Cross-validatory Z-Residual for Diagnosing Shared Frailty Models

Tingxuan Wu
Department of Mathematics and Statistics, University of Saskatchewan
and
Cindy Feng*
Department of Community Health and Epidemiology, Dalhousie University
and
Longhai Li† ‡
Department of Mathematics and Statistics, University of Saskatchewan

September 25, 2024

Abstract

Accurate model performance assessment in survival analysis is imperative for robust predictions and informed decision-making. Traditional residual diagnostic tools like martingale and deviance residuals lack a well-characterized reference distribution for censored regression, making numerical statistical tests based on these residuals challenging. Recently, the introduction of Z-residuals for diagnosing survival models addresses this limitation. However, concerns arise from conventional methods that utilize the entire dataset for both model parameter estimation and residual assessment, which may cause optimistic biases. This paper introduces cross-validatory Z-residuals as an innovative approach to address these limitations. Employing a cross-validation (CV) framework, the method systematically partitions the dataset into training and testing sets to reduce the optimistic bias. Our simulation studies demonstrate that, for goodness-of-fit tests and outlier detection, cross-validatory Z-residuals are significantly more powerful (e.g. power increased from 0.2 to 0.6). and more discriminative (e.g. AUC increased from 0.58 to 0.85) than Z-residuals without CV. We also compare the performance of Z-residuals with and without CV 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, which Z-residuals without CV fail to identify. The CV Z-residual is a more

*ORCID: https://orcid.org/0000-0003-4030-7413
†ORCID: https://orcid.org/0000-0002-3074-8584
‡Correspondence to: longhai.li@usask.ca

powerful tool than the No-CV Z-residual for checking survival models, particularly in goodness-of-fit tests and outlier detection. We have published a generic function, which is collected in an R package called `Zresidual`, for computing CV Z-residual for the output of the widely used `survival` R package.

Keywords: cross-validation, Z-residual, goodness-of-fit, model checking, residual diagnosis, survival models

List of Abbreviations: AIC, Alkaike's Information Criterion; CHF, cumulative hazard function; CS, Cox-Snell; GOF, goodness-of-fit; CV, cross-validation or cross-validatory; No-CV Z-residuals, Z-residuals without CV; SW, Shapiro-Wilk; RSP, randomized survival probability.

1 Introduction

Residual diagnosis is a critical step in statistical modelling for checking the validity of model assumptions. However, traditional residual analysis, relying on comparing observed and predicted values using the entire dataset, poses limitations. This method employs the same data for both fitting the model and assessing its adequacy, potentially introducing an optimistic bias in the evaluation of model fit. The issues stemming from the double use of the dataset have gained significant attention, particularly in the context of Bayesian model comparison (see Marshall and Spiegelhalter (2003, 2007); Piironen and Vehtari (2017); Vehtari et al. (2024, 2017); Smith et al. (2022); Gelman et al. (2014); Li et al. (2015, 2017), for example).

To address the issue of optimistic bias, the CV method is crucial in model development, mitigating reliance on the same dataset for both model fitting and assessment, particularly for predictive performance and detecting overfitting. K-fold CV and leave-one-out CV (LOOCV) are the most commonly used methods. In K-fold CV, the dataset is divided into k subsets. Iteratively, the model is fitted (trained) on $k-1$ subsets and then the fitted model is tested on the left-out subset. LOOCV is a specific case where k is set to be the sample size n. The key advantage of CV is that the information of the test dataset is rigorously excluded from the fitted model. In machine learning, CV is typically used to evaluate the predictive performance of the fitted model with unseen data, which is proven effective in detecting overfitting. Despite its widespread use in evaluating predictive performance, assessing model generalization, and identifying overfitting, CV techniques are less commonly utilized specifically for residual diagnosis. For residual diagnosis, its primary application is within the realm of diagnosing linear and generalized linear regression models (McCullagh and Nelder, 1989; Pierce and Schafer, 1986). In the context of linear regression, various diagnostic measures involved the CV concept — omitting a subset of observations, refitting the model, finally assessing the changes in residuals/coefficient estimates/fitted values. For example, the studentized deleted residuals (McCullagh and Nelder, 1989) and Cook's distances (Cook and Weisberg, 1982; Cook, 1986) are widely used in practice and prove more powerful in outlier detection and the identification of subtle patterns that may otherwise be concealed than the counterparts without CV. While these methodologies are

applied to detect influential observations in generalized linear models, the utilization of CV techniques for direct residual diagnosis in survival models remains relatively uncommon. This rarity is attributed to the complexity found within these models.

Several residual diagnostic tools have been commonly used for checking the survival models (Collett, 2015b), including Cox-Snell (CS) (Cox and Snell, 1968), martingale (Therneau et al., 1990), deviance (Therneau, 2000; McCullagh, 1989), Schoenfeld (Collett, 2015b; Schoenfeld, 1982) and scaled Schoenfeld (Grambsch and Therneau, 1994) residuals. However, there is a lack of residuals with a characterized reference distribution for censored regression. Li et al. (2021) and Wu et al. (2024+b) 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, has greater statistical power, and is more informative than the existing residual diagnostic tools. However, Wu et al. (2024+b) showed that overall goodness-of-fit tests based on Z-residuals have relatively low powers in detecting subtle non-linear covariate effects, especially in complex scenarios. The low power or conservatism is presumably attributed to the bias from the double use of the same dataset in calculating the Z-residuals.

To address the conservatism problem, this paper introduces an innovative approach — cross-validatory Z-residual for shared frailty models and investigates the difference between Z-residuals with and without cross-validation (CV) in overall goodness-of-fit tests and outlier detection. Shared frailty models incorporate random effects (frailties) to address unobserved heterogeneity (Vaupel et al., 1979; Therneau and Grambsch, 2013; Balan and Putter, 2020), and these frailties are shared among individuals within a cluster or group (Duchateau and Janssen, 2008; Karagrigoriou, 2011; Hanagal, 2015). We have developed a generic R function for calculating CV Z-residuals for the output from fitting a survival model using the `coxph` function in the `survival` R package (Therneau and Grambsch, 2013). Our CV approach ensures adequate representation of groups and other covariates in each fold. In our study design, we calculate Z-residuals using three methods: the full dataset (No-CV), 10-fold CV (10-fold), and LOOCV. Simulation studies are conducted to investigate the performance of these Z-residuals in detecting nonlinear covariate effects and identifying outliers through graphical visualization and Shapiro-Wilk (SW) tests. We fur-

ther compare the performance of No-CV Z-residuals and LOOCV Z-residuals in identifying outliers using a kidney infection dataset (Mcgilchrist and Aisbett, 1991).

