blindness in a diabetic retinopathy study, the Z-residual diagnostic tools capture a severe violation of the PH assumption for the dataset, which is, however, not detected by Schoenfeld residuals and the related tests.

5.1 Introduction

The Cox proportional hazard regression models, developed by Cox [22], are commonly used in epidemiological and clinical research to analyze time-to-event data. The models rely on the assumption of proportional hazards (PH), which posits that the ratios of hazards remain constant or independent of time across the entire follow-up time with different predictors or covariate levels. However, in real-world applications, the PH assumption may not hold, and the hazard ratio may be time-dependent. For instance, in studies of coronary atherosclerosis [75], the hazard ratio for treatment vs. placebo is close to one in the first six months, followed by a downward trend. In another study on the Norwegian colorectal cancer prevention trial [76], the hazard ratio for flexible sigmoidoscopy screening vs. no screening is greater than one in the early stages and less than one thereafter.

These examples illustrate that the PH assumption may not always be valid, and it is crucial to examine whether it holds in a particular application. Failure to assess the PH assumption can lead to biased parameter estimates, incorrect inference, and incorrect conclusions. Therefore, it is essential to evaluate the validity of the PH assumption and consider alternative models when necessary. The most commonly used graphical diagnostic methods to evaluate the PH assumption of a Cox PH model are Schoenfeld residuals [11, 16] and scaled Schoenfeld residuals [11, 17]. Schoenfeld residuals represent the difference between the observed and expected values of the covariate(s) given the risk set at that time. The PH assumption holds if the scatterplot of Schoenfeld residuals versus survival time is flat and centred around zero without any pattern. Scaled Schoenfeld residuals [11, 17] are Schoenfeld residuals adjusted by the inverse of the covariance matrix of the Schoenfeld residuals. This scaling makes the residuals more sensitive to departures from the PH assumption,

especially in cases where the variance of the covariate changes over time or is very small. By scaling, the residuals are put on a similar scale, making it easier to compare and identify patterns of deviation from the PH assumption. The graphical diagnostics based on scaled Schoenfeld residuals commonly compare the plot of the Scaled Schoenfeld residuals versus the observed survival time. In principle, the Schoenfeld residuals are independent of time if the PH assumption holds. Therefore, a scatterplot of scaled Schoenfeld residuals that shows a random pattern around a horizontal line indicates the plausibility of the PH assumption. A numerical test, known as the Grambsch and Therneau (GT) test, has therefore been developed to assess the PH assumption by testing the null hypothesis that the regression coefficient between the scaled Schoenfeld residuals and a function of time is equal to zero. The GT test can evaluate the PH assumption for each covariate individually and/or for all covariates included in the Cox regression model [11, 16]. The GT test can be regarded as an approximation of the general score test [77]. From version 3.0, the `survival` package [5] began to use the score test for testing the PH assumption. Checking the PH assumption using the score test involves performing a goodness-of-fit (GOF) test that compares the observed values of the score function to their expected values under the null hypothesis of the PH assumption. A more recent version of the `survival` package in R computes the score test [5].

Although the above-mentioned statistical methods for checking the PH assumption are widely used, they are most appropriate for detecting violations of the PH assumption due to time-varying covariate effects [78–81]. The Cox models with time-varying effects are assumed to be the true models in these tests. However, the time-varying covariate effect in the Cox regression model is only one specific cause of violation of the PH assumption. The survival time may not necessarily be adequately modelled by a Cox regression model even with time-varying covariate effects. The accelerated failure time (AFT) model [23] incorporates a wide variety of survival time distributions, which is used in many fields as an alternative to the Cox PH model if the PH assumption is not tenable [82, 83]. However, little research has been conducted to evaluate the performance of score tests for testing PH when the survival time does not fall under a Cox regression modelling framework. We demonstrated that in the situation when the survival time does not follow the form of a Cox model but rather an AFT model, these traditional residual diagnostic tools may fail to detect Non-PH. An analogy in medical diagnostics is the diagnostic tool that can detect only breast cancer may have low power for detecting other kinds of cancers. To this end, we develop a residual diagnostic tool with a numerical test to detect Non-PH not only in the situation with time-varying covariate effects in a Cox regression model but also more broadly, under alternative modelling methods other than Cox regression model, such as AFT models.

In this chapter, we present evidence that the Z-residual diagnostic tool is highly effective in detecting Non-PH due to both time-varying covariate effects and accelerated failure times. Firstly, we conducted simulation studies to investigate the performance of the Z-residual in detecting Non-PH due to time-varying covariate effects in the Cox PH model with frailty, and compared it with score tests based on Schoenfeld residuals. Our results showed that the overall GOF tests based on Z-residuals and the non-homogeneity test to assess the group variances of Z-residuals had similar powers and type I error rates to those of the score tests. In the scenario where the survival times are simulated from AFT models, we found that the overall GOF test based on Z-residuals had a much higher power in detecting Non-PH than the score test. We also demonstrate the use of Z-residuals in testing the PH assumption in a real data application. The results based on the Z-residuals revealed that the PH assumption was violated, which was however not identified using the score test.

The rest of this chapter is organized as follows. In Section 5.2, we review the existing residuals and test methods for testing the PH assumption. In Section 5.3, we present the PH assumption diagnostics based

on Z-residuals. In Section 5.4, we conduct simulation studies to investigate the performances of Z-residuals. Section 5.5 presents the results of applying the Z-residual for detecting the Non-PH in a real data application. The conclusion is provided in Section 5.6.

5.2 Review of Existing Methods for Checking PH Assumption

5.2.1 Graphical Strategy

A common approach to evaluating the validity of the PH assumption in a Cox PH model is to plot the log-cumulative hazard functions against the logarithm of the survival time [11]. This involves grouping the survival data into levels of factors. In the case of a categorical covariate, the Kaplan-Meier estimate of the survivor function of the time $S(t)$ is calculated for each group, and the log-cumulative hazard functions $\log(H(t))$ is obtained by transforming the survivor function of the time $S(t)$ to $\log(-\log(S(t)))$. If the hazards are proportional across the different groups, the plot of the log-cumulative hazard functions against the logarithm of the survival time will exhibit constant differences and display parallel curves. The continuous covariate can be transformed into a categorical variable by cutting it into groups. Then, the plot of the log-cumulative hazard functions against the logarithm of the survival time should show approximately parallel lines across different groups if the PH assumption holds.

Although the visual inspection of the plot of log-cumulative hazard functions is simple to use, it has some drawbacks. When there are multiple groups in a categorical covariate, the plot becomes cluttered, making it challenging to detect non-proportionality [84]. Additionally, it is challenging to determine the best way to group continuous covariates as categorical variables, and the visual inspection of PH is subjective and difficult to quantify [85].

5.2.2 Schoenfeld Residuals

In this section, we review the most commonly used residual for checking the PH assumption in survival analysis. Suppose there are g groups of individuals with n_i individuals in the ith group, $i = 1, 2, \ldots, g$. All subjects in the ith group share the same frailty value z_i. Suppose t_{ij} is the true failure time for the jth individual from the ith group, which we assume to be a continuous random variable in this article, where $j = 1, 2, \cdots, n_i$. We can observe that t_{ij} is greater than a value c_{ij}, where c_{ij} is the corresponding censoring time. The observed failure times are denoted by the pair (y_{ij}, δ_{ij}), where $y_{ij} = \min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} < c_{ij})$. The observed data can be written as $y = (y_{11}, \cdots, y_{gn_g})$ and $\delta = (\delta_{11}, \cdots, \delta_{gn_g})$. The Cox PH model with shared frailty assumes a hazard function at time t, for the jth individual, $j = 1, 2, \cdots, n_i$, in the ith group, is then

$$h_{ij}(t) = z_i \exp(x_{ij}\beta) h_0(t); \tag{5.1}$$

where x_{ij} is a row vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \cdots, x_{gn_g})$; β is a column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, and z_i is the frailty term that is common for all n_i individuals within the ith group. This model implies that the hazards of the two groups remain proportional over time and the regression coefficient (β) does not change over time. To test the PH assumption in equation (5.1), we can consider the following time-varying covariate

effects models with a hazard function given by

$$h_{ij}(t) = z_i \exp\left(\sum_{p=1}^{P} x_{ijp}\beta_p(t)\right) h_0(t); \tag{5.2}$$

where $\beta_p(t) = \beta_p + g_p(t)$ is a function of time. When $g_p(t)$ differs from a constant function, the hazard ratio for the covariate changes over time, that is, the PH assumption is violated.

Schoenfeld residuals [11, 16] were originally termed partial residuals. It can overcome two disadvantages of the Cox Snell (CS) residuals [12], which depend heavily on the observed survival time and require an estimate of the cumulative hazard function. In [16], Schoenfeld residual is perceived as the difference between the observed and expected values of the x_{ijp} given the risk set at the time y_{ij}, one for each covariate in the fitted Cox regression model, defined as

$$r_{S_{ijp}} = \delta_{ij}(x_{ijp} - \hat{a}_{ijp}), \tag{5.3}$$

where x_{ijp} is the value of the pth covariate, $p = 1, 2, \cdots, P$, for the jth individual of the ith group in the study,

$$\hat{a}_{ijp} = \frac{\sum_{lh \in R(y_{ij})} x_{lhp} z_l \exp\left(x_{lh}\hat{\beta}\right)}{\sum_{lh \in R(y_{ij})} z_l \exp\left(x_{lh}\hat{\beta}\right)}, \tag{5.4}$$

and $R(t_{ij})$ is the set of all individuals at risk at time y_{ij}. If the PH assumption is met, the scatter plot of Schoenfeld residuals versus survival time should be flat, centred around zero without exhibiting any pattern. The primary drawback of Schoenfeld residuals is that it only considers a non-proportional departure in one covariate, and there is no numeric test to measure the practical and statistical significance of the degree of PH assumption violation.

Scaled Schoenfeld residuals, proposed by Grambsch and Therneau (1994) [17], are a variation of Schoenfeld residuals that are scaled by the variance of the parameter estimates [11, 17]. The purpose of scaling is to make the residuals more sensitive to departures from the PH assumption, especially in cases where the variance of the parameter estimates changes over time or is very small. By scaling, the residuals are put on a similar scale, which makes it easier to compare them and identify patterns of deviation from the PH assumption. The widely used scaled Schoenfeld residuals are defined as

$$r^{\star}_{S_{ij}} = d\text{var}(\hat{\beta}) r_{S_{ij}}, \tag{5.5}$$

where d is the number of events among all observed failure times, and $\text{var}(\hat{\beta})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. The method by [17] is widely used and adopted in the `survival` package of R software before version 3.0. It is shown [17] that the expected value of the ijth scaled Schoenfeld residual for the pth covariate is given by $E(r^{\star}_{S_{ijp}}) \approx \beta_p(t_{ij}) - \hat{\beta}_p$, where $\beta_p(t_{ij}) = \beta_p + g_p(t_{ij})$ is the value of a time-vary coefficient for x_p at the survival time of the ijth individual, and $\hat{\beta}_p$ is the estimated coefficient in the fitted Cox regression model. Diagnostics based on scaled Schoenfeld residuals commonly use a plot of the scaled Schoenfeld residuals against the observed survival time. A plot that shows a random pattern around a horizontal line without showing a trend indicates the plausibility of the PH assumption. A numerical test, called the GT test, was developed to assess the PH assumption by testing whether the $E(r^{\star}_{S_{ijp}})$ is time-dependent against the event times. More specifically, the time-varying coefficient for x_p at the survival time t of the ijth individual can be expressed by taking $\beta_p(t_{ij}) = \beta_p + g_p(t) = \beta_p + \nu_p(t_{ij} - \bar{t})$,

where ν_p is an unknown slope and $\bar{t}$ is the average of all event time. With this particular choice of $\beta_p(\cdot)$, the expected value of the ijth scaled Schoenfeld residual can be expressed as $E(r^{\star}_{S_{ijp}}) = \nu_p(t_{ij} - \bar{t})$. If ν_p is non-zero, the coefficient of x_p is time-dependent and the PH assumption is violated. Letting $t_{11}, t_{12}, \cdots, t_{mn}$ be the time of all observed event times d in the data set, the GT test for x_p is given by

$$\frac{[\sum_{ij=11}^{mn}(t_{ij} - \bar{t})r^{\star}_{S_{ijp}}]^2}{d\text{var}(\hat{\beta}_p)\sum_{ij=11}^{mn}(t_{ij} - \bar{t})^2}. \tag{5.6}$$

The null hypothesis is set as $H_0 : \nu_p = 0$. This leads to an asymptotic χ^2 distribution with degree freedom 1 for x_p, and significantly large values mean the PH assumption can be rejected. For an overall test of the PH assumption in the model, each x_p in expression (5.6) can be aggregated. The overall test statistic follows

$$\frac{(t - \bar{t})' r_S \text{var}(\hat{\beta}) r_S'(t - \bar{t})}{\sum_{ij=11}^{mn}(t_{ij} - \bar{t})^2/d}, \tag{5.7}$$

where t is the column vector of all observed event times, and r_S is a matrix of Schoenfeld residuals for all covariates with rows representing different times in t and columns representing different covariates. The test statistic in the expression (5.7) has a χ^2 distribution with degree freedom P when the assumption of proportional hazards across all P explanatory variables is true. The function $g_p(t)$ in expressions (5.6) and (5.7) can be replaced by others, for example, by the logarithms of the event times, by the rank order of the event times, or by the survival function at each event time based on the Kaplan-Meier estimation.

5.2.3 Score Tests

The GT test can be regarded as an approximation of the general score test [77]. From version 3.0, the `survival` package began to use the score test for testing the PH assumption. The score test [77] is a general test for testing whether an encompassing model, often called a full model, can provide a significantly better fit than a reduced model, which is a special model of the full model, based on the score function and information matrix of the full model. The score function is the gradient of the log-likelihood function with respect to the parameters, and the information matrix is the Hessian matrix of the negative log-likelihood function, both evaluated at the maximum likelihood estimate of the model parameters under the null hypothesis. Under certain conditions, the score test statistic asymptotically follows a chi-squared distribution with the degrees of freedom equal to the difference in the number of parameters of the two nested models. For testing the PH assumption, the full model is the time-varying effect model as given in Equation (5.2) and the reduced model is the Cox PH model. In the actual implementation, the full model is treated as a Cox model by including $x_{ijp}g_p(t_{ijp})$ as an additional linear predictor term in the Cox PH model. If $g_p(t_{ijp}) = \nu_p(t_{ij} - \bar{t})$ is chosen, the score test is applied to test the null hypothesis expressed as $\nu_p = 0$.