The subsequent sections of this paper are organized as follows. Section 2 provides a brief review of shared frailty models. Section 3 presents the definition of CV Z-residuals along with a discussion of the algorithm for computing them. Section 4 details the results of simulation studies, exploring the performances of 10-fold and LOOCV Z-residuals. In Section 5, we present the results of applying LOOCV Z-residuals to identify outliers in a kidney infection dataset. Finally, Section 6 concludes the article with a discussion of future work.

2 Shared frailty models

In survival analysis, the hazard function describes the instantaneous risk of the event of interest for an individual, provided the individual has not previously experienced the event. The hazard function indirectly characterizes the distribution of the time to the event. The most widely used model for survival data is the Cox proportional hazard model (Cox, 1972). In practice, survival data are often not independent even after controlling for fixed-effect covariates in the model. The effect of unobserved heterogeneity of lifetimes is referred to as frailty, which constitutes an unobserved random effect that multiplicatively influences the hazard. This variance of the random effects indicates the degree of unobserved heterogeneity. The frailty model quickly gained popularity in public health, epidemiological, medical science, and environmental research (Karagrigoriou, 2011; Henderson, 2001; Duchateau and Janssen, 2008; Hougaard, 1995).

A shared frailty model is commonly used to model clustered survival data, where the frailties are common or shared among individuals within groups (Henderson, 2001; Duchateau and Janssen, 2008; Hougaard, 1995). For instance, unobserved genetic and environmental background factors that family members share often result in correlations among outcomes within family members, even after accounting for the observed covariates. 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, $i = 1, 2, \ldots, g$. The true failure time for the jth individual from the ith group is de-

noted as T_{ij}^*, which we assume to be a continuous random variable, where $j = 1, 2, \ldots, n_i$. Let t_{ij}^* be 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} is the corresponding censoring time and assumed to be independent of T_{ij}^*. In the scenario of right censoring, the observed failure times are represented by the pairs (T_{ij}, δ_{ij}), where the observed event time and the non-censoring indicator are denoted as $T_{ij} = \min(T_{ij}^*, C_{ij})$ and $\delta_{ij} = I(T_{ij}^* < C_{ij})$, respectively. The observed data can be succinctly expressed as $t = (t_{11}, \ldots, t_{gn_g})$ and $\delta = (\delta_{11}, \ldots, \delta_{gn_g})$.

For a shared frailty model, the conditional hazard function of the failure time T_{ij}^* for the jth individual, $j = 1, 2, \ldots, n_i$, in the ith group, denoted as $h_{ij}(t|x_{ij}, z_i)$ and abbreviated as $h_{ij}(t)$ for simplicity, is given by

$$h_{ij}(t) = z_i \exp(\beta^T x_{ij}) h_0(t), \tag{1}$$

where x_{ij} is a column vector of values of p explanatory variables for the jth individual in the ith group; β is a column vector of regression coefficients for x_{ij}; $h_0(t)$ is a baseline hazard function; and z_i is the frailty term that is common for all n_i individuals within the ith group. Shared frailty models require a distribution for the frailty z_i, which is often assumed to be a gamma distribution (Collett, 2015b). The gamma distribution (Johnson and Kotz, 1977) makes it easy to obtain a closed-form representation of the observable survival, cumulative density, and hazard functions due to the simplicity of the Laplace transform (Balan and Putter, 2020). The conditional survival function for the jth individual of the ith group at time t, denoted as $S_{ij}(t|x_{ij}, z_i)$ and abbreviated as $S_{ij}(t)$ for simplicity, is given as follows:

$$S_{ij}(t) = \exp\left\{ -\int_0^t h_{ij}(\tau)\, \mathrm{d}\tau \right\} = \exp\left\{ - z_i \exp(\beta^T x_{ij}) H_0(t) \right\}, \tag{2}$$

where $H_0(t)$ is the baseline cumulative hazard function (CHF) of $h_0(t)$. Our definition of Z-residual is based on an estimate of the above conditional survival function with an estimate of z, β, and $H_0(t)$.

Shared frailty models can be categorized as either semiparametric or parametric based on the underlying nature of the baseline hazard. In semiparametric models, no specific assumptions are made regarding the baseline hazard. However, parametric models rely on

a predefined parametric distribution, and flexible parametric methodologies utilize spline-based estimators to model the baseline hazard (Hougaard, 1995; Balan and Putter, 2020). For this study, we focus on the semiparametric shared frailty models, the baseline CHF is estimated using the Breslow estimator (Lin, 2007a), suitable for continuous event times with few or no tied event times.

Shared frailty models can be fitted with many methods, which include penalized partial likelihood (Ripatti and Palmgren, 2000), the EM algorithm, the pseudo-likelihood approach (Gorfine et al., 2006), and the h-likelihood method (Ha et al., 2001). Wu et al. (2024+a) provides a comprehensive comparison study of available R packages for fitting shared frailty models with extensive simulation studies. Based on the results of Wu et al. (2024+a), we chose the `coxph` function from the widely used `survival` R package, which employs the penalized partial likelihood method to estimate model parameters.

3 Cross-validatory Z-residual

The Z-residuals are derived from the concept of randomized survival probability (RSP) as introduced in Li et al. (2021) and Wu et al. (2024+b). The RSP concept involves substituting the survival probability (SP) of a censored failure time with a uniformly distributed random number between 0 and the SP of the censored time. RSPs exhibit a uniform distribution on the interval (0, 1) under the true model with the true generating parameters. The RSP for t_{ij} for the jth individual in the ith group in a shared frailty model is defined as:

$$S_{ij}^{R}(t_{ij},\delta_{ij},U_{ij}) = \begin{cases} \hat{S}_{ij}(t_{ij}), & \text{if } t_{ij} \text{ is uncensored, i.e., } \delta_{ij}=1, \\ U_{ij}\,\hat{S}_{ij}(t_{ij}), & \text{if } t_{ij} \text{ is censored, i.e., } \delta_{ij}=0, \end{cases} \tag{3}$$

where U_{ij} represents a uniform random number in the range (0, 1), and $\hat{S}_{ij}(\cdot)$ is an estimated survival function for t_{ij} given x_{ij} and z_i as defined in eqn. (2). $S_{ij}^{R}(t_{ij},\delta_{ij},U_{ij})$ is a random number between 0 and $S_{ij}(t_{ij})$ when t_{ij} is censored. RSPs have been proven to be independently and uniformly distributed on the interval $(0,1)$ given x_{ij} and z_i under the true shared frailty model for clustered survival data (Wu et al., 2024+b). Therefore, these RSPs can be transformed into residuals with any desired distribution. We transform RSPs

with the negative normal quantile:

$$r_{ij}^{Z}(t_{ij}, \delta_{ij}, U_{ij}) = -\Phi^{-1}(S_{ij}^{R}(t_{ij}, \delta_{ij}, U_{ij})), \tag{4}$$

where $\Phi^{-1}(\cdot)$ represents the quantile function of the standard normal distribution. The residuals, as defined in (4), are called Z-residuals, which approximately follow the standard normal distribution when the model (i.e.,$S_{ij}(\cdot)$) is correctly specified. The negative sign before $\Phi^{-1}(\cdot)$ is added on purpose. The survival function $S_{ij}(t)$ is a decreasing function of t. Transforming $S_{ij}(t)$ with a positive normal quantile results in that smaller survival times t_{ij} corresponds to larger Z-residuals. Adding a negative sign reverses this relationship so that smaller survival times correspond to smaller Z-residuals, enabling a similar interpretation of Z-residuals as Pearson's residuals.

The CV method involves partitioning a dataset into training and testing sets, commonly implemented through K-fold CV or Leave-one-out CV (LOOCV). LOOCV is a straightforward method in which each observation is used as a test case for testing or evaluating the model fitted to the remaining data. LOOCV repeats training the model for each observation held out as a test case, hence, it could be slow for datasets with large sample sizes. Due to the time-consuming computational process of LOOCV, the K-fold CV method is often preferred for datasets with large sample sizes. In contrast, K-fold CV randomly divides the observations into k folds. It uses the observations in $k-1$ folds for fitting a model and the remaining fold is used to validate the fitted model, which is less time-consuming than LOOCV as the model fitting process needs to be repeated for only k times rather than n in LOOCV. In our study, we employ both LOOCV and 10-fold CV techniques to compute CV Z-residuals and compare their performance.

In LOOCV Z-residual, one observation, t_{ij}^{test}, is excluded from the dataset with n observations. The remaining observations, acting as the training dataset, are used for parameter estimation in the shared frailty model. Fitting the model to the training dataset produces the estimated regression coefficients, $\hat{\beta}$, and frailty effects, $\hat{z}_i$. The Breslow estimator (Lin, 2007b) estimates the cumulative baseline hazard function, $\hat{H}_0(\cdot)$. The survival function $\hat{S}_{ij}(t_{ij})$ for the test observation t_{ij}^{test} of the jth individual in the ith group is computed using:

$$\hat{S}_{ij}(t_{ij}^{test}) = \exp\{-\hat{z}_i \exp(\hat{\beta}^T x_{ij}) \hat{H}_0(t_{ij}^{test})\}. \tag{5}$$

Subsequently, the RSP for the observed t_{ij}^{test} is defined as:

$$\hat{S}_{ij}^{R}(t_{ij}^{test}, d_{ij}, U_{ij}) = \begin{cases} \hat{S}_{ij}(t_{ij}^{test}), & \text{if } t_{ij}^{test} \text{ is uncensored, i.e., } d_{ij} = 1, \\ U_{ij}\,\hat{S}_{ij}(t_{ij}^{test}), & \text{if } t_{ij}^{test} \text{ is censored, i.e., } d_{ij} = 0. \end{cases} \tag{6}$$

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

$$\hat{r}_{ij}^{Z}(t_{ij}^{test}, d_{ij}, U_{ij}) = -\Phi^{-1}(\hat{S}_{ij}^{R}(t_{ij}^{test}, d_{ij}, U_{ij})). \tag{7}$$

We repeat the above calculation of LOOCV Z-residual for each observation in each group, i.e., t_{ij}, being held out as a test observation t_{ij}^{test}. In particular, the estimate of the survival function, $\hat{S}_{ij}(\cdot)$, is different when different observations in the same group are held out as a test case.

In K-fold CV, the full dataset is divided into k groups of approximately equal size. One group is designated as the test dataset, and the other $k-1$ groups form the training dataset for fitting the shared gamma frailty model. The steps for obtaining estimates ($\hat{\beta}$, $\hat{z}_i$, and $\hat{H}_0$) and calculating Z-residuals are identical to LOOCV, for which each observation t_{ij} forms a fold.

To ensure well-distributed observations within groups and consistent values in both training and test datasets, the process aims to match cluster identities and categorical covariate values between the two sets. Additionally, it avoids situations where certain clusters or categorical covariate categories have no observed failure times. For cluster-based or categorical covariate values, specific considerations are employed during the LOOCV and K-fold CV methods. Clusters with only one observation cannot be included in the training dataset, and similar requirements are imposed on categorical covariates. As such, CV Z-residuals for these observations are designated as NA in the implementation.

We have included an R function for computing CV Z-residuals for the output of the `coxph` function in the `survival` R package in the Supplementary Materials of this paper. We have also collected this R function into an R package called `Zresidual`, which can be downloaded and installed directly from GitHub via this link: https://github.com/tiw150/Zresidual . For further details on using the function and the `Zresidual` package, please refer to our demonstration available on this webpage: https://tiw150.github.io/CV_Zresidual_demo.html.

4 Simulation Studies and Results

4.1 Detecting Non-linear Covariate Effects

In this section, we use simulated datasets to investigate the difference between CV and No-CV Z-residuals in detecting non-linear covariate effects.

4.1.1 Generating Datasets with Logarithmic Effects and Model Fitting

The original failure times t_{ij} are generated from a Weibull regression model as

$$t_{ij} = \left(\frac{-\log(v_{ij})}{\lambda z_i \exp(x_{ij}^{(1)} + \beta_2 \log(x_{ij}^{(2)}) + 0.5 x_{ij}^{(3)})} \right)^{1/\alpha} \tag{8}$$

for the jth individual in the ith group, where $i \in \{1, \ldots, 10\}$ and $j \in \{1, \ldots, m\}$, and v_{ij} is simulated from Uniform(0, 1). We chose the shape parameter $\alpha = 3$ and the scale parameter $\lambda = 0.007$ by following the work by Hirsch and Wienke (2011). The equation (8) is derived by following the inverse-CDF method for generating random numbers from the Weibull distribution; a detailed derivation is given in the Supplementary Materials.