Suppose we expand the Cox PH model by including $x_{ijp}g_p(t_{ijp})$ for $p = 1, \ldots, P$ as additional linear predictor terms and the form of the time-varying effect for covariate x_p is a linear function $\beta_p(t_{ij}) = \beta_p + \nu_p(t_{ij} - \bar{t})$, where ν_p is an unknown slope, $\bar{t}$ is the average of all event times. Since there are P covariates in the Cox PH model, the expanded model has P covariates. The new parameter vector is denoted by $\theta = (\beta, \nu) = (\beta_1, \beta_2, \cdots, \beta_P, \nu_1, \nu_2, \cdots, \nu_P)'$. The partial likelihood function of the expanded model is denoted by $L_1(\theta)$. The score test evaluates the score function and information matrix at the estimate of θ under the null hypothesis (ie, $\nu_1 = 0, \ldots, \nu_P = 0$), which is given by $\hat{\theta}_0 = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_P, \hat{\nu}_1 = 0, \hat{\nu}_2 = 0, \cdots, \hat{\nu}_P = 0)'$, where the $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_P)'$ is indeed the estimate of β

in the reduced Cox PH model (under the null hypothesis). The score function evaluated at $\hat{\theta}_0$ is given by $U_1(\hat{\theta}_0) = \frac{\partial \log L_1(\theta)}{\partial \theta} \mid_{\hat{\theta}_0} = (\frac{\partial \log L_1(\theta_0)}{\partial \beta_1}, \cdots, \frac{\partial \log L_1(\theta)}{\partial \beta_P}, \frac{\partial \log L_1(\theta)}{\partial \nu_1}, \cdots, \frac{\partial \log L_1(\theta)}{\partial \nu_P}) \mid_{\hat{\theta}_0}$. The $\hat{\theta}_0 = (\hat{\beta}, \nu = 0)$ can be regarded the maximizer of $\log(L_1(\theta))$ under the restriction $\nu_1 = 0, \ldots, \nu_P = 0$. Following the argument with the Lagrange multiplier optimization [86], we can see that the score vector $U_1(\hat{\theta}_0)$ has the following form:

$$\mathbf{U}_1(\hat{\theta}_0) = (0, 0, \cdots, 0, u_1, u_2, \cdots, u_P). \tag{5.8}$$

Similarly, we also need to evaluate the information matrix at $\hat{\theta}_0$, which is denoted by $\mathbf{I}_1(\hat{\theta}_0)$. The score test statistic for testing the PH assumptions for all the covariates is given by

$$\mathbf{U}_1'(\hat{\theta}_0)\mathbf{I}_1^{-1}(\hat{\theta}_0)\mathbf{U}_1(\hat{\theta}_0). \tag{5.9}$$

The above score statistic asymptotically follows a chi-square distribution with P degrees of freedom when the PH assumption holds.

The score test for a single covariate can also be defined similarly by expanding the Cox PH model with only a single $x_{ijp}g_p(t_{ijp})$ term. The score test statistic for a single covariate follows a chi-square distribution with 1 degree of freedom when the PH assumption holds.

5.3 Z-residuals

5.3.1 Definition of Z-residual and censored Z-residual

In this paper, we use `Z-residual` [19, 87] for detecting Non-PH in survival analysis. The key idea of Z-residual is to replace the survival probability of a censored failure time with a uniform random number between 0 and the survival probability of the censored time. For fitting the Cox PH or AFT models with the shared frailty, the normalized randomized survival probabilities (RSPs) for y_{ij} are defined as:

$$S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij}) = \begin{cases} S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is uncensored, i.e., } \delta_{ij} = 1, \\ U_{ij}\, S_{ij}(y_{ij}), & \text{if } y_{ij} \text{ is censored, i.e., } \delta_{ij} = 0, \end{cases} \tag{5.10}$$

where U_{ij} is a uniform random number on $(0, 1)$, and $S_{ij}(\cdot)$ is the postulated survival function for t_{ij} given x_{ij}. $S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij})$ is a random number between 0 and $S_{ij}(y_{ij})$ when y_{ij} is censored. It is proved that the RSPs are uniformly distributed on $(0, 1)$ given x_i under the true model [19]. Therefore, the RSPs can be transformed into residuals with any desired distribution. We prefer to transform them with the normal quantile:

$$r_{ij}^Z(y_{ij}, \delta_{ij}, U_{ij}) = -\Phi^{-1}(S_{ij}^R(y_{ij}, \delta_{ij}, U_{ij})), \tag{5.11}$$

which is normally distributed under the true model, so we can conduct model diagnostics with Z-residuals for censored data in the same way as conducting model diagnostics for a normal linear regression model. There are a few advantages of transforming RSPs into Z-residuals. First, the diagnostics methods for checking normal linear regression are rich in the literature. Second, transforming RSPs into normal deviates facilitates the identification of extremely small and large RSPs. The frequency of such small RSPs may be too small to be highlighted by the plots of RSPs. However, the presence of such extreme SPs, even very few, is indicative of model misspecification. Normal transformation can highlight such extreme RSPs.

For checking the PH assumption, the **censored Z-residual** [19, 87] can be applied. Transforming survival

probabilities using the quantile of standard normal distribution [50] defined as $r_{ij}^n(y_{ij}) = -\Phi^{-1}(S_{ij}(y_{ij}))$, where y_{ij} is the observed failure time or censoring time. The diagnosis of the GOF of survival probabilities can be converted to the diagnosis of the normality of the **censored Z-residuals**.

5.3.2 Checking PH assumption Based on Z-residuals

To detect Non-PH across all terms included in a survival model, we can assess the model's overall goodness-of-fit (GOF). A QQ plot of Z-residuals can be used to graphically evaluate the overall GOF. The Shapiro-Wilk (SW) normality test, applied to Z-residuals, can be used to numerically test the overall GOF of the model. Additionally, censored Z-residuals can be used to detect Non-PH in the survival model by assessing the overall GOF of the model. The function `gofTestCensored` in R package `EnvStats` [51, 52] provides a Shapiro-Franciards (SF) test for testing the normality of multiply censored data. Hence, `gofTestCensored` can be applied to check the normality of censored Z-residuals (denoted by CZ) for checking the overall GOF of survival models. The overall GOF test methods used for detecting Non-PH across all terms are denoted as "R-T" with "R" denoting residual name and "T" denoting test method. For example, Z-SW is the test method in which the normality of Z-residuals is tested using the SW test. The CZ-CSF method tests the normality of censored Z-residuals using an extended SF method for censored observations.

The conditional distribution of Z-residual given x_i is approximately a standard normal and is homogeneous at varying levels of covariates when a model is correctly specified. The plot of Z-residuals against linear predictors can be considered another graphical diagnostic method to assess the model's overall GOF for detecting Non-PH. When the PH assumption is tenable, we expect that there is no trend in these scatterplots. Moreover, the plot of Z-residuals against covariates can be used for checking the individuals' covariate effect of the PH assumption. Only such a graphical examination is difficult to determine whether the observed trend in Z-residuals is caused by chance or by the violation of the PH assumption. Therefore, we desire a formal test to quantify the statistical and practical significance of the difference between the observed trend and the expected horizontal line at 0. To this end, we propose the following diagnostic procedure. The Z-residuals can be divided into k groups by cutting the covariates or linear predictors into equally-spaced intervals. We can check whether the Z-residuals of the k groups are homogeneously distributed using the scatterplot to detect Non-PH. The homogeneity will be expected that the Z-residuals are randomly scattered without showing differential means or variances in groups. Numerical methods to assess the homogeneity of such grouped Z-residuals are to test the equality of group means and the equality of group variances of the Z-residuals. We apply the F-test in ANOVA to test the equality of means of grouped Z-residuals and Bartlett's test to test the equality of variances of grouped Z-residuals. These proposed quantitative methods for checking the PH assumption are also given a short nomenclature for ease of reference. Z-AOV-LP and Z-BL-LP are the methods of applying ANOVA and Bartlett to examine the equality of the means of Z-residuals against the groups formed with the linear predictor (LP). For a specific covariate, Z-AOV-x_1, Z-AOV-x_2, Z-BL-x_1 and Z-BL-x_2 are the methods of testing the equality of the means and variances of Z-residuals against the covariates x_1 and x_2, respectively.

5.3.3 A P-value Upper Bound for Assessing Replicated Z-residuals GOF Test p-values

A difficulty in conducting statistical tests with Z-residuals is the randomness in the test p-values. Given a fitted model, we can generate many sets of Z-residuals and obtain replicated test p-values. According to the

distribution of order statistics of correlated random variables [61, 62], we can obtain the following inequality for the rth order statistics $p_{(r)}$:

$$P(p_{(r)} < t) \leq \min\left(1, t\frac{J}{r}\right). \tag{5.12}$$

Based on (5.12), a p-value upper bound for observed (simulated) rth statistics $p_{(r)}^{\text{obs}}$ is given by $\min\left(1, p_{(r)}^{\text{obs}}\frac{J}{r}\right)$. To avoid the selection of r, we report the minimal upper bound for $r = 1, \ldots, J$, denoted by $p_{\min}$:

$$p_{\min} = \min_{r=1,\ldots,J} \min\left(1, p_{(r)}^{\text{obs}}\frac{J}{r}\right). \tag{5.13}$$

The $p_{\min}$ is rather conservative for assessing model fit because of its generality. When a model has a small $p_{\min}$, it is highly suspected that the model can be improved to better fit the dataset. Considering the conservatism of $p_{\min}$, a rule of thumb for declaring model failure in practice should be much larger, say 0.25 as suggested by Yuan and Johns [63], than the conventional 0.05 for exact p-values.

5.4 Simulation Studies

5.4.1 Detection of Non-PH Due to Time-varying Covariate Effects

In this section, we present simulation studies to demonstrate the effectiveness of Z-residual in detecting Non-PH due to time-varying covariate effects. The Cox PH model can be extended by incorporating random effects (frailties) [11], where the frailties are common or shared among individuals within a cluster or group [1–4]. We generate the failure times from a Cox PH model with shared frailty with the following hazard:

$$h_{ij}(t) = z_i \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_2(t) x_{ij}^{(2)}\Big) h_0(t). \tag{5.14}$$

Two covariates are generated including $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ from a normal with mean 0 and standard deviation 2, where $\beta_1 = 0.3$, $\beta_2(t) = \beta_a$ for $t < \pi/2$ and $\beta_2(t) = \beta_b$ for $t \geq \pi/2$. We generate the true failure times from a Weibull regression model with shape parameter (α=3) and scale parameter (λ=0.007); then the data generator is given by

$$t_{ij} = \begin{cases} \left\{ \frac{-\log(v_{ij})}{\lambda z_i \exp\left(\beta_1 x_{ij}^{(1)} + \beta_a x_{ij}^{(2)}\right)} \right\}^{(1/\alpha)}, \\ \text{if } -\log(v_{ij}) < z_i \lambda \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_a x_{ij}^{(2)}\Big)(\pi/2)^\alpha, \\ \left\{ \frac{-\log(v_{ij}) - z_i \lambda \exp\left(\beta_1 x_{ij}^{(1)} + \beta_a x_{ij}^{(2)}\right)((\pi/2)^\alpha) + z_i \lambda \exp\left(\beta_1 x_{ij}^{(1)} + \beta_b x_{ij}^{(2)}\right)((\pi/2)^\alpha)}{z_i \lambda \exp\left(\beta_1 x_{ij}^{(1)} + \beta_b x_{ij}^{(2)}\right)} \right\}^{(1/\alpha)}, \\ \text{if } -\log(v_{ij}) \geq z_i \lambda \exp\Big(\beta_1 x_{ij}^{(1)} + \beta_a x_{ij}^{(2)}\Big)(\pi/2)^\alpha, \end{cases} \tag{5.15}$$

where $i = \{1, \cdots, \text{g}\}$ and $j = \{1, \cdots, n_i\}$ and a random number v_{ij} is simulated from Uniform[0, 1]; the frailty term z_i is generated from a gamma distribution with a variance of 0.5. The censoring times c_{ij} is simulated from an exponential distribution, exp(7.5), so that the censoring rate is about $c \approx 50\%$. Then, the observed survival time is equal to $min(t_{ij}, c_{ij})$. We have two simulation settings. The first simulation setup is that the data with time-varying effects ($\beta_a \neq \beta_b$). We generate datasets with varying degrees of Non-PH by setting the covariate effects $(\beta_a, \beta_b) \in \{(0.35, -0.35), (0.65, -0.65), (1, -1), (1.35, -1.35)\}$, to investigate if the performance of Z-residual depends on the degree of Non-PH. We name this dataset as time-varying data.

For the second simulation setting, we set $\beta_a = \beta_b$ to obtain the datasets with time-constant covariate effect, by setting the covariate effects $(\beta_a, \beta_b) \in \{(0.35, 0.35), (0.65, 0.65), (1, 1), (1.35, 1.35)\}$. The dataset is named time-constant data. We consider fitting the Cox PH model with shared frailty with the time-varying data and time-constant data to investigate the power of Z-residuals compared to conventional diagnostic tools for detecting Non-PH. For each pair of (β_a, β_b) in two simulation settings, the data size is fixed as 10 clusters of 50 observations with the censoring rates $c \approx 50\%$, we then generated 1000 datasets for estimating model rejection rates using different residual diagnostic methods.

We first present the results of graphical methods for checking the PH assumption across all covariates in two simulated datasets, one with time-constant effect $\beta_a = \beta_b = 1.35$, and the other with time-varying effect $\beta_a = 1.35$ and $\beta_b =$ -1.35. As shown in the panels of the first row of Figure 5.1, the QQ plot of the Z-residuals for the time-constant data aligns well along the 45° straight line, but the QQ plot for the time-varying data indicates the deviation of Z-residuals from the normal. The second and third rows of Figure 5.1 display the scatterplots of Z-residuals against the linear predictor. Under the time-constant data, the Z-residuals are mostly bounded between -3 and 3 as the standard normal deviates without any discernible pattern, with the LOWESS curve roughly horizontal around 0. Under the time-varying data, a slight non-linear pattern in the Z-residual scatterplot is observed, and the smoothed LOWESS curve in the residual plot deviates slightly from a horizontal straight line. In the third row, Z-residuals was divided into $k = 10$ groups by cutting the linear predictors into equally spaced intervals. The boxplots of the grouped Z-residuals indicate that the Z-residuals are homogeneous across groups under the time-constant data, but exhibit varying group means and variances under the time-varying data.

We further checked the scatterplots and boxplots of Z-residuals against x_1 and x_2 under the time-constant and time-varying datasets, as shown in Figures 5.2 and 5.3. Under the time-constant data, the scatterplots of Z-residuals against x_1 and x_2 are randomly scattered without exhibiting any pattern with the LOWESS curves roughly a horizontal line at 0. Under the time-varying data, the scatterplots of Z-residuals against x_1 are mostly bounded between -3 and 3 as standard normal deviates with a roughly straight LOWESS line. However, the Z-residuals against x_2 clearly exhibit a pattern with a downward trend in the LOWESS curve as x_2 increases. We further divided Z-residuals into $k = 10$ groups by cutting x_1 and x_2 into equally spaced intervals, respectively. Under the time-constant data, the boxplots of Z-residuals against x_1 and x_2 across all the groups are fairly homogeneous. For the time-varying data, the Z-residuals against x_1 for each group are homogeneous across groups, but the group means of Z-residuals decrease as the x_2 increases. These plots suggest that the PH assumption is violated for the covariate x_2 in the time-varying data.

As a comparison, we also show the performance of scaled Schoenfeld residuals for checking the PH assumption in the time-constant and time-varying datasets. If the PH assumption holds, we expect a horizontal line in a plot of the scaled Schoenfeld residuals against the observed survival time. Figure 5.3 shows the scaled Schoenfeld residuals for covariates x_1 and x_2 against observed survival time under the time-constant and time-varying datasets, respectively. Under the time-constant data, the scaled Schoenfeld residuals against the covariates x_1 and x_2 show that the smoothed LOWESS curves deviate little from horizontal lines, suggesting the validity of the PH assumption. Under the time-varying data, the plot of the scaled Schoenfeld residuals for the covariates x_1 shows randomly scattered with a roughly straight smoothed LOWESS line; however, the scaled Schoenfeld residuals for the covariates x_2 clearly show a pattern with a curved smoothed LOWESS line. These plots suggest a non-constant hazard ratio for the covariate x_2 under the time-varying data.