The censoring times C_i is simulated from an exponential distribution, $\exp(\theta)$, 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. The methods for generating the covariates and frailties also follow the work by Hirsch and Wienke (2011).

We consider fitting a wrong model $h_{ij}(t) = z_i \exp(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)}) h_0(t)$ and the true model $h_{ij}(t) = z_i \exp(\beta_1 x_{ij}^{(1)} + \beta_2 \log(x_{ij}^{(2)}) + \beta_3 x_{ij}^{(3)}) h_0(t)$ to the simulated datasets.

4.1.2 Visualizing CV and No-CV Z-residuals of a Single Dataset

We first examine the CV and No-CV[1] Z-residuals of a single dataset in which a strong non-linear covariate effect ($\beta_2 = -2$) is present. The dataset comprises 10 clusters with 50 observations (total sample size $n = 500$). The scatterplots displaying Z-residuals against the covariate $x_{ij}^{(2)}$ are depicted in Fig. 1. Under the true model, the scatterplots of CV

[1]We use the term "No-CV Z-residuals" throughout this paper to stand for "Z-residuals without CV".

and No-CV Z-residuals exhibit a random distribution without any specific patterns, mainly concentrated within the interval (-3, 3), as expected for a random sample from the standard normal[1]. Notably, most Z-residuals cluster to the left of the x-axis, given that $x_{ij}^{(2)}$ was generated from a positive-normal (0, 1) distribution. In contrast, under the wrong model, the scatterplots of all three types of Z-residuals reveal a discernible non-linear pattern. However, there is a notable disparity in the magnitude of CV and No-CV Z-residuals for a specific observation, for which the CV Z-residual is close to the value of 6, whereas the corresponding No-CV Z-residual is about only 3. Additionally, more CV Z-residuals than No-CV Z-residuals are greater than 3, indicating a more conservative nature of the No-CV Z-residuals.

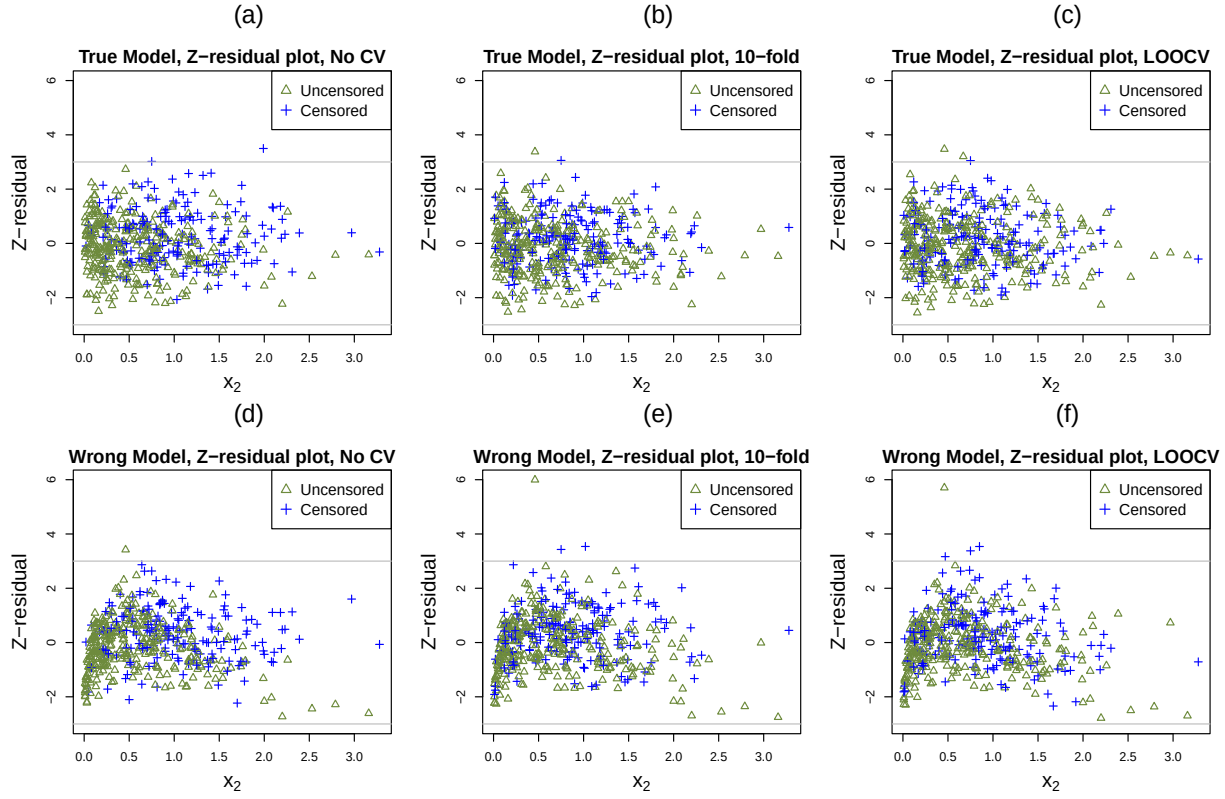

Figure 1: The scatterplots depict the No-CV, 10-fold, and LOOCV Z-residuals for a simulated dataset with a non-linear covariate effect, detailed in Section 4.1. The dataset comprises 10 clusters of 50 observations with a censoring rate $\approx$ 50%. Gray horizontal lines are drawn at values 3 and -3 for reference.

[1] 99.73% of observations from the standard normal should be within (-3,3)

All the QQ plots of CV and No-CV Z-residuals under the true model (Fig. S1 in the Supplementary Materials) align closely with the 45° straight line, affirming their normality, as expected when a correct model is fitted. In contrast, the QQ plots of 10-fold and LOOCV Z-residuals under the wrong model reveal increased deviations in the upper tail than No-CV Z-residuals. The difference in the QQ plots indicates that the CV Z-residuals have greater power in detecting non-linear covariate effects than the No-CV Z-residuals.

4.1.3 Comparing the Model Rejection Rates of Z-residual-based SW Tests

We further used multiple simulated datasets to investigate the distinction between CV and No-CV Z-residuals in GOF tests. The Shapiro–Wilk (SW) test was applied to assess the normality of the three types of Z-residuals for checking the overall GOF for fitted models. In our investigation, we generated 1000 datasets, each comprising 10 clusters of m observations. The value of m varied within the range of $10, 20, \ldots, 100$. Additionally, we set two distinct values for β_2 (-2 and -1), representing strong and moderate non-linear covariate effects.

Fig. 2 illustrates the results for the scenario with a pronounced non-linearity effect ($\beta_2 = -2$). Under the true model, the model rejection rates for the No-CV Z-residuals approximate the 0.05 nominal level. Conversely, under the wrong model, although the model rejection rates for the No-CV Z-residuals gradually increase with sample size, they remain significantly low. However, under the wrong model, the model rejection rates of LOOCV or 10-fold CV Z-residuals (Fig. 2(b)(c)) are notably higher than those of No-CV Z-residuals (Fig. 2(a)), across all sample sizes. In addition, we also see that the difference in the means of SW p-values between the true and wrong models (the gap between blue and red lines as depicted in Fig. 2(d)-(f)) is notably larger for CV Z-residuals compared to No-CV Z-residuals across all sample sizes.

We observed that SW tests employing CV Z-residuals exhibit slightly higher type-I error rates than the nominal level, especially noticeable with small sample sizes (Fig. 2). This increase is likely due to finite-sample errors encountered in estimating model parameters. Notably, parameter estimation can be particularly challenging with reduced sample sizes within the CV approach. Theoretically, Z-residuals conform to an exact normal distribution

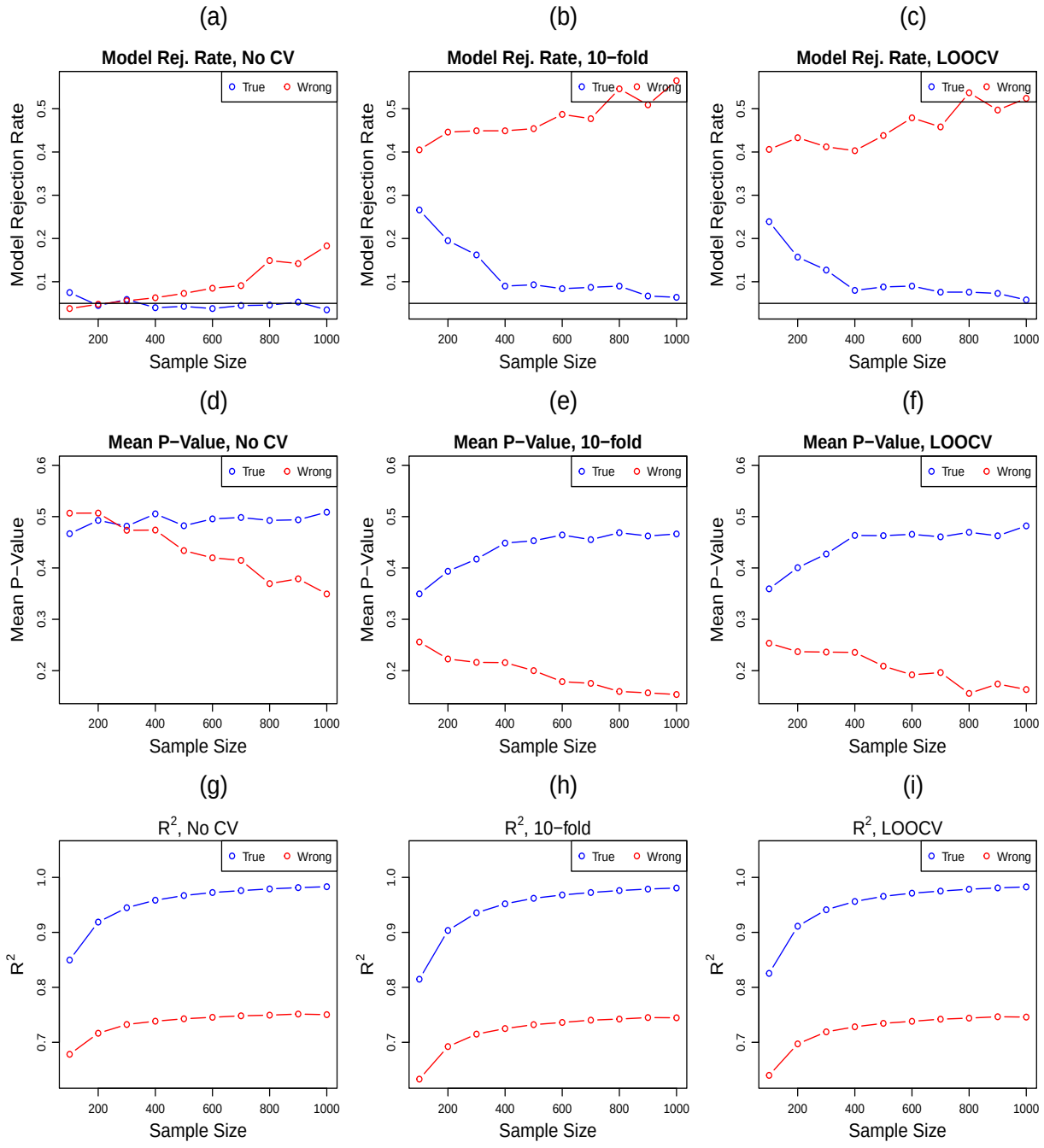

Figure 2: Comparison of model rejection rates (proportions of Z-residual-based SW test p-values ≤ 0.05) and the means of Z-residual-based SW p-values for detecting the non-linear covariate effect. Fig. (g)-(i) show the values of R^2 for measuring the agreement between the survival probabilities calculated with the fitted models and the survival probabilities calculated with the true generating models.

when calculated with the true model and corresponding parameters. However, when the sample size is small, a fitted model may not accurately represent the true model due to the errors in estimating the parameters, although the model specification (e.g. family and covariate functional form) is correct.

To delineate the discrepancy between the fitted and true models, we calculated the R^2, the square of the correlation, between the actual survival probabilities and the estimated survival probabilities. The actual survival probability is computed using the true values of the parameters ($\alpha = 3, \lambda = 0.007, \beta_1 = 1, \beta_2 = -2$ and $\beta_3 = 0.5$), which were used to simulate the datasets. The estimated survival probability is calculated by using the estimated parameters (i.e., $\hat{\alpha}, \hat{\lambda}, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$) from fitting either the true or wrong models. The plots in Fig. 2(g)(h)(i) present the average of the R^2 values across 1000 datasets and various CV folds for each simulation setting. We see that the R^2 for the cases with small sample sizes is pretty small, e.g. a value of 0.8. The difference in the fitted model and the true model may explain why the type-I error rates of the SW tests with CV Z-residuals are larger than the nominal level of 0.05.