In addition to the graphical checking, we also examine the performances of numerical tests introduced in

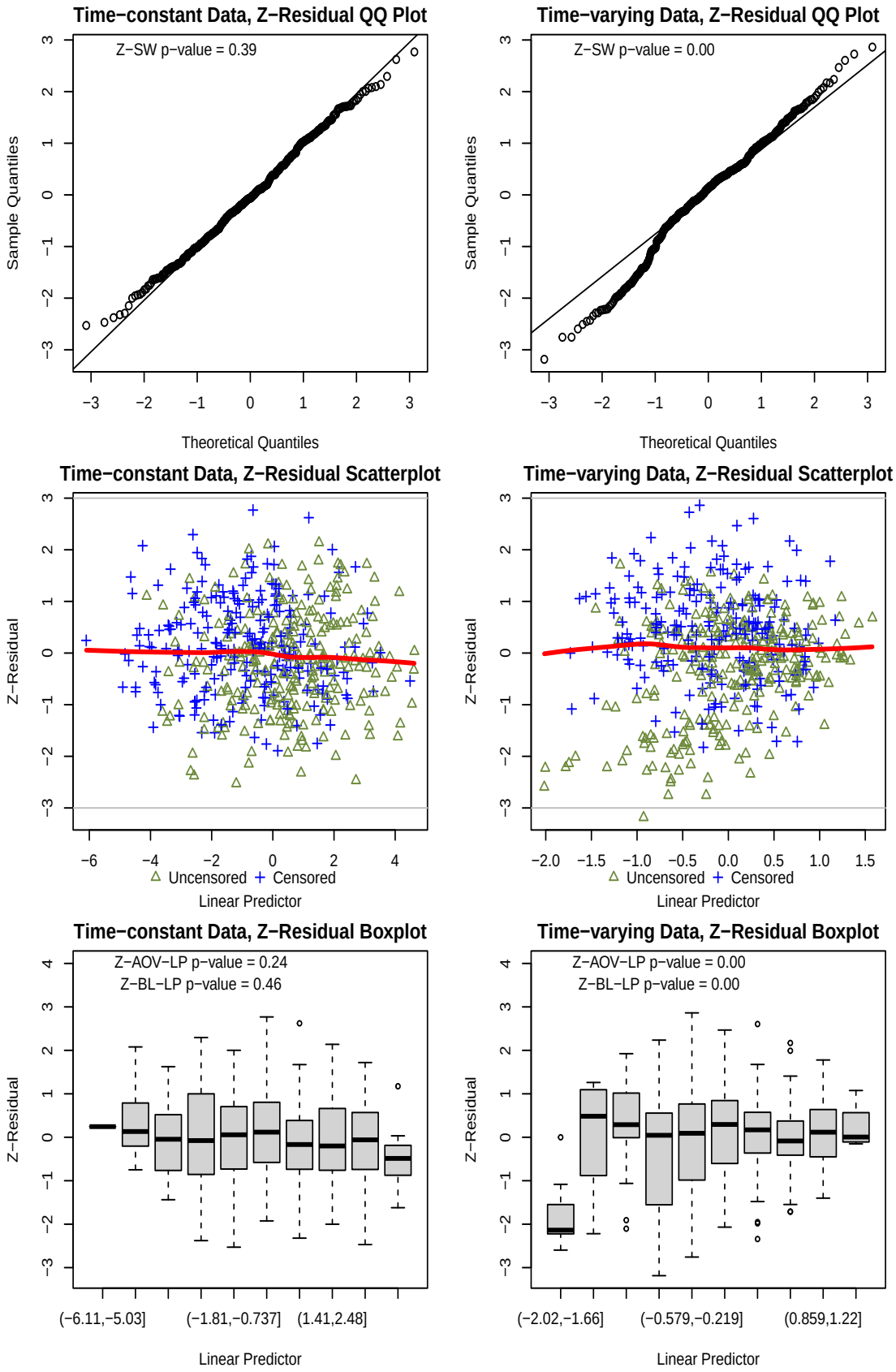


Figure 5.1: Performance of the Z-residuals as graphical tools for checking the violation of the global PH assumption due to time-varying effect. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$

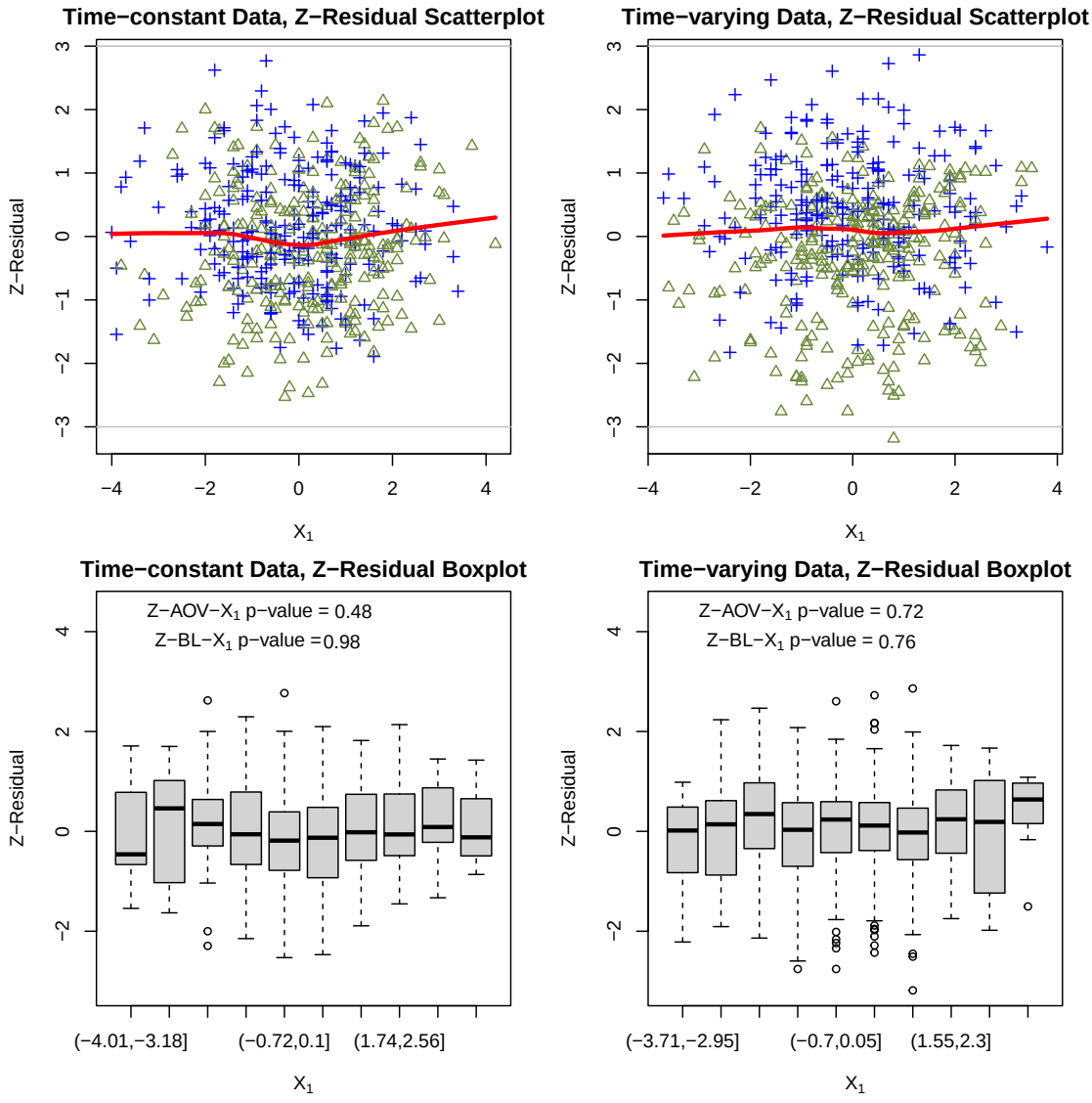


Figure 5.2: Performance of the Z-residuals as graphical tools for checking the violation of the PH assumption for covariate x_1 due to time-varying effect. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$.

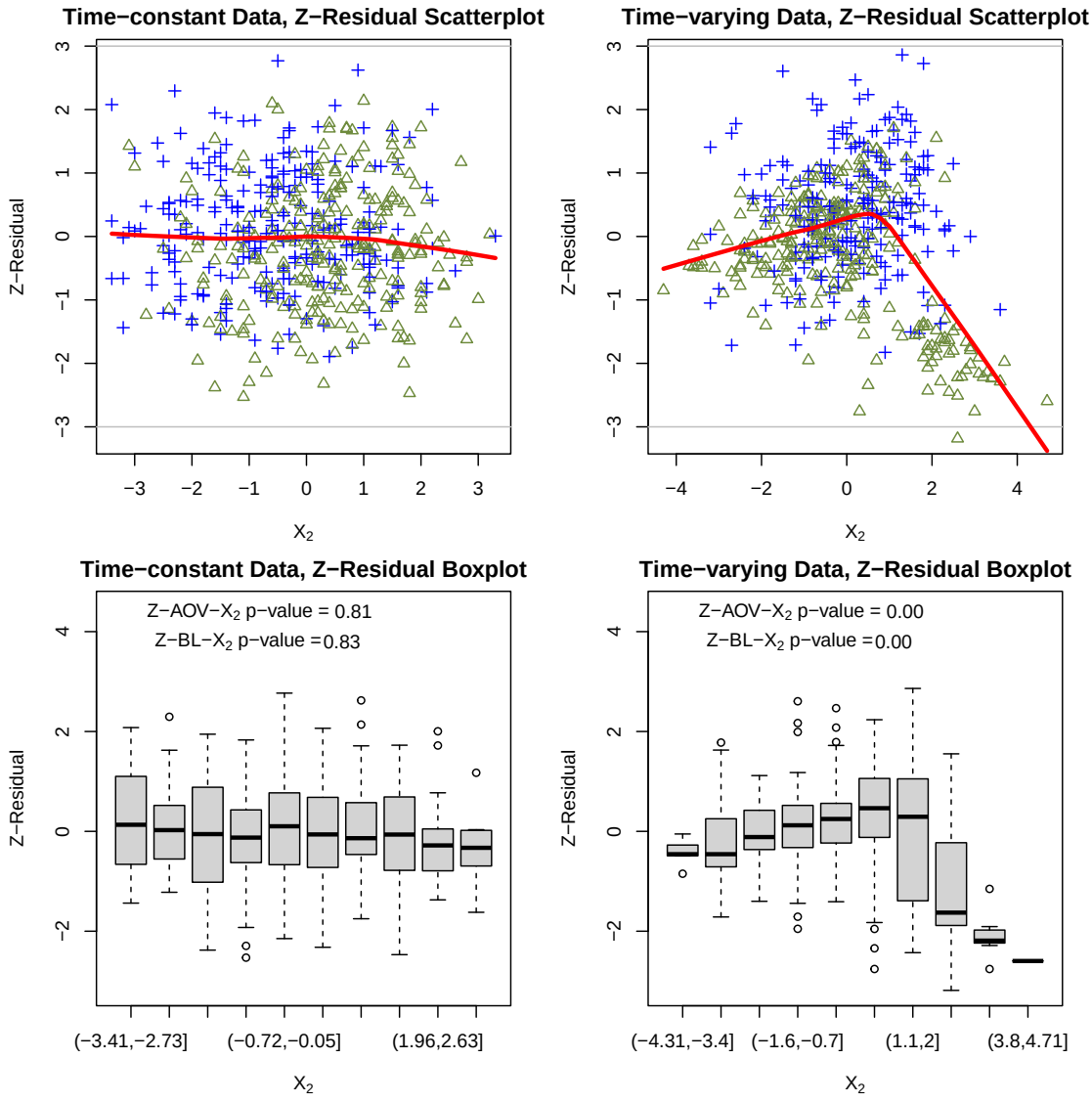


Figure 5.3: Performance of the Z-residuals as graphical tools for checking the violation of the PH assumption for covariate x_2 due to time-varying effect. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$.

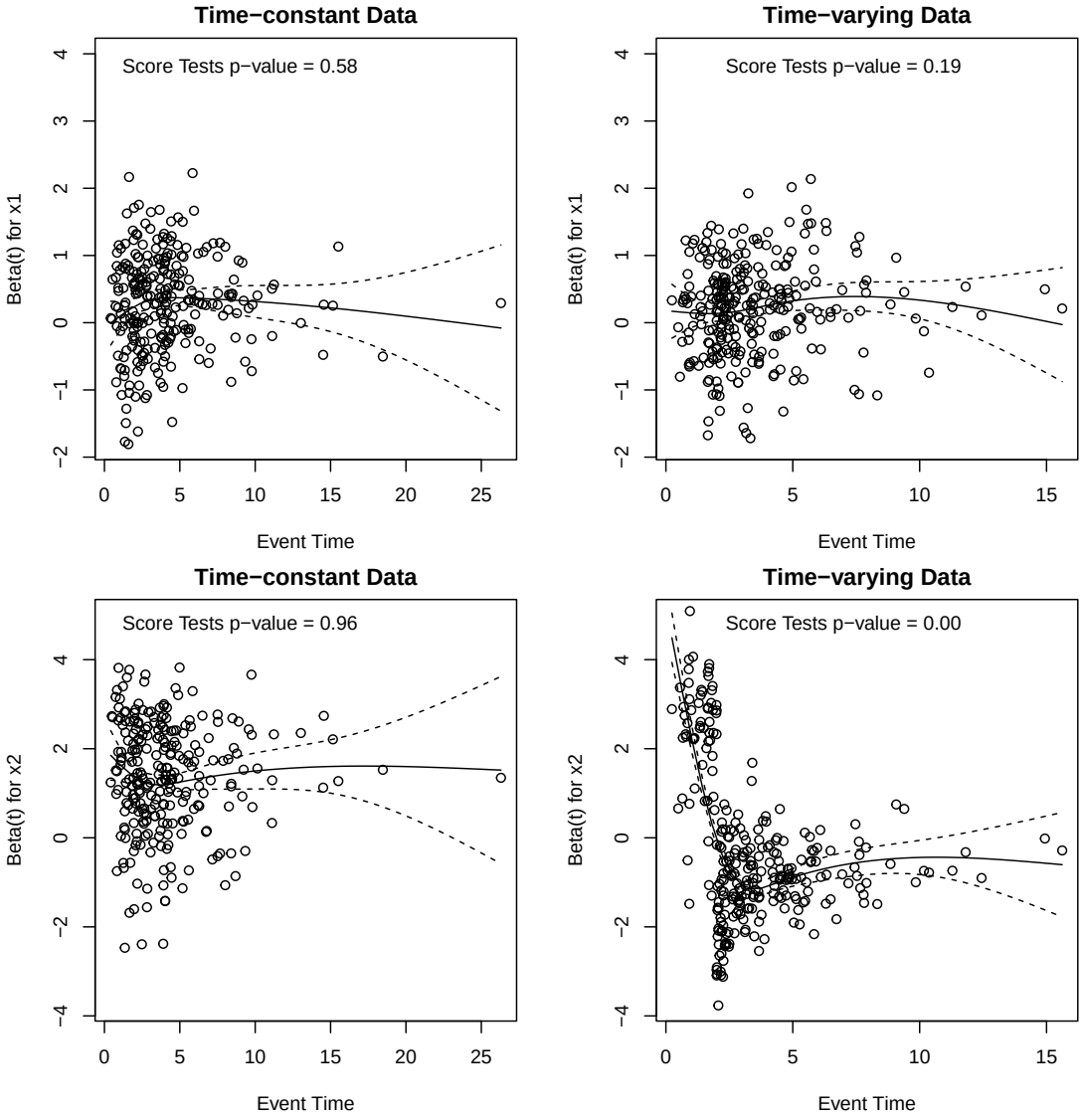


Figure 5.4: Scaled Schoenfeld residuals for the covariates x_1 and x_2 with 95% confidence interval. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$.

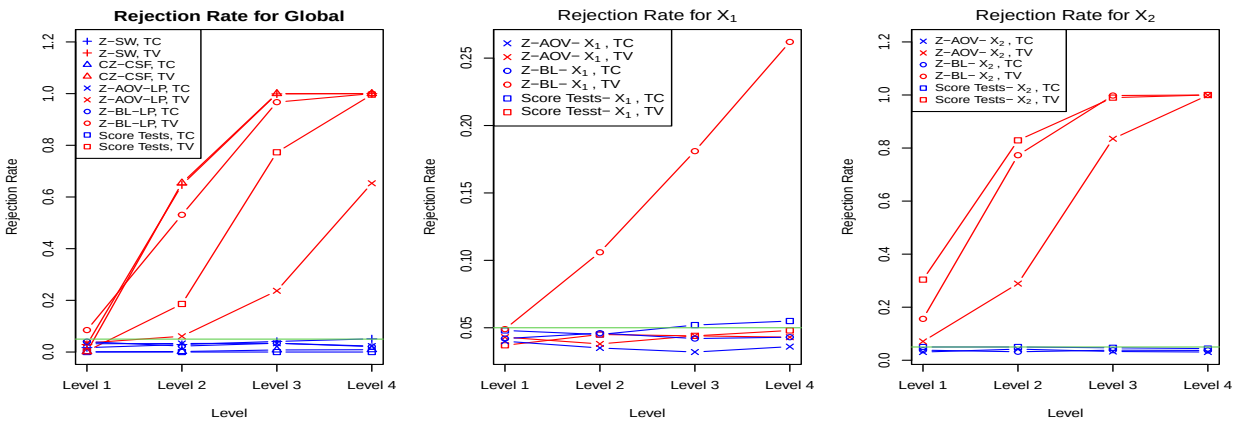


Figure 5.5: Rejection rates of various statistical tests based on Z-residual and score test. A model is rejected when the test p-value is smaller than 5%. Note that we use a random Z-residual test p-value rather than the $p_{\min}$.

Section 3.2 for detecting Non-PH based on Z-residuals. For demonstrating the performances of the numerical tests using Z-residuals and score tests, we replicated the time-constant and time-varying datasets 1000 times. Under each scenario, the rejection rate of each test method was estimated as the proportion of the test p-values less than 0.05 for the tests based on the residuals from fitting a Cox PH regression model to the simulated datasets. The rejection rates of the overall GOF tests for the global test of the PH assumption methods are shown in the first plot of Figure 5.5. The global test of the PH assumption methods, including Z-SW, CZ-CSF, Z-AOV-LP, Z-BL-LP and score tests, have false-positive rates (reject PH models when PH models are true models) close to the nominal level of 5% for all scenarios and have increased power (reject PH models when Non-PH models are true models) when the degree of Non-PH is moderate to strong. We also note that the power of detecting Non-PH increases as the degree of Non-PH increases. Moreover, the powers of the overall Z-SW, CZ-CSF and Z-BL-LP tests are significantly greater than the power of the global score test and Z-AOV-LP test. The second plot of Figure 5.5 shows the rejection rate for the covariate x_1 under the time-constant and time-varying datasets. The Z-AOV-x_1, Z-BL-x_1 and score tests methods for the covariate x_1 under the time-constant and time-varying datasets have low false-positive rates and powers; the rejection rate of Z-BL-x_1 method is slightly higher than the rejection rates of others under the time-varying data. The third plot of Figure 5.5 shows the performances of the Z-AOV-x_2, Z-BL-x_2 and score test methods for the covariate x_2 under the time-constant and time-varying datasets. Under the time-constant data, the false positive rates for all methods are close to the nominal level of 5%. Under the time-varying data, the power of score tests, Z-BL-x_2 and Z-AOV-x_2 increase as the degree of Non-PH increases; but the powers of the score test and Z-BL-x_2 is higher than the power of Z-AOV-x_2 method when the degree of Non-PH is severe. Therefore, the PH assumption for the covariates x_2 has been violated under the time-varying data.

5.4.2 Detection of Non-PH Due to Accelerated Failure

In this section, we investigate the performance of the Z-residuals for detecting Non-PH due to accelerated failure time. The covariate x_{ij} is generated from a normal distribution with mean 0 and standard deviation 1. The frailty term $z_i = \exp(u_i)$ is generated from a gamma distribution with a variance of 0.5. The accelerated failure model (AFT) with shared frailty can be fitted if the baseline hazard function is fully specified [11]. We generate the true failure times from the Lognormal AFT regression model with shared frailty with parameters

mean μ=0 and standard deviation σ, then the data generator is given by:

$$t_{ij} = \frac{1}{\exp(x_{ij} + u_i)} \exp\big(\sigma\phi^{-1}(v_{ij}) + \mu\big), \tag{5.16}$$

where $i = \{1,\cdots, \text{g}\}$ and $j = \{1,\cdots, n_i\}$ and a random number v_{ij} was simulated from Uniform[0, 1]. The censoring times c_{ij} were generated from exp(3.8) to obtain the censoring rate $c \approx 50\%$. The Lognormal AFT model with shared frailty of parameter σ was set to three different values 1, 1.5, and 2, to investigate if the performance of Z-residual depends on the different parameter values. These datasets are named Lognormal AFT data. We then generate another failure time from the Weibull Cox PH regression model with shared frailty with a shape parameter of 0.48 and scale parameter of 1.67, then the data generator is given by:

$$t_{ij} = \Big(\frac{-\log(1 - v_{ij})}{\lambda \exp(-(x_{ij} + u_i))}\Big)^{1/\alpha} \tag{5.17}$$

where $i = \{1,\cdots, \text{g}\}$ and $j = \{1,\cdots, n_i\}$, and a random number v_{ij} is simulated from Uniform[0, 1]. The censoring times c_{ij} were generated from exp(0.2), which can obtain the censoring rates $c \approx 50\%$. This dataset has been named Weibull PH data to differentiate it from the previous dataset. We fixed the number of clusters g=10 and cluster size n_i to be 50 with the censoring rate $c \approx 50\%$. For each parameter setting, we generated 1000 datasets for estimating the rejection rates of different diagnostic methods. We consider fitting these datasets with a Cox PH model with shared frailty.

The performance of graphical methods based on Z-residuals for detecting Non-PH was first evaluated on two single simulated datasets from Lognormal(0, 1) and Weibull(0.48, 1.67), respectively. The first row of Figure 5.6 displays the QQ plot of Z-residuals under the Lognormal AFT and Weibull PH datasets. The QQ plot of the Z-residuals for the Weibull data aligns well along the 45° straight line. By contrast, there are observable deviations from the diagonal line in the lower tail of the QQ plot of Z-residuals for the Lognormal dataset, indicating a clear violation of the global PH assumption. In the second and third rows of Figure 5.6, scatterplots of Z-residuals and boxplots of Z-residuals by cutting the linear predictor into 10 equally spaced intervals against the linear predictor are displayed. The residuals are mostly bounded between -3 and 3 and without any discernible pattern for both the Weibull PH and Lognormal AFT datasets. The Z-residuals are homogeneous across groups under both datasets, as indicated in the boxplots. Similar evaluations were carried out for the scatterplots and boxplots of Z-residuals against the covariate x as displayed in Figure 5.7. The results are mostly bounded between -3 and 3 with a roughly straight LOWESS line in the scatterplots, and the boxplots show that the Z-residuals across all groups are homogeneous under both datasets. These results suggest that the global Non-PH assumption is violated in the Lognormal AFT dataset.

We demonstrate the use of scaled Schoenfeld residuals for checking the PH assumption in the Lognormal AFT and Weibull PH datasets. The plots of the scaled Schoenfeld residuals against the observed survival time for covariates x under the Lognormal AFT and Weibull PH datasets are shown in Figure 5.8. The points appear randomly dispersed without exhibiting any pattern and the smoothed LOWESS curves only slightly deviate from horizontal lines. Therefore, the plots of scaled Schoenfeld residuals can not clearly detect Non-PH under both datasets.

The estimation of the rejection rate was performed by generating 1000 datasets ($n = 500$ and censoring percentage $\approx 50\%$) from different Lognormal AFT models, including {Lognormal^1(0, 2), Lognormal^2(0, 1.5), Lognormal^3(0, 1)}, and Weibull PH model with the parameter (0.48, 1.67). The results are presented in Table 5.1. The performance of several tests for checking the overall PH assumption, including Z-residuals

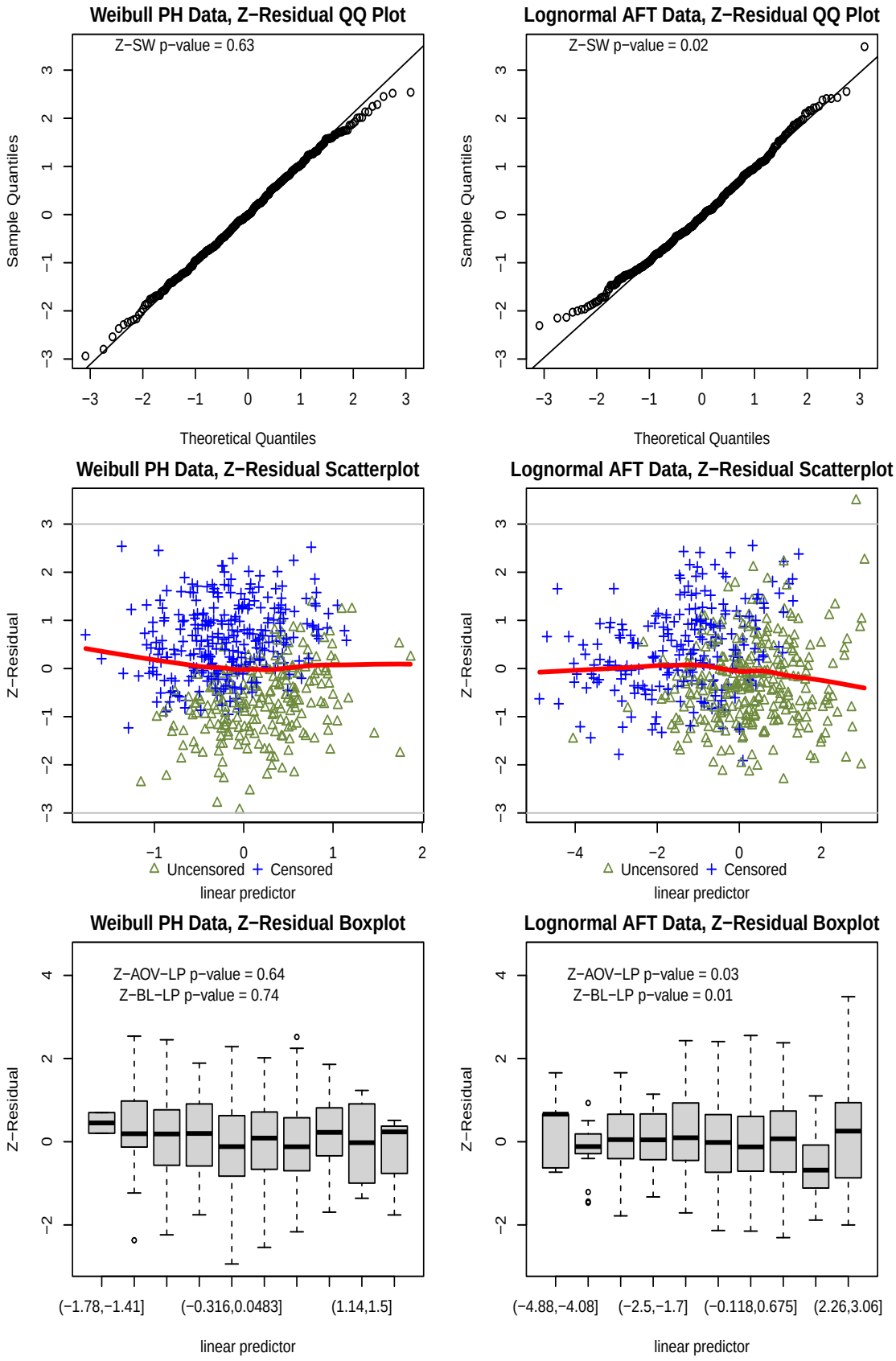


Figure 5.6: Performance of the Z-residuals as graphical tools for detecting global Non-PH due to accelerated failure time. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$.

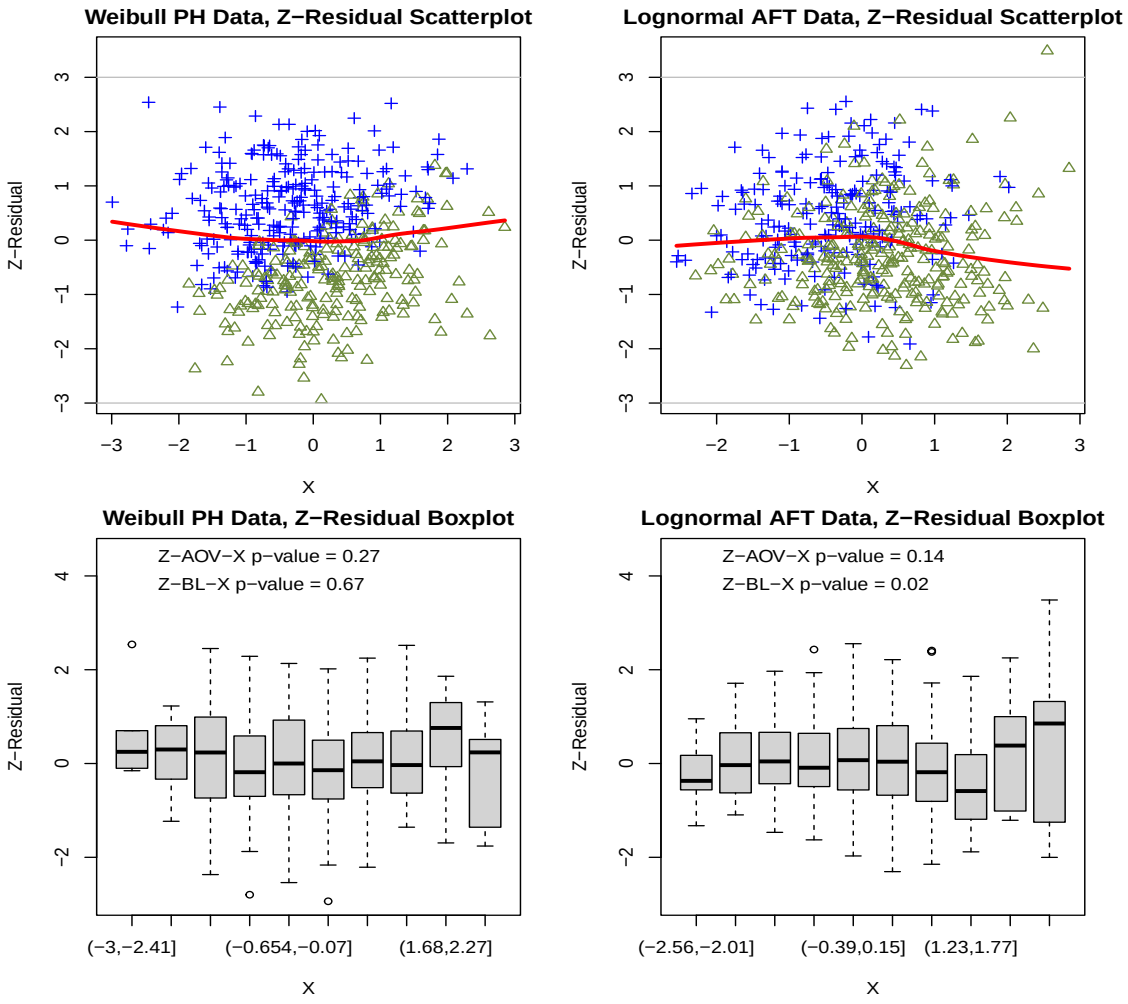


Figure 5.7: Performance of the Z-residuals as graphical tools for detecting Non-PH for covariate x due to accelerated failure time. The dataset was generated with 10 clusters of 50 observations in each cluster and a censoring rate $c \approx 50\%$.