4.1.4 Comparing the Discriminativeness of Z-residual-based SW Tests

We further employ the area under the ROC curve (AUC) to quantify the separability of the two sets of 1000 SW test p-values: one from fitting the true model and the other from fitting the wrong model. AUC is a quantity to measure the correlation between a binary response variable and a continuous variable called predictive probability, which is between 0 and 1. We create an artificial response variable by labelling 1 for the SW test p-values associated with the true model and 0 for those with the wrong model. The SW test p-values themselves are treated as the predicted probabilities. We employ the `pROC` package (Robin et al., 2023) to calculate the AUC for the response values and predictive probabilities as defined above. High AUC values indicate that the SW p-values effectively distinguish between the correct and incorrect models. Fig. 3 displays the AUC values for scenarios with strong and moderate non-linear covariate effects in the left and right plots, respectively. As the sample size increases, the AUC of all three methods rises, with the AUC values of the 10-fold and LOOCV Z-residuals closely aligned. Notably, the AUC

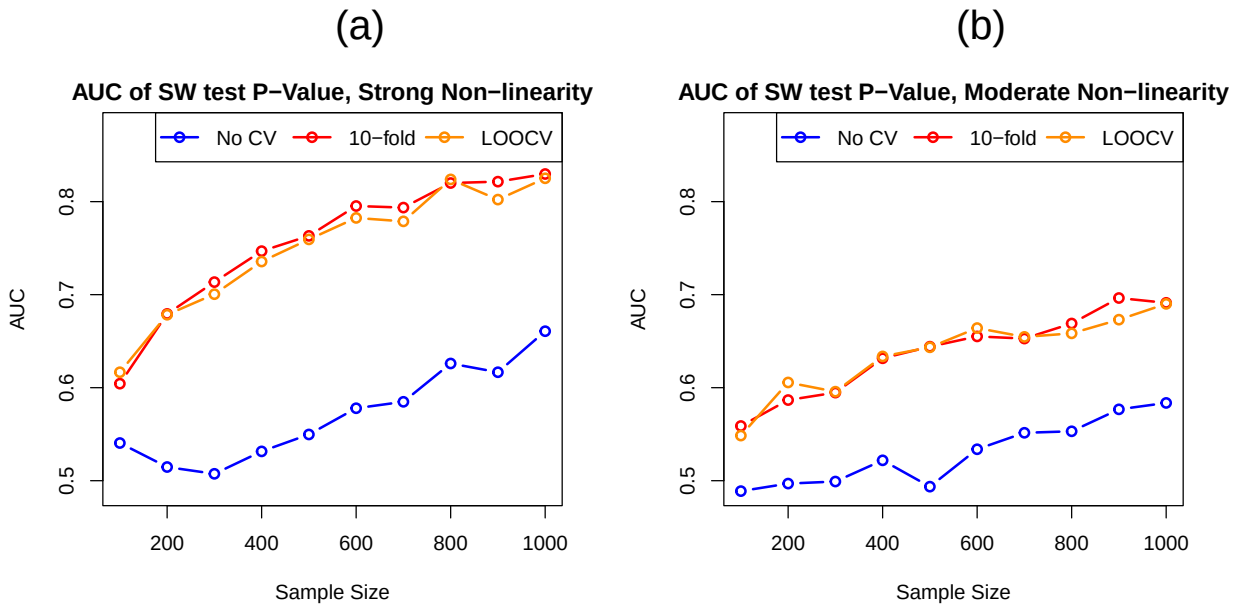

Figure 3: Comparison of the AUC values of Z-residual-based SW test p-values in predicting the correct and incorrect models for simulated datasets with logarithmic covariate effects.

values for 10-fold and LOOCV Z-residuals consistently surpass those of No-CV Z-residuals. Moreover, the superior performance of CV Z-residuals remains steady even with increased sample sizes, at least up to 1000. This finding is noteworthy, considering the assumption that the bias due to the double use of the dataset might diminish with larger sample sizes. In summary, the SW p-values computed with CV Z-residuals exhibit higher discriminative abilities in distinguishing the correct and incorrect models compared to those calculated with No-CV Z-residuals.

4.2 Detecting Outliers

4.2.1 Generating Datasets with Outliers and Model Fitting

We generate a clean dataset based on a Weibull model and then introduce perturbations to create a corresponding contaminated dataset, where the outlier identities are known. The clean datasets are constructed by generating true failure times from a Weibull regression model as follows:

$$t_{ij} = \left(\frac{-\log(v_{ij})}{\lambda z_i \exp(x_{ij}^{(1)} - 2x_{ij}^{(2)} + 0.5x_{ij}^{(3)})} \right)^{1/\alpha} \tag{9}$$

for the jth individual in the ith group, where $i \in \{1, \ldots, 10\}$ and $j \in \{1, \ldots, m\}$, and v_{ij} is simulated from Uniform (0, 1). We chose the shape parameter $\alpha = 3$ and the scale parameter $\lambda = 0.007$ by following Hirsch and Wienke (2011). The equation (9) is derived by following the inverse-CDF method for generating random numbers from the Weibull distribution; a detailed derivation is given in the Supplementary Materials.

The censoring times C_{ij} is simulated from an exponential distribution, $\exp(\theta)$, with θ set to obtain censoring rates approximately equal to 50%. The three covariates are generated in a similar way as used by Hirsch and Wienke (2011): $x_{ij}^{(1)}$ 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.

Jitters, meant to represent outliers, are added according to the formula $\max(w, e)$, where e is a random number from an exponential distribution with a rate of 1, and w is set to 2 or 4 to indicate moderate and strong jitters, respectively. This approach ensures that the jitters are at least greater than w. Additionally, we consider two methods for adding jitters to the clean datasets: one entails adding jitters to randomly chosen 10% event times, while the other involves adding jitters 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.

The Cox model with hazard function given by $h_{ij}(t) = z_i \exp(\beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \beta_3 x_{ij}^{(3)}) h_0(t)$ is a true model for the clean datasets. We fit this model to both clean datasets and contaminated datasets to compare the performance of CV and No-CV Z-residuals. When the true model is fitted to contaminated datasets, it is a wrong model for the datasets due to the the added jitters.

4.2.2 Visualizing CV and No-CV Z-residuals of a Single Dataset

To illustrate the performance of CV and No-CV Z-residuals, we first examine the Z-residuals of a pair of clean and contaminated datasets with a cluster size of $m = 20$. In the contaminated dataset, strong jitters are introduced to ten randomly selected failure times. Fig. 4 indicates that the Z-residuals for the clean dataset mainly fall within the range of -3 and 3 without any unusual patterns, behaving as independent standard normal variates.