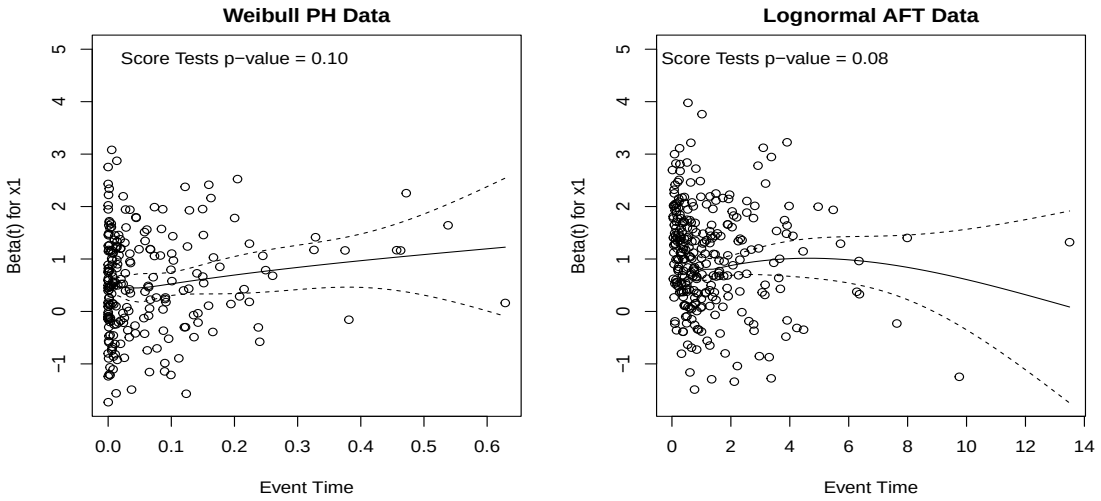


Figure 5.8: Scaled Schoenfeld residuals with 95% confidence interval for detecting Non-PH for the covariates X.

and score test methods, was first evaluated. It was observed that the rejection rates for the Weibull PH data were lower than the nominal level of 5% for all methods. For the Lognormal AFT datasets, the powers of the Z-SW and CZ-CSF tests increased as the parameter decreased. The rejection rates with the Z-AOV-LP, Z-BL-LP and score-G tests were very low under all Lognormal AFT datasets. Of all the methods considered for checking overall PH assumption, Z-SW and CZ-CSF performed the best, achieving a high rejection rate for the AFT Lognormal datasets and a low rejection rate for the Weibull PH dataset. Additionally, the individual covariate effects on the PH assumption were investigated. For the Weibull PH data, the rejection rates with the Z-AOV-X, Z-BL-X, and score-X tests remained at the nominal level of 5%. For all Lognormal AFT datasets, the rejection rate with the score-X test method was higher than those with the Z-AOV-X, and Z-BL-X tests, but only around 40%. Since the distribution shape of the model has changed due to accelerated failure times, the PH assumption is not tenable for the overall model. The overall GOF tests, namely Z-SW and CZ-CSF, demonstrate good power in detecting Non-PH under the Lognormal AFT datasets. Although all models only contain one covariate x, the frailty terms could affect the results of the score test method. If the score test method included the frailty term, the results of the overall PH check were lower than those of the score test method for checking the individual covariate effect. However, this phenomenon did not occur in the Z-residual method.

Table 5.1: Comparison of the percentages of model rejections based on Z-SW, score test for global (Score-G), Z-AOV-X, Z-BL-X and score test for X (Score-X). The response variables are simulated from varying models: AFT Weibull(0.48, 1.67), AFT Lognormal[1](0, 2), AFT Lognormal[2](0, 1.5), AFT Lognormal[3](0, 1), respectively.

	Checking Overall GOF					Checking Individual Covariate		
Model	Z-SW	CZ-CSF	Z-AOV-LP	Z-BL-LP	Score-G	Z-AOV-X	Z-BL-X	Score-X
Weibull	0.8	0.3	3.4	4.3	0.1	3.8	4.5	5.0
Lognormal[1]	15.8	9.2	8.8	10.0	1.4	6.6	7.7	30.3
Lognormal[2]	44.5	46.1	10.7	14.0	1.1	7.0	12.2	38.0
Lognormal[3]	84.2	90.6	9.0	13.5	1.8	5.6	11.9	45.3

5.5 Real Data Example

This section presents a real data application of detecting Non-PH using the proposed Z-residual diagnostics tools. The dataset used for this illustration is from the diabetic retinopathy study [88] that followed 1742 patients with diabetic retinopathy in both eyes. One randomly selected eye in each patient was treated and the other was untreated. The study was initiated in 1971 and followed over several years for the occurrence of blindness in their two eyes. The study collected data on the time to blindness for diabetic retinopathy and prognostic factors, including the treatment for which one randomly selected eye in each patient, how the age at onset of diabetes affects the outcome, the type of laser used for treatment, and the type of diabetes. More descriptions of these variables can be found in [88]. In our analysis, we selected 197 patients with 394 observations, which were previously analyzed by various authors [89–91]. Each patient is considered a cluster due to the frailty shared between the two eyes. The censoring rate is 60.66% and is caused by death, dropout, or end of the study. The outcome of interest is the time to blindness in months. The Cox PH model with frailty and the AFT Lognormal model was fitted to this dataset.

This study was deemed exempt from ethics approval since the data utilized were publicly available in Modelling Paired Survival Data with Covariates [88], and no personally identifiable information was collected or used. This research adhered to the principles outlined in TCPS 2-2nd Edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans [64].

The graphical of scaled Schoenfeld residuals and score test is demonstrated to provide adequate assessments of the PH assumption in the Cox PH model with frailty when applied to the Diabetic Retinopathy study dataset. It is important to note that these methods can only be applied to a Cox PH model. Figure 5.9 displays the scaled Schoenfeld residuals for each covariate against the event time, to detect a violation of the PH assumption. The smoothed curves in the plots show little deviation from horizontal lines, indicating that the PH assumption is not violated for each covariate. The score test gives p-values of 0.18, 0.24, 0.34 and 0.18, respectively, for each particular covariate. Moreover, the global score test produced a p-value of 1.00, as shown in Table 5.2. These results suggest the validity of the PH assumption in the Cox PH model with frailty.

Using Z-residuals, we first present the results of graphical checking and the overall PH testing for the Cox PH model with frailty and the Lognormal AFT model. The first row of Figure 5.10 displays the QQ plot of the Z-residual for these two models. The QQ plot of Z-residual for the Cox PH model with frailty deviates from the 45° straight line in the upper and lower tails, while the plot for the AFT Lognormal model aligns well with the 45-degree line. The Z-SW tests give p-values of 0.00 and 0.78 for the Cox PH model with frailty and the AFT log-normal model, respectively, supporting the results of the graphical checking. The second and third rows of Figure 5.10 show the scatterplots and boxplots of the Z-residuals against the linear predictor. Under the AFT Lognormal model, the LOWESS curve appears to be a horizontal line at 0, and the grouped Z-residuals appear to have equal means and variances across groups. In contrast, the Cox PH model with frailty displays a pattern with a curved LOWESS curve and grouped Z-residuals with different means across groups. To assess the statistical significance of the observed trends, we apply Z-AVO-LP and Z-BL-LP to test the equality of the means and variances of the grouped Z-residuals. The Z-AVO-LP method gives p-values of 0.00 and 1.00 for the Cox PH model with frailty and the AFT Lognormal model, respectively. On the other hand, the Z-BL-LP method provides p-values of 0.00 and 0.75 for the two models, respectively. The very small p-value of the Z-SW and Z-AVO-LP tests for the Cox PH model with frailty strongly suggests that the PH assumption does not hold in this real application.

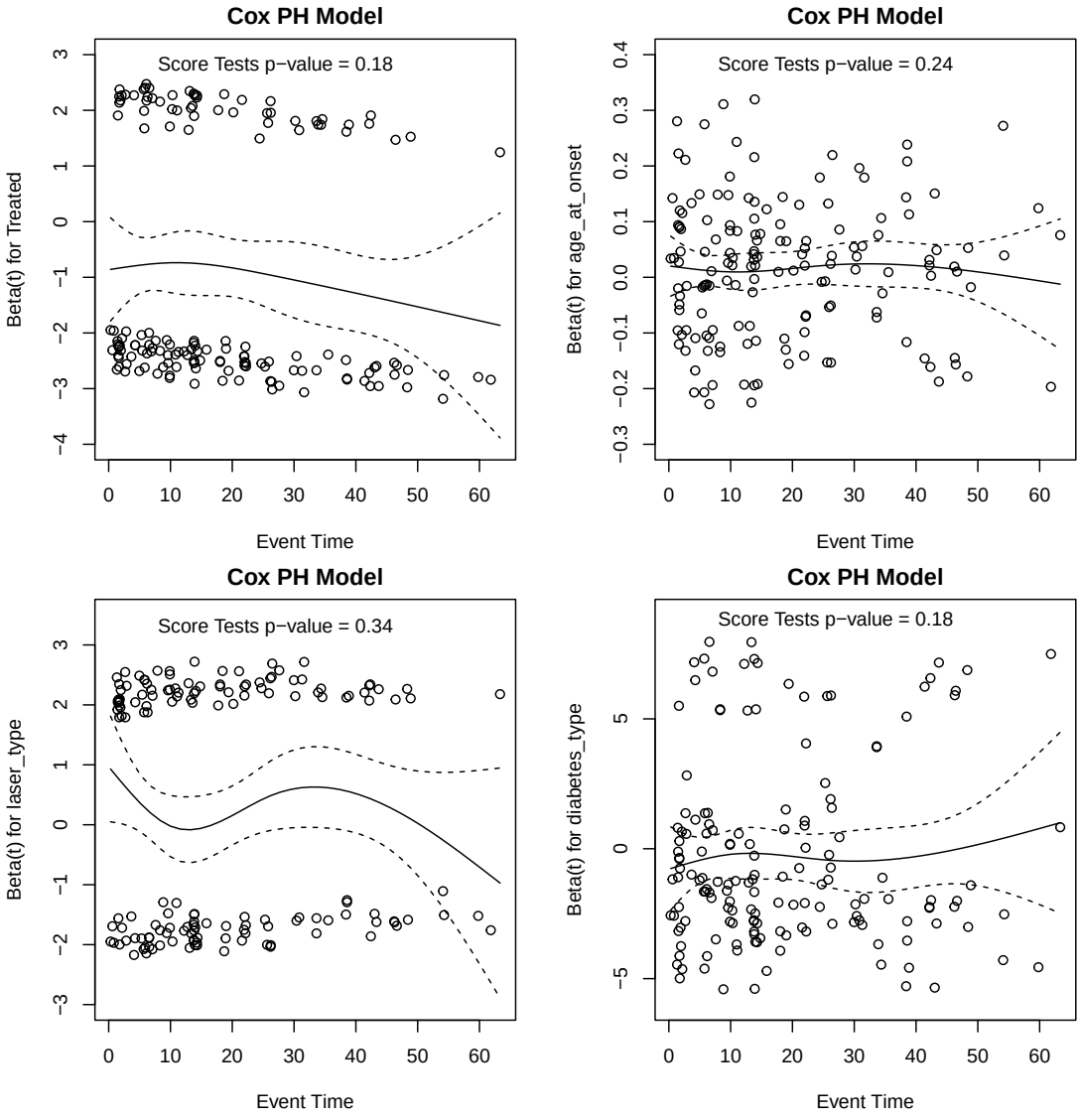


Figure 5.9: Scaled Schoenfeld residuals for all covariates in the Cox PH model with frailty fitted to the diabetic retinopathy study dataset. The dashed lines represent the 95% confidence interval.

Table 5.2: AIC, $p-$values for the CZ-CSF test, $p_{\min}$ values for Z-SW, Z-AOV-LP and Z-BL-LP test for the Cox PH and AFT Lognormal models, respectively, for the diabetic retinopathy study data.

Model	AIC	Score tests	CZ-CSF $p-$value	Z-SW $p_{\min}$	Z-AOV-LP $p_{\min}$	Z-BL-LP $p_{\min}$
Cox PH	1703.391	1.00	0.0005	0.011	$<$**0.00001**	0.021
AFT LN	1674.423	-	0.336	0.817	0.999	0.999

Table 5.3: $p_{\min}$ values for Z-AOV-Treat (Z-AOV-T), Z-BL-Treat (Z-BL-T), Z-AOV-Age (Z-AOV-A), Z-BL-Age (Z-BL-A), Z-AOV-Laser (Z-AOV-L), Z-BL-Laser (Z-BL-L), Z-AOV-Diabete (Z-AOV-D) and Z-BL-Diabete (Z-BL-D) test for the Cox PH and AFT Lognormal models, respectively, for the diabetic retinopathy study data.

Model	Z-AOV-T	Z-BL-T	Z-AOV-A	Z-BL-A	Z-AOV-L	Z-BL-L	Z-AOV-D	Z-BL-D
Cox PH	0.995	0.066	0.998	0.995	0.998	0.995	0.990	0.259
AFT LN	1.000	0.935	1.000	0.993	1.000	0.993	1.000	0.426

We also verify the PH assumption for each covariate using Z-residuals. The scatterplots and boxplots of Z-residuals against each covariate under these two models are shown in Figure 5.12 (in the supplementary materials), and the results of statistical tests are displayed in the corresponding boxplots. The graphical checking and statistical tests consistently demonstrate the validity of the PH assumption for each covariate.

The Z-residuals p-values are subject to randomness. Hence, to verify the robustness of the results, the Z-residuals were randomly generated 1000 times to obtain 1000 replicated p-values for each test method. The histograms of the 1000 replicated p-values for the overall PH check based on Z-residuals are shown in Figure 5.11. The red vertical lines in these histograms indicate the upper bound of these replicated p-values, $p_{\min}$, which are presented in Table 5.2. As seen in the second column of the histograms, the p-values of Z-SW, Z-AVO-LP and Z-BL-LP tests for the AFT Lognormal model are mostly greater than 0.05, resulting in larger $p_{\min}$ values. Table 5.2 shows that the $p_{\min}$ values for each test method under the AFT Lognormal model are close to 1. In contrast, the first column of the histograms displays that a large proportion of the replicated Z-SW, Z-AVO-LP, and Z-BL-LP p-values under the Cox PH model with frailty are smaller than 0.05. Table 5.2 shows that the $p_{\min}$ values of the Z-SW, Z-AVO-LP, and Z-BL-LP tests for the Cox PH model with frailty are significantly smaller than 0.05, further confirming that the PH assumption is invalid. Figures 5.13 and 5.14 (in the supplementary materials) display the histograms of 1000 replicated Z-AOV-Treat, Z-BL-Treat, Z-AOV-Age, Z-BL-Age, Z-AOV-Laser, Z-BL-Laser, Z-AOV-Diabete, Z-BL-Diabete tests p-values, which show that a large proportion of p-values under these two models are greater than 0.05. Table 5.3 shows that all the $p_{\min}$ values for each covariate test method for diagnosing the two models(shown with red lines in Figures 5.13 and 5.14) are larger than 0.05.

Furthermore, Table 5.2 also reports the non-random CZ-CSF test p-values for the two models and the AIC values for comparing these two models. The p-value of the CZ-CSF test for the Cox PH model with frailty is smaller than 0.05, while the p-value of the CZ-CSF test for the AFT Lognormal model is 0.336. The AIC value of the AFT Lognormal model is 1674.423, and the AIC value of the Cox PH model with frailty is 1703.391, indicating that the AFT Lognormal model provides a better fit to the data. This conclusion is consistent with the results from the model diagnostics based on Z-residuals. Hence, all results provide strong evidence that the Cox PH model with frailty does not fit the dataset well, but the AFT Lognormal model performs better.