For the contaminated dataset, all No-CV Z-residuals of the contaminated dataset remain constrained within the -3 to 3 range. Consequently, these Z-residuals are unable to identify outliers if the criterion for outliers is based on Z-residuals outside the $(-3, 3)$ interval. By contrast, three outliers were identified with CV Z-residuals exceeding the (-3, 3) interval. This comparison indicates that CV Z-residuals exhibit superior capability in identifying outliers, despite not capturing all outliers due to their failure times not exceeding the model's limits after the introduction of jitters.

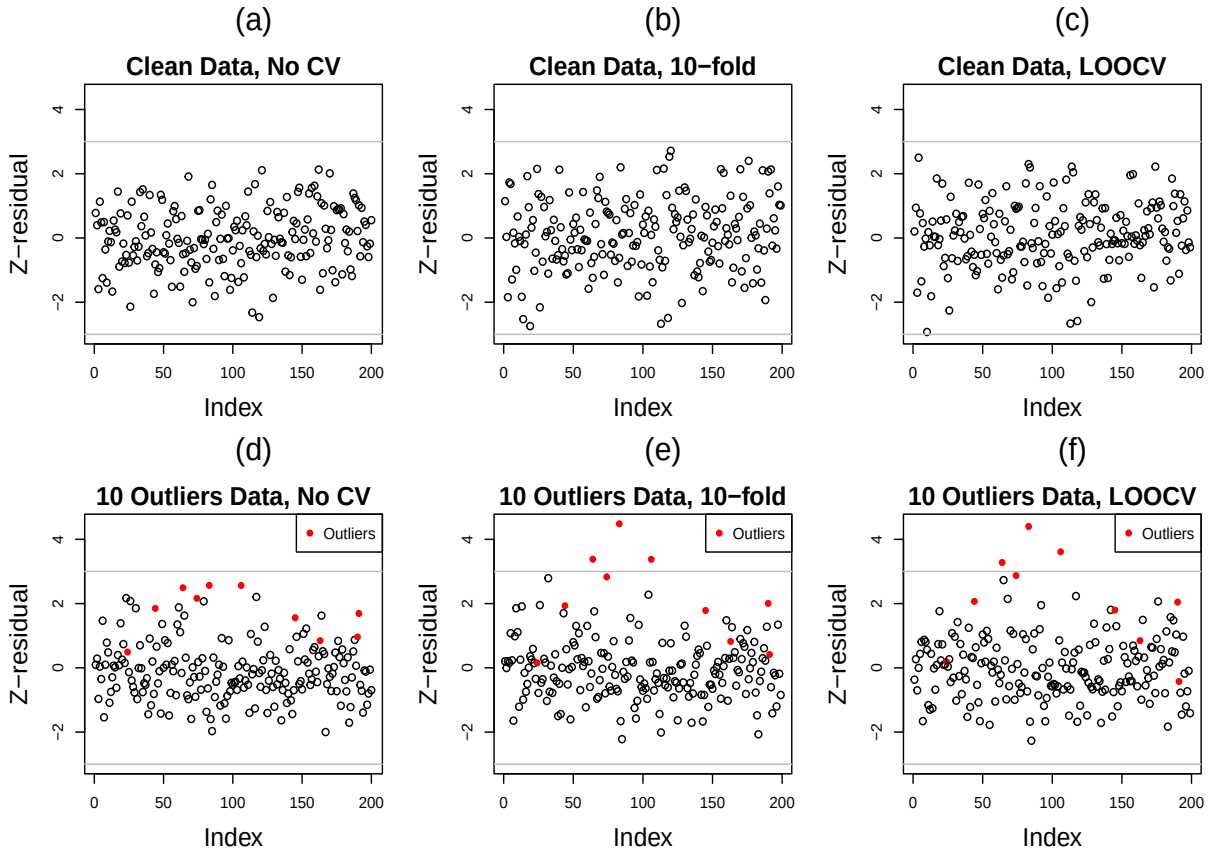

Figure 4: Comparison of the performance of Z-residuals in detecting outliers on a pair of clean and contaminated datasets. The datasets have 10 clusters each with 20 observations.

4.2.3 Comparing the Model Rejection Rates of Z-residual-based SW Tests

We conduct repeated simulations involving 1000 datasets, each comprised of 10 clusters with m observations. The cluster size m ranges from 10 to 100, allowing us to explore the relationship between Z-residual performance and cluster size variation. We aim to

assess the performance of SW tests based on three types of Z-residuals in identifying model inadequacy for contaminated datasets, for which the true model is a wrong model due to the outliers. We fit the true model for both clean and contaminated datasets. Similar to Section 4.1, we utilize 1000 simulated datasets for each simulation setting to evaluate the proportion of SW test p-values below 0.05 and to determine the mean of SW p-values.