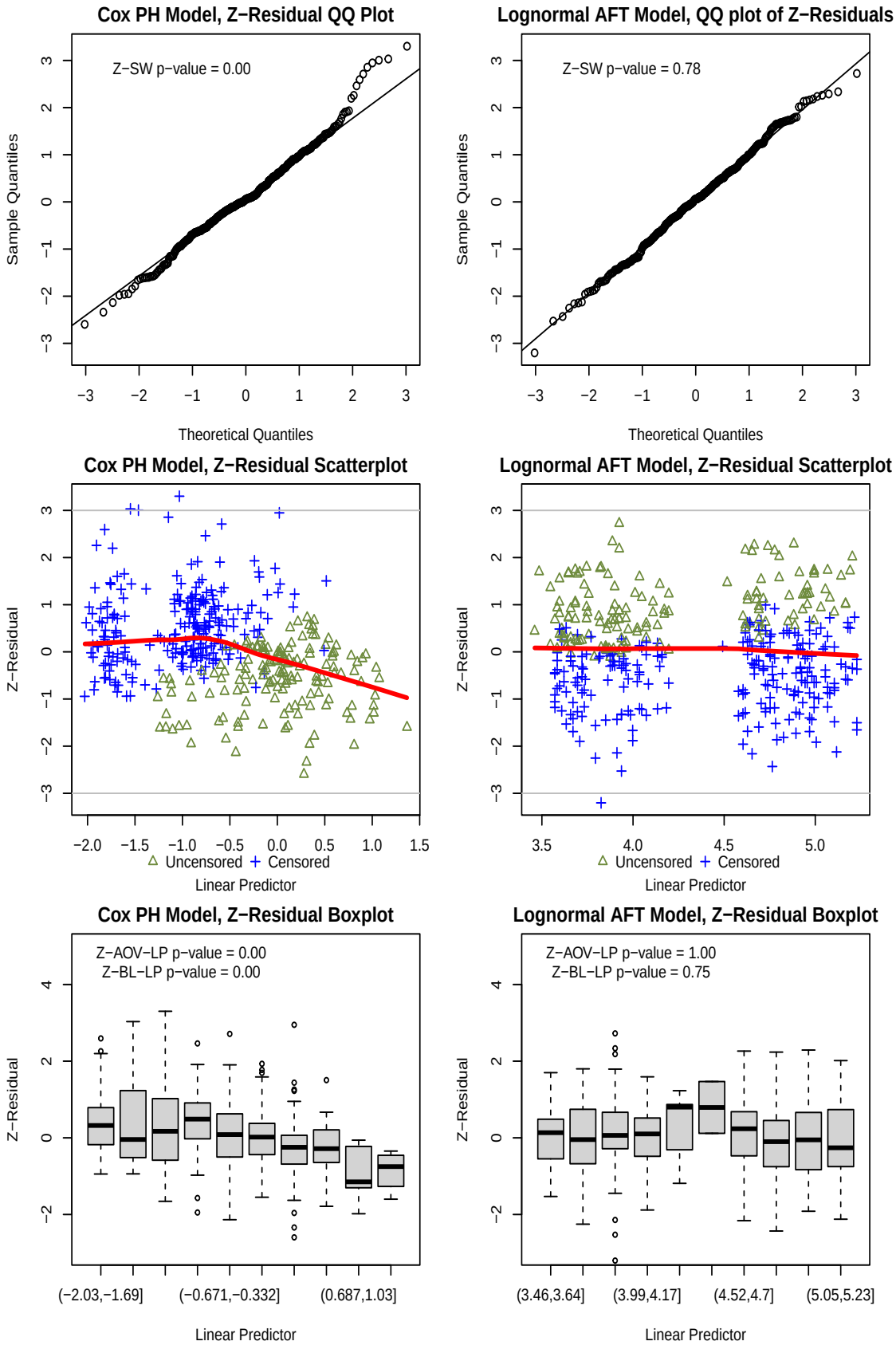


Figure 5.10: The global PH assumption diagnostics results for the Cox PH model with frailty (left panels) and the AFT Lognormal model (right panels) fitted to the diabetic retinopathy study dataset.

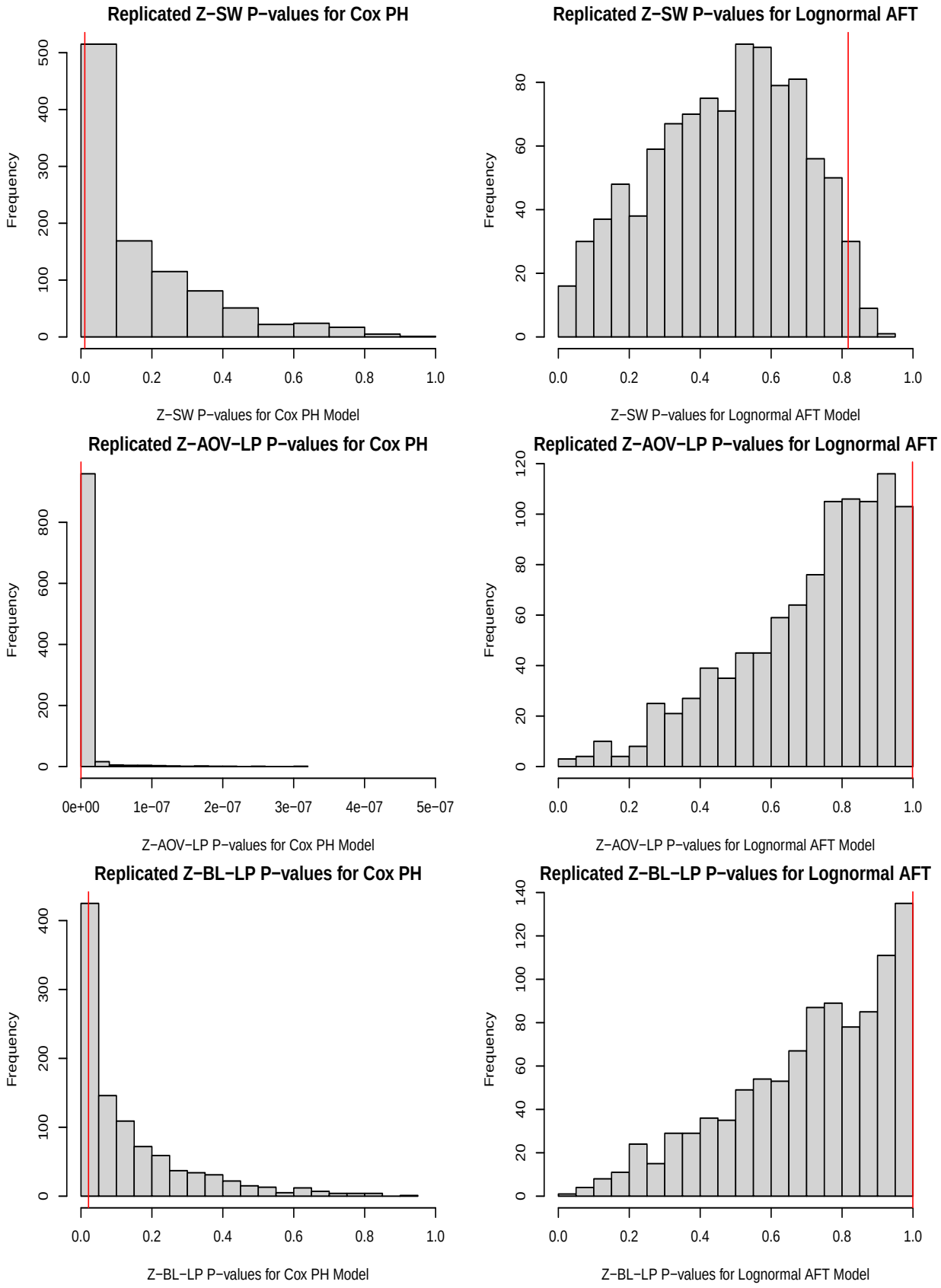


Figure 5.11: Histograms of 1000 replicated Z-SW, Z-AOV-LP and Z-BL-LP p-values for the Cox PH model with frailty (left panels) and the AFT lognormal model (right panels) fitted with the diabetic retinopathy study dataset. The vertical red lines indicate $p_{\min}$ for 1000 replicated p-values. Note that the upper limit of the x-axis for Z-AOV-LP p-values for the Cox PH model with frailty is 0.005, not 1 for others.

5.6 Discussions and Conclusions

In many epidemiological and medical studies, diagnosis of the Cox regression model is simply through checking the PH assumption. Schoenfeld residuals and score tests are commonly used graphical and quantitative diagnostic tools for checking the PH assumption. However, those diagnostic tools for detecting Non-PH were developed under the framework of the Cox regression model with time-varying covariate effects. In the situation when the Cox regression model does not fit the data, those diagnostic methods may not adequately detect Non-PH. In this paper, through extensive simulation studies, we showed that the power of the score test is generally high when the Non-PH is due to time-varying covariate effects in a Cox regression model, but it lacks power if the Non-PH is caused by other reasons, for instance, accelerated failure time.

To this end, we proposed the graphical and quantitative methods based on Z-residual to detect non-PH in a survival model. We developed overall GOF tests based on Z-residual for detecting Non-PH across all terms covariates included in a model, and also proposed non-homogeneity tests to test for trends in the Z-residuals against each covariate in order to detect non-PH for an individual covariate. Our simulation studies show that the overall GOF tests have high power and satisfactory type I error in detecting Non-PH due to both time-varying covariate effects and accelerated failure times. Additionally, the non-homogeneity test for the group variances of Z-residuals by the levels of the covariate values has high power in detecting Non-PH due to time-varying effects for individual covariates. The real data analysis shows that the Z-residual diagnostics successfully detected non-PH, which was not captured by the Schoenfeld residuals and score tests. Therefore, we view the Z-residual diagnostic tool and these tools as complementary rather than competitive in checking the PH assumption in the survival model.

For checking PH assumption, a number of diagnostics tools have been proposed in the literature as well. Lin, et.al [92] proposed the cumulative sums of Martingale residuals to check the validity of the PH assumption in the Cox Model. Lee, et. al [93] proposed model-checking tools based on the cumulative sums of mean zero stochastic processes for testing the PH assumption graphically and analytically under the setting of length-biased sampling. Scheike, et al [94] proposed new tests for testing the time-dependent effects of the covariates in the proportional hazards model, in which the tests are a natural and integrated part of an extended version of the Cox model. Comparing Z-residual for diagnosing the proportional hazard assumption with these existing residual diagnostics tools warrants a research topic in the future.

5.7 Additional Figures and Tables

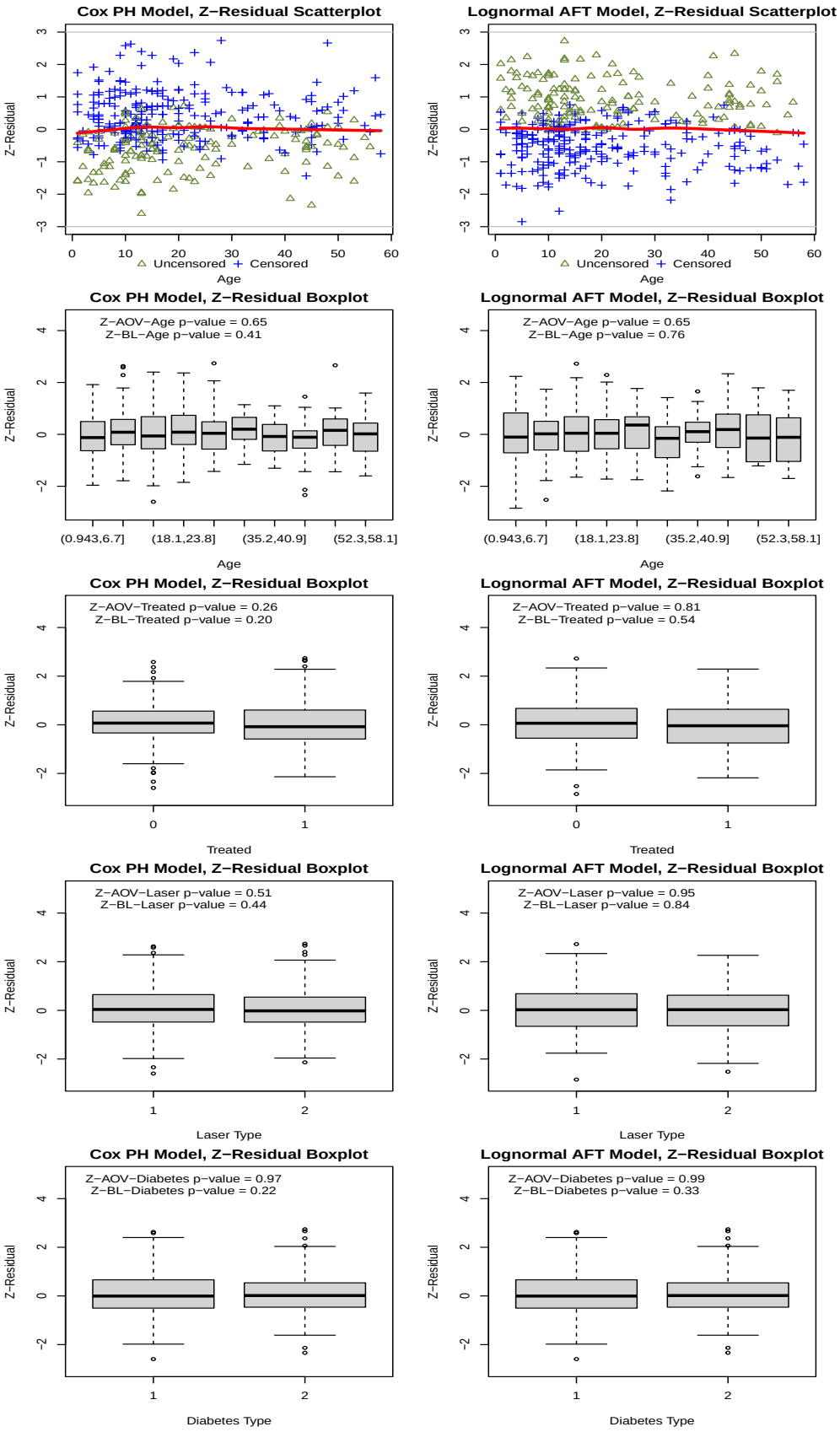


Figure 5.12: The PH assumption for each covariate diagnostics results for the Cox PH model with frailty (left panels) and the AFT lognormal model (right panels) fitted to the Diabetic Retinopathy study dataset.

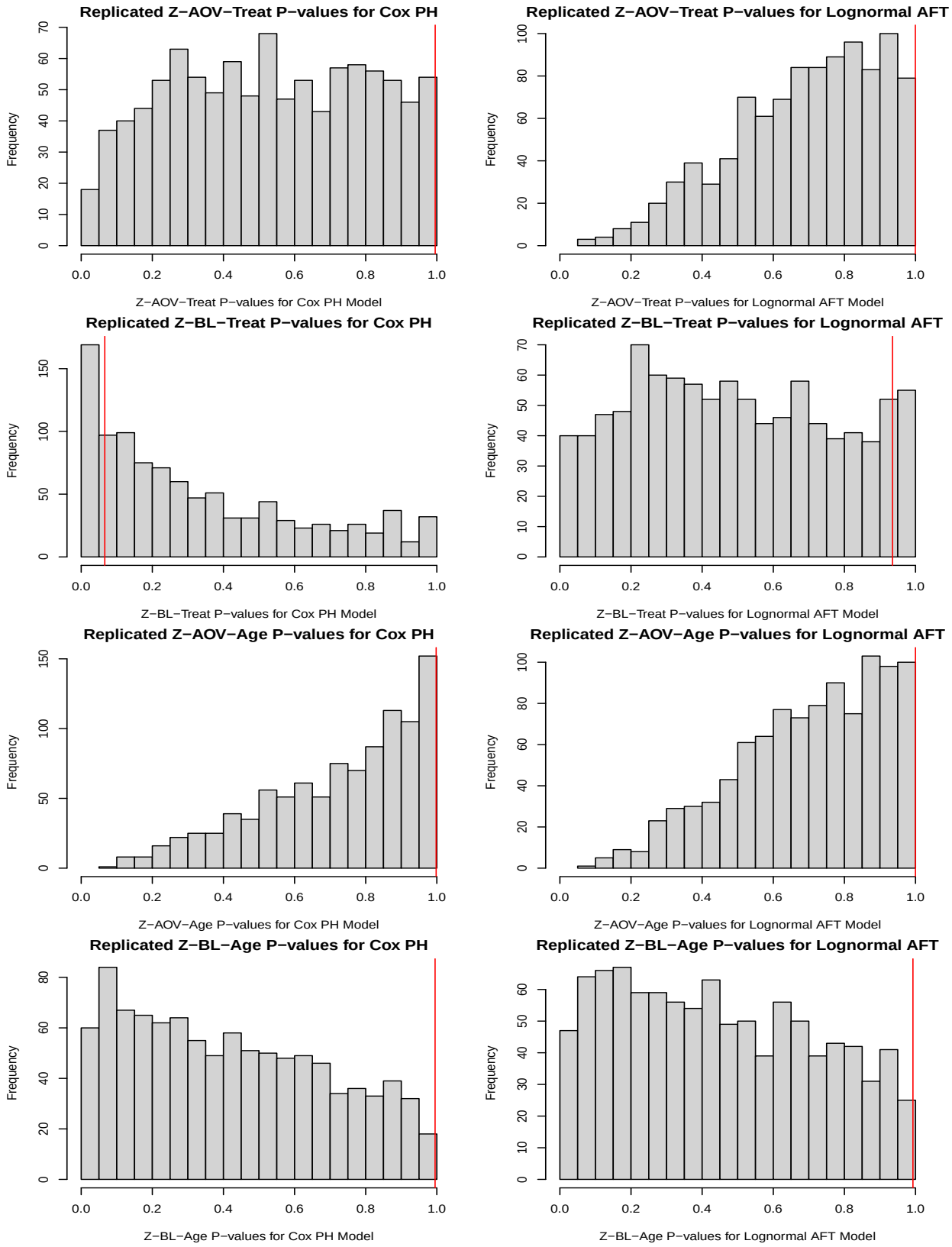


Figure 5.13: The histograms of 1000 replicated Z-AOV-Treat, Z-BL-Treat, Z-AOV-Age, and Z-BL-Age p-values for the Cox PH model with frailty (left panels) and the AFT lognormal model (right panels) fitted with the Diabetic Retinopathy study dataset. The vertical red lines indicate $p_{\min}$ for 1000 replicated p-values.

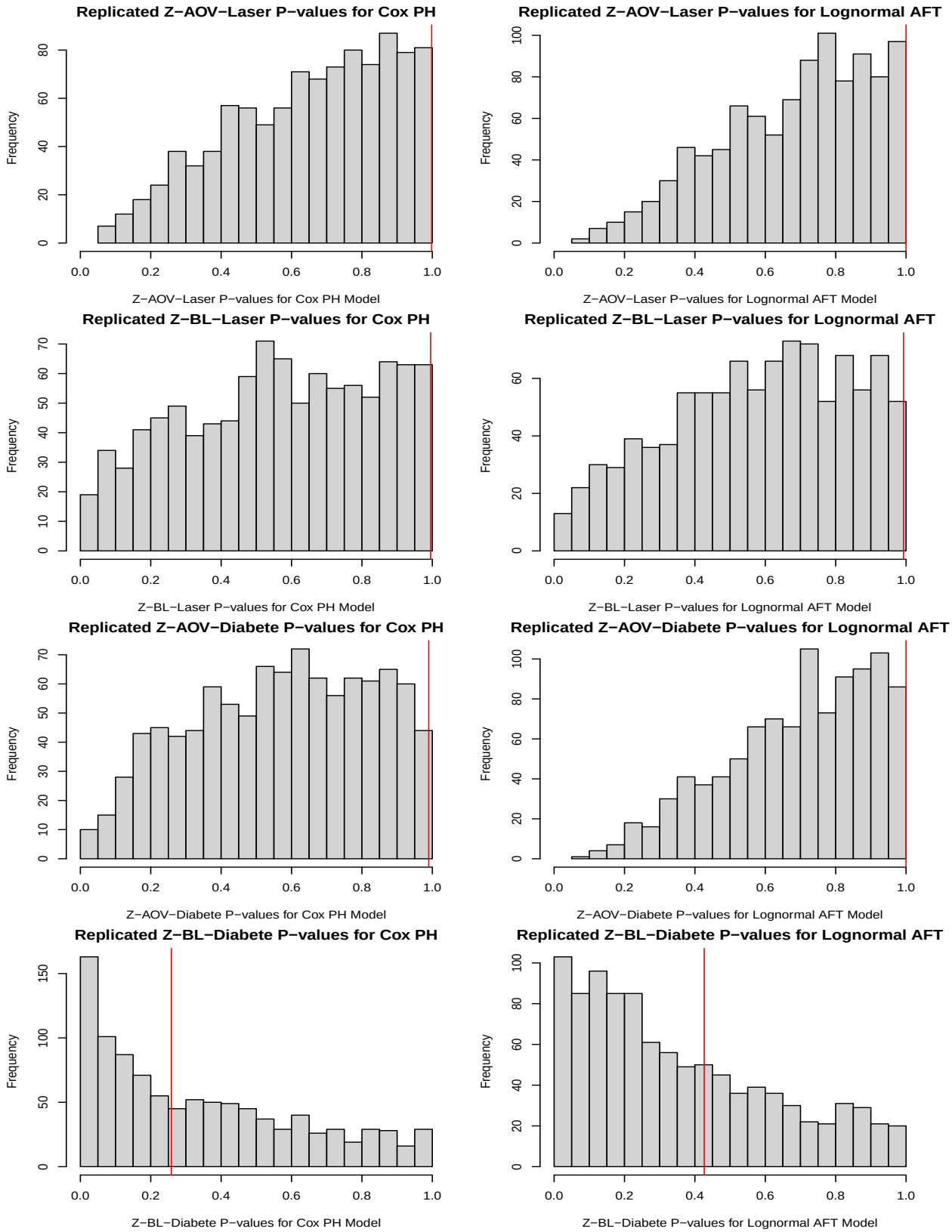


Figure 5.14: The histograms of 1000 replicated Z-AOV-Laser, Z-BL-Laser, Z-AOV-Diabete and Z-BL-Diabete p-values for the Cox PH model with frailty (left panels) and the AFT lognormal model (right panels) fitted with the Diabetic Retinopathy study dataset. The vertical red lines indicate $p_{\min}$ for 1000 replicated p-values.

6 Discussions and Conclusions

6.1 Discussions

In this thesis, Chapter 2 investigated currently available R packages used for fitting shared frailty models. The results showed that all packages provided similar and unbiased parameter estimates for fixed-effect regression coefficients, regardless of cluster sizes and censoring rates. However, there were differences in the estimation of the variance parameter for the frailty term, which was consistently underestimated for all R packages. The bias was less pronounced as the censoring rate increased, but subject to more variability, which leads to higher MSE. Increasing the number of clusters resulted in higher precision of the estimated variance parameter of the frailty term. Convergence rates were over 97% in most packages, except for the `parfm` and `frailtyEM` packages in the scenario with a small sample size and high censorship. Among all the packages, the `survival` package was found to be the best choice for estimating frailty model parameters due to its computational speed and high convergence rates in almost all simulation scenarios. However, it did not provide an estimate of the standard error for the variance component of the frailty. The `frailtyEM` package can be used instead of `survival` if the standard error of the frailty variance is required in a real application. The study proposed a new approach to compute the confidence interval for the frailty variance using the standard error of $\log\left(\hat{\theta}\right)$, which was found to have a much higher coverage probability than the standard error-based confidence interval. Most packages did not provide the standard error of $\log\hat{\theta}$, and the proposed approach provided a solution using the Fisher information approach. The study recommends adding this approach to R packages to calculate a more reliable 95% confidence interval for the frailty variance in frailty models. In summary, the `survival` package is the best choice for estimating frailty model parameters, and the proposed approach can be used to compute a reliable confidence interval for the frailty variance. Other packages, such as `frailtyEM`, `parfm`, `frailtySurv`, `frailtyHL`, and `frailtypack`, have their advantages and disadvantages and can be used depending on the specific needs of the application.

After fitting a shared frailty model, residual diagnosis is crucial to verify the validity of the model assumption. In Chapter 3, a diagnostic tool called Z-residual was developed to assess the covariate functional form in semi-parametric shared frailty models, building on the concept of randomized survival probability. This diagnostic tool can provide both graphical and numerical tests for examining the covariate functional form in the model. A general function was developed to calculate Z-residuals based on the output from the coxph function in the survival package in R. A non-homogeneity test was proposed to evaluate the trend in Z-residuals for checking the covariate functional form. Our simulation studies showed that proposed non-homogeneity tests based on Z-residuals were more effective in detecting misspecification in covariate functional form than overall goodness-of-fit tests. In an analysis of a real dataset, Z-residual diagnostics revealed that a model with log transformation was inappropriate for modelling the survival time of acute myeloid leukemia patients, which was not detected by other diagnostic methods. The p-values obtained from Z-residual tests were found to be random because of the randomization in generating Z-residuals, so a method for obtaining a p-value upper bound $p_{\min}$ was described. Although $p_{\min}$ can be informative when the model departure is evident,

it tends to be conservative, so the bootstrap method could be a more powerful alternative that warrants future research. In addition, Lin et al. [57] proposed the cumulative sums of martingale residuals to check the covariate functional form. Comparing the Z-residual with the cumulative sums of martingale residuals tool warrants another research topic in the future.

Survival analysis typically calculates residuals based on the entire dataset. However, this approach can lead to conservatism bias and hinder the ability to detect model fit inadequacy and outliers, particularly with small sample sizes or high censoring rates. To address this issue, cross-validation methods can be used. In Chapter 4, we developed cross-validation methods to compute Z-residuals for detecting model inadequacy and identifying outliers in the context of shared frailty models. We created a general function that calculates cross-validated Z-residuals using the output from the `coxph` function in the `survival` package in R. To compare the performance of cross-validatory (10-fold and LOOCV) Z-residuals and No-CV Z-residuals for the overall GOF test and outlier detection, we conducted simulation studies and applied the method to a real dataset. The results showed that residual diagnosis without cross-validation tends to be conservative for detecting model misspecification, while the CV method improves the power of the SW-test with Z-residuals in detecting model inadequacy and identifying outliers. However, cross-validation may slightly elevate the type-I error rates in SW tests with Z-residuals in shared frailty models. To address this issue, we can explore methods for improving the computation of cross-validatory Z-residuals or conducting SW tests with Z-residuals. Marginalizing the frailties when calculating the randomized survival probability could potentially reduce the type-I error rates. Investigating the performance of methods for computing Z-residuals with and without this marginalization could be a promising research topic for the future. Furthermore, since the proposed cross-validatory residuals have the potential to diagnose various types of regression models, it would be valuable to investigate their broader applicability in future research.

Previous studies on semiparametric shared frailty models have mostly assumed proportional hazards, but when the model includes time-varying coefficients or time-dependent explanatory variables, the assumption may not hold. Schoenfeld residuals and related tests are commonly used for diagnosing the proportional hazards assumption, but these methods only consider a specific type of violation caused by time-varying covariate effects. This limitation may affect the power of these methods to detect other types of violations, such as accelerated failure time. In Chapter 5, we proposed graphical and quantitative methods based on Z-residual to detect non-PH in survival models. We developed overall goodness-of-fit tests based on Z-residuals to detect non-PH across all covariates terms included in a model and also proposed non-homogeneity tests to test to identify trends in the Z-residuals against each covariate to detect non-PH for individual covariates. Our simulation studies show that, compared to the score tests related to Schoenfeld residuals, the tests based on Z-residuals have similar powers and type I error rates in the scenarios with time-varying covariate effects, but significantly higher power in scenarios with accelerated failure time. Our analysis of real data showed that the Z-residual diagnostics successfully detected non-PH, which was missed by the Schoenfeld residuals and score tests. Therefore, we view the Z-residual diagnostic tools as complementary rather than competitive in checking the PH assumption in survival models. Additionally, a number of diagnostic tools have been proposed in the literature for checking PH assumption and comparing Z-residuals for diagnosing proportional hazard assumption with other existing residual diagnostic tools warrants future research.

6.2 Conclusions

This thesis extensively investigated currently available R statistical packages for fitting shared frailty models and model diagnostic tools. The study found that the survival package is the best choice for estimating shared frailty models due to its superior performance in terms of accuracy and efficiency. Additionally, the thesis proposed a novel confidence interval for frailty variance estimates that can significantly enhance the reliability of such estimates, and it is recommended that this interval be incorporated into future R packages for frailty models. Furthermore, the study demonstrated the usefulness of the Z-residual as a diagnostic tool for assessing model fit and model assumptions, such as detecting the functional form of covariates, outlier detection, and proportional hazard assumption. The study also showed that the CV method can further improve the power of Z-residuals in detecting model inadequacy and identifying outliers. Overall, this thesis made significant contributions to the advancement of statistical methods for analyzing frailty models.

Availability of R Code and Datasets

R functions for computing Z-residuals and cross-validatory Z-residuals for the `coxph` objects with demonstration examples and the datasets used in this thesis are available on github: `https://github.com/tiw150/Z-residual.git`.

Bibliography

[1] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978. ISSN 0006-3444.

[2] Luc Duchateau and Luc. Duchateau. *The frailty model*. Statistics for biology and health. Springer Verlag, New York, 2008. ISBN 9780387728353.

[3] Alex Karagrigoriou. Frailty models in survival analysis. *Journal of Applied Statistics*, 38(12):2988–2989, 2011. ISSN 0266-4763.

[4] DD Hanagal. Modeling survival data using frailty models. *Statistical methods in medical research*, 24 (6):936–936, 2015. ISSN 0962-2802.

[5] Terry M Therneau. *A Package for Survival Analysis in R*, 2022. URL `https://CRAN.R-project.org/package=survival`. R package version 3.3-1.

[6] Theodor Adrian Balan and Hein Putter. **frailtyEM** : An *R* Package for Estimating Semiparametric Shared Frailty Models. *J. Stat. Soft.*, 90(7), 2019. ISSN 1548-7660. doi: 10.18637/jss.v090.i07.

[7] John V. Monaco, Malka Gorfine, and Li Hsu. General Semiparametric Shared Frailty Model: Estimation and Simulation with **frailtySurv**. *J. Stat. Soft.*, 86(4), 2018. ISSN 1548-7660. doi: 10.18637/jss.v086.i04.

[8] Il Ha, Do, Maengseok Noh, and Youngjo Lee. frailtyHL: A Package for Fitting Frailty Models with H-likelihood. *The R Journal*, 4(2):28, 2012. ISSN 2073-4859. doi: 10.32614/RJ-2012-010.

[9] Marco Munda, Federico Rotolo, and Catherine Legrand. **Parfm** : Parametric Frailty Models in *R*. *J. Stat. Soft.*, 51(11), 2012. ISSN 1548-7660. doi: 10.18637/jss.v051.i11.

[10] Virginie Rondeau, Yassin Mazroui, and Juan R. Gonzalez. **Frailtypack** : An *R* Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *J. Stat. Soft.*, 47(4), 2012. ISSN 1548-7660. doi: 10.18637/jss.v047.i04.

[11] David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 2015.

[12] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B, Methodological*, 30(2):248–275, 1968. ISSN 0035-9246.

[13] Terry M. Therneau, Patricia M. Grambsch, and Thomas R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, March 1990. ISSN 0006-3444. doi: 10.1093/biomet/77.1.147.

[14] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media, November 2013. ISBN 978-1-4757-3294-8.

[15] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. CRC Press, August 1989. ISBN 978-0-412-31760-6.

[16] DAVID SCHOENFELD. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982. ISSN 0006-3444.

[17] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[18] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[19] Longhai Li, Tingxuan Wu, and Cindy Feng. Model diagnostics for censored regression via randomized survival probabilities. *Statistics in medicine*, 40(6):1482–1497, 2021. ISSN 0277-6715.

[20] Rob Henderson. Analysis of Multivariate Survival Data. Philip Hougaard, Springer, New York, 2000. No. of pages: Xvii+542. Price: $84.95. ISBN 0-387-98873-4. *Statist. Med.*, 20(16):2533–2534, August 2001. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.938.

[21] Philip Hougaard. Frailty models for survival data. *Lifetime Data Anal*, 1(3):255–273, 1995. ISSN 1380-7870, 1572-9249. doi: 10.1007/BF00985760.

[22] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B, Methodological*, 34(2):187–220, 1972. ISSN 0035-9246.

[23] D. Y. LIN, L. J. WEI, and ZHILIANG YING. Accelerated failure time models for counting processes. *Biometrika*, 85(3):605–618, 1998. ISSN 0006-3444.

[24] James W. Vaupel, Kenneth G. Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, August 1979. ISSN 0070-3370, 1533-7790. doi: 10.2307/2061224.

[25] Luc Duchateau and Paul Janssen. Penalized Partial Likelihood for Frailties and Smoothing Splines in Time to First Insemination Models for Dairy Cows. *Biometrics*, 60(3):608–614, September 2004. ISSN 0006341X. doi: 10.1111/j.0006-341X.2004.00209.x.

[26] Samuli Ripatti and Juni Palmgren. Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood. *Biometrics*, 56(4):1016–1022, December 2000. ISSN 0006341X. doi: 10.1111/j.0006-341X.2000.01016.x.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B, Methodological*, 39(1):1–38, 1977. ISSN 0035-9246.

[28] JP KLEIN. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):795–806, 1992. ISSN 0006-341X.

[29] David M. Zucker, Malka Gorfine, and Li Hsu. Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory. *Journal of Statistical Planning and Inference*, 138(7):1998–2016, July 2008. ISSN 03783758. doi: 10.1016/j.jspi.2007.08.005.

[30] Malka Gorfine, David M. Zucker, and Li Hsu. Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika*, 93(3):735–741, 2006. ISSN 0006-3444.

[31] Il Do Ha, Youngjo Lee, and Jae-kee Song. Hierarchical likelihood approach for frailty models. *Biometrika*, 88(1):233–233, 2001. ISSN 0006-3444.

[32] Gerard J. van den Berg and Bettina Drepper. Inference for shared-frailty survival models with left-truncated data. *Econometric reviews*, 35(6):1075–1098, 2016. ISSN 0747-4938.

[33] K. F. Lam and Anthony Y. C. Kuk. A marginal likelihood approach to estimation in frailty models. *Journal of the American Statistical Association*, 92(439):985–990, 1997. ISSN 0162-1459.

[34] P Joly, D Commenges, and L Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics*, 54(1):185–194, 1998. ISSN 0006-341X.

[35] Virginie Rondeau, Daniel Commenges, and Pierre Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2):139–153, 2003. ISSN 1380-7870.

[36] Katharina Hirsch and Andreas Wienke. Software for semiparametric shared gamma and log-normal frailty models: An overview. *Computer methods and programs in biomedicine*, 107(3):582–597, 2011. ISSN 0169-2607.

[37] D. Y. Lin. On the Breslow estimator. *Lifetime Data Anal*, 13(4):471–480, December 2007. ISSN 1380-7870, 1572-9249. doi: 10.1007/s10985-007-9048-y.

[38] Philip S. Rosenberg. Hazard Function Estimation Using B-Splines. *Biometrics*, 51(3):874, September 1995. ISSN 0006341X. doi: 10.2307/2532989.

[39] J. O. Ramsay. Monotone regression splines in action. *Statistical science*, 3(4):425–441, 1988. ISSN 0883-4237.

[40] C. A. McGilchrist. REML Estimation for Survival Models with Frailty. *Biometrics*, 49(1):221, March 1993. ISSN 0006341X. doi: 10.2307/2532615.

[41] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B, Methodological*, 58(4):619–678, 1996. ISSN 0035-9246.

[42] IL DO HA, MAENGSEOK NOH, and YOUNGJO LEE. Bias reduction of likelihood estimators in semiparametric frailty models. *Scandinavian journal of statistics*, 37(2):307–320, 2010. ISSN 0303-6898.

[43] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. ISSN 0368-4245.

[44] Charles E. McCulloch and John M. Neuhaus. Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical science*, 26(3):388–402, 2011. ISSN 0883-4237.

[45] Liselotte Petersen, Thorkild I. A. Sørensen, Gert G. Nielsen, and Per Kragh Andersen. Inference Methods for Correlated Left Truncated Lifetimes: Parent and Offspring Relations in an Adoption Study. *Lifetime Data Anal*, 12(1):5–20, March 2006. ISSN 1380-7870, 1572-9249. doi: 10.1007/s10985-005-7217-4.

[46] Theodor A. Balan, Marianne A. Jonker, Paul C. Johannesma, and Hein Putter. Ascertainment correction in frailty models for recurrent events data: Ascertainment correction in frailty models for recurrent events data. *Statist. Med.*, 35(23):4183–4201, October 2016. ISSN 02776715. doi: 10.1002/sim.6968.

[47] E. H ESTEY, YU SHEN, and P. F THALL. Effect of time to complete remission on subsequent survival and disease-free survival time in aml, raeb-t, and raeb. *Blood*, 95(1):72–77, 2000. ISSN 0006-4971.

[48] Robin Henderson, Silvia Shimakura, and David Gorst. Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 97(460):965–972, 2002. ISSN 0162-1459.

[49] Theodor A Balan and Hein Putter. A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11):3424–3454, 2020. ISSN 0962-2802.

[50] Alessandra Nardi and Michael Schemper. New Residuals for Cox Regression and Their Application to Outlier Screening. *Biometrics*, 55(2):523–529, June 1999. ISSN 1541-0420. doi: 10.1111/j.0006-341X.1999.00523.x.

[51] Steven P. Millard. EnvStats: Package for Environmental Statistics, Including US EPA Guidance, July 2018. URL `https://CRAN.R-project.org/package=EnvStats`.

[52] Steven P. Millard. *EnvStats: An R Package for Environmental Statistics*. Springer, New York, NY, 2nd ed. 2013.. edition, 2013. ISBN 978-1-4614-8456-1.

[53] Yingwei Peng and Jeremy M. G. Taylor. Residual-based model diagnosis methods for mixture cure models. *Biometrics*, 73(2):495–505, 2017. doi: 10.1111/biom.12582.

[54] Sündüz Keleş and Mark R. Segal. Residual-based tree-structured survival analysis. *Statistics in Medicine*, 21(2):313–326, 2002. doi: 10.1002/sim.981.

[55] C. P. Farrington. Residuals for proportional hazards models with interval-censored survival data. *Biometrics*, 56(2):473–482, 2000. doi: 10.1111/j.0006-341X.2000.00473.x.

[56] A. C. Davison and A. Gigli. Deviance residuals and normal scores plots. *Biometrika*, 76(2):211–221, June 1989. ISSN 0006-3444. doi: 10.1093/biomet/76.2.211.

[57] D. Y. Lin, L. J. Wei, and Z. Ying. Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. *Biometrika*, 80(3):557–572, 1993. ISSN 00063444. doi: 10.2307/2337177.

[58] Martin Law and Dan Jackson. Residual plots for linear regression models with censored outcome data: A refined method for visualizing residual uncertainty. *Communications in Statistics - Simulation and Computation*, 46(4):3159–3171, April 2017. ISSN 0361-0918. doi: 10.1080/03610918.2015.1076470.

[59] Bryan E. Shepherd, Chun Li, and Qi Liu. Probability-scale residuals for continuous, discrete, and censored data. *The Canadian journal of statistics = Revue canadienne de statistique*, 44(4):463–479, December 2016. ISSN 0319-5724. doi: 10.1002/cjs.11302.

[60] Stephen L. Hillis. Residual plots for the censored data linear regression model. *Statistics in Medicine*, 14(18):2023–2036, 1995. ISSN 0277-6715. doi: 10.1002/sim.4780141808.

[61] G Caraux and O Gascuel. Bounds on distribution functions of order statistics for dependent variates. *Statistics & probability letters*, 14(2):103–105, 1992.

[62] Tomasz Rychlik. Stochastically extremal distributions of order statistics for dependent samples. *Statistics & probability letters*, 13(5):337–341, 1992.

[63] Ying Yuan and Valen E. Johnson. Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models. *Biometrics*, 68(1):156–164, 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01668.x.

[64] Ethics, 2023. URL `https://nursing.usask.ca/research-tools/ethics.php#top`.

[65] E. C. Marshall and D. J. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in medicine*, 22(10):1649–1660, 2003. ISSN 0277-6715.

[66] E. C. Marshall and D. J. Spiegelhalter. Identifying outliers in bayesian hierarchical models: a simulation-based approach. *Bayesian analysis*, 2(2), 2007. ISSN 1936-0975.

[67] Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, May 2017. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-016-9649-y. URL `https://link.springer.com/article/10.1007/s11222-016-9649-y`.

[68] Aki Vehtari and Andrew Gelman. Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*, July 2015. URL `http://arxiv.org/abs/1507.02646`.

[69] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, September 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL `https://doi.org/10.1007/s11222-016-9696-4`.

[70] Anna L. Smith, Tian Zheng, and Andrew Gelman. Prediction scoring of data-driven discoveries for reproducible research. *Statistics and Computing*, 33(1):11, December 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10154-7. URL `https://doi.org/10.1007/s11222-022-10154-7`.

[71] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, November 2014. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-013-9416-2.

[72] Longhai Li, Shi Qiu, Bei Zhang, and Cindy X Feng. Approximating cross-validatory predictive evaluation in bayesian latent variable models with integrated is and waic. *Statistics and computing*, 26(4):881–897, 2015. ISSN 0960-3174.

[73] Longhai Li, Cindy X. Feng, and Shi Qiu. Estimating cross-validatory predictive p-values with integrated importance sampling for disease mapping models. *Statistics in Medicine*, 36(14):2220–2236, January 2017. ISSN 1097-0258. doi: 10.1002/sim.7278.

[74] C. A MCGILCHRIST and C. W AISBETT. Regression with frailty in survival analysis. *Biometrics*, 47 (2):461–466, 1991. ISSN 0006-341X.

[75] John R Downs, Michael Clearfield, Stephen Weis, Edwin Whitney, Deborah R Shapiro, Polly A Beere, Alexandra Langendorfer, Evan A Stein, William Kruyer, Antonio M Gotto, Jr, and for the AFCAPS/TexCAPS Research Group. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of afcaps/texcaps. *JAMA : the journal of the American Medical Association*, 279(20):1615–1622, 1998. ISSN 0098-7484.

[76] Øyvind Holme, Magnus Løberg, Mette Kalager, Michael Bretthauer, Miguel A Hernán, Eline Aas, Tor J Eide, Eva Skovlund, Jørn Schneede, Kjell Magne Tveit, and Geir Hoff. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: A randomized clinical trial. *JAMA : the journal of the American Medical Association*, 312(6):606–615, 2014. ISSN 0098-7484.

[77] Terry Therneau. *A package for survival analysis in R*, 2023. URL `https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf`. R package version.

[78] Carine A Bellera, Gaëtan MacGrogan, Marc Debled, Christine Tunon de Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. Variables with time-varying effects and the cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10(1):20–20, 2010. ISSN 1471-2288.

[79] Xiaonan Xue, Xianhong Xie, Marc Gunter, Thomas E Rohan, Sylvia Wassertheil-Smoller, Gloria Y F Ho, Dominic Cirillo, Herbert Yu, and Howard D Strickler. Testing the proportional hazards assumption in case-cohort analysis. *BMC medical research methodology*, 13(1):88–88, 2013. ISSN 1471-2288.

[80] J.G Aerts, H Coington, N Lankheet, S Burgers, B Biesma, A.M Dingemans, A.D Vincent, O Dalesio, H.J Groen, and E.F Smit. A randomized phase ii study comparing erlotinib versus erlotinib with alternating chemotherapy in relapsed non-small-cell lung cancer patients: the nvalt-10 study. *Annals of oncology*, 24(11):2860–5, 2013. ISSN 0923-7534.

[81] MD Garassino, Marina Chiara, MD Martelli, Olga, PhD Broggini, Massimo, MD Farina, Gabriella, PhD Veronese, Silvio, PhD Rulli, Eliana, MD Bianchi, Filippo, MD Bettini, Anna, MD Longo, Flavia, MD Moscetti, Luca, MD Tomirotti, Maurizio, PhD Marabese, Mirko, PhD Ganzinelli, Monica, PhD Lauricella, Calogero, MD Labianca, Roberto, PhD Floriani, Irene, MD Giaccone, Giuseppe, MD Torri, Valter, MD Scanni, Alberto, and MD Marsoni, Silvia. Erlotinib versus docetaxel as second-line treatment of patients with advanced non-small-cell lung cancer and wild-type egfr tumours (tailor): a randomised controlled trial. *The lancet oncology*, 14(10):981–988, 2013. ISSN 1470-2045.

[82] Philippe Lambert, Dave Collett, Alan Kimber, and Rachel Johnson. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in medicine*, 23 (20):3177–3192, 2004. ISSN 0277-6715.

[83] L. J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992. ISSN 0277-6715.

[84] Terry M Therneau, Patricia Grambsch, and Jui-Chung Allen Li. Modeling survival data: Extending the cox model. *Sociological Methods & Research*, 32(1):117–120, 2003. ISSN 0049-1241.

[85] Michael Schemper. Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(4):455–465, 1992. ISSN 0039-0526.

[86] T. S. Breusch and A. R. Pagan. The lagrange multiplier test and its applications to model specification in econometrics. *The Review of economic studies*, 47(1):239–253, 1980. ISSN 0034-6527.

[87] Tingxuan Wu, Longhai Li, and Cindy Feng. Z-residual diagnostics for detecting misspecification of the functional form of covariates for shared frailty models. *arXiv*, page 2302.09106, 2023.

[88] W. J HUSTER, R BROOKMEYER, and S. G SELF. Modelling paired survival data with covariates. *Biometrics*, 45(1):145–156, 1989. ISSN 0006-341X.

[89] Kung-Yee Liang, Steven G. Self, and Yue-Cune Chang. Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society. Series B, Methodological*, 55(2):441–453, 1993. ISSN 0035-9246.

[90] D. Y. Lin. Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in medicine*, 13(21):2233–2247, 1994. ISSN 0277-6715.

[91] C. F. SPIEKERMAN and D. Y. LIN. Checking the marginal cox model for correlated failure time data. *Biometrika*, 83(1):143–156, 1996. ISSN 0006-3444.

[92] D. Y. LIN, L. J. WEI, and Z. YING. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572, 1993. ISSN 0006-3444.

[93] Chi Hyun Lee, Jing Ning, and Yu Shen. Model diagnostics for the proportional hazards model with length-biased data. *Lifetime data analysis*, 25(1):79–96, 2018. ISSN 1380-7870.

[94] Thomas H. Scheike and Torben Martinussen. On estimation and tests of time-varying effects in the proportional hazards model. *Scandinavian journal of statistics*, 31(1):51–62, 2004. ISSN 0303-6898.