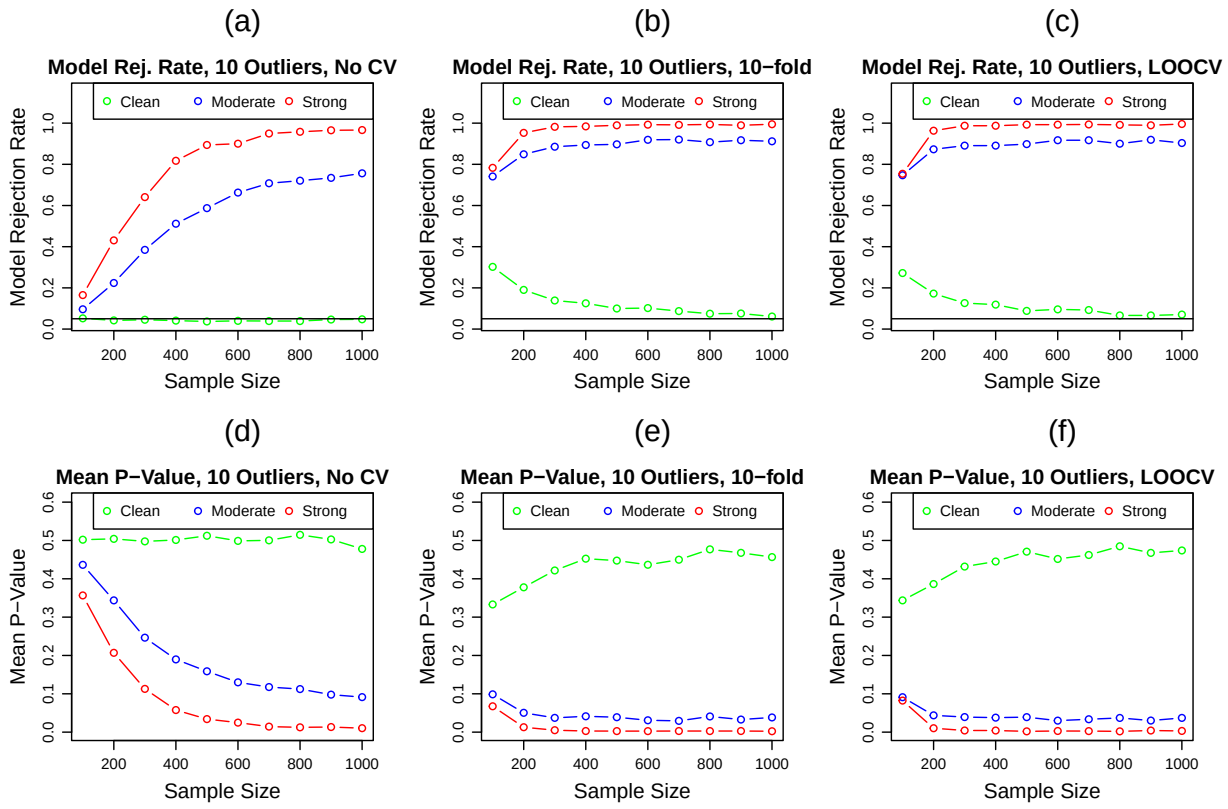

Figure 5: Comparison of model rejection rates based on SW test p-values ≤ 0.05, and the mean of SW p-values for datasets containing 10 outliers. The horizontal lines for the model rejection rate indicate the nominal type-I error rate 0.05.

Fig. 5 (a)-(c) presents the results for the scenario involving ten strong outliers. The results demonstrate that the rejection rates of No-CV Z-residuals for clean datasets (green curves) maintain the nominal level of 0.05 across all scenarios. However, for contaminated datasets, the powers of No-CV Z-residuals are notably lower than those of 10-fold and LOOCV Z-residuals. This power reduction is particularly significant when the sample size is below 300. SW tests using CV Z-residuals exhibit slightly higher type-I error rates for clean datasets when the sample size is small. Yet, these rates tend to approach 0.05 as

the sample size increases. As shown by Fig. 5 (d)-(f), the mean SW p-values of No-CV Z-residuals are significantly higher than those of CV Z-residuals.

4.2.4 Comparing the Discriminativeness of Z-residual-based SW Tests

To assess the discriminative capabilities of SW test p-values, we utilized the AUC to measure the distinction between the SW test p-values derived from clean and contaminated datasets. We create an artificial binary variable to indicate whether the dataset is clean or contaminated for calculating AUC. We consider four combinations of two jitter levels and two schemes for jitter introduction.

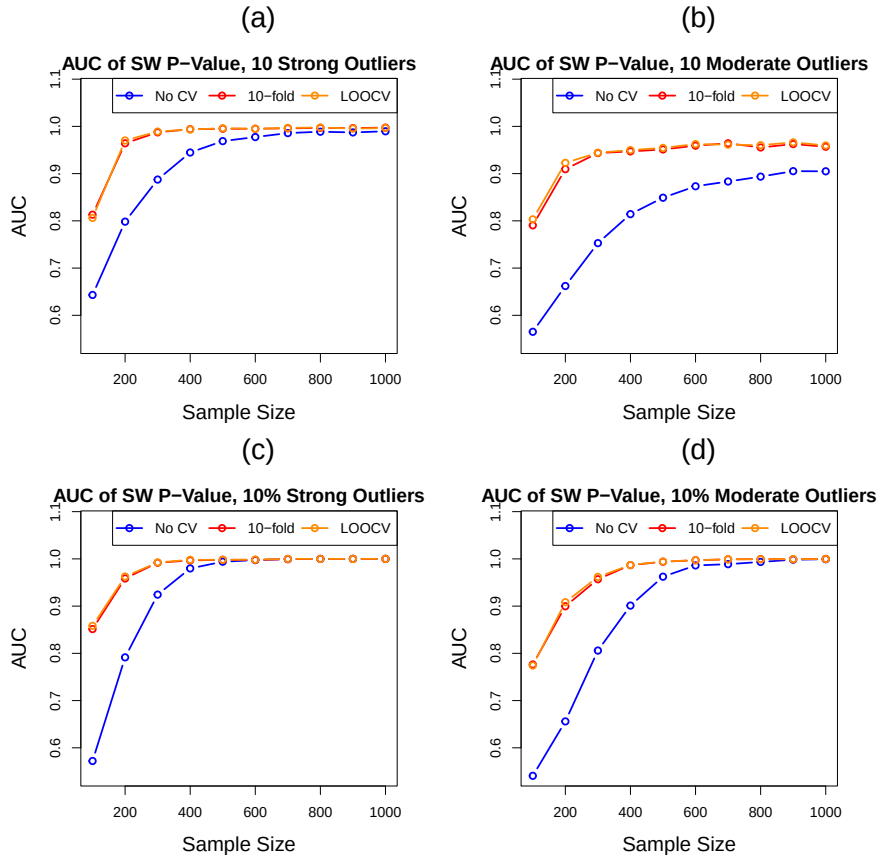

Figure 6: Comparison of the AUC values of SW test p-values based on Z-residuals for simulation datasets with outliers.

From Fig. 6, we consistently observed significantly higher AUC values for 10-fold and LOOCV Z-residuals in comparison to No-CV Z-residuals. Notably, when the sample size is around 100, the AUC values for No-CV Z-residuals tend to approach 0.5, indicating

a lack of discriminative power. In contrast, the AUC values for 10-fold and LOOCV Z-residuals are approximately 0.8 for the same sample size, demonstrating substantially greater discriminatory ability. Moreover, we observed that the difference in AUC values between CV and No-CV Z-residuals diminishes as the sample size increases in three out of the four scenarios. However, in the case where a fixed number of moderate outliers (10) is present, the discrepancy in AUC values persists even at a sample size of 1000.

4.2.5 Comparing the Outlier Detection Rates of Z-residuals

Finally, we compare the sensitivity and false positive rate (FPR) for detecting outliers using CV and No-CV Z-residuals. Our criterion for identifying an outlier is an absolute Z-residual greater than 3. Sensitivity is the proportion of true outliers correctly identified as outliers, while the FPR is the proportion of non-outliers falsely identified as outliers.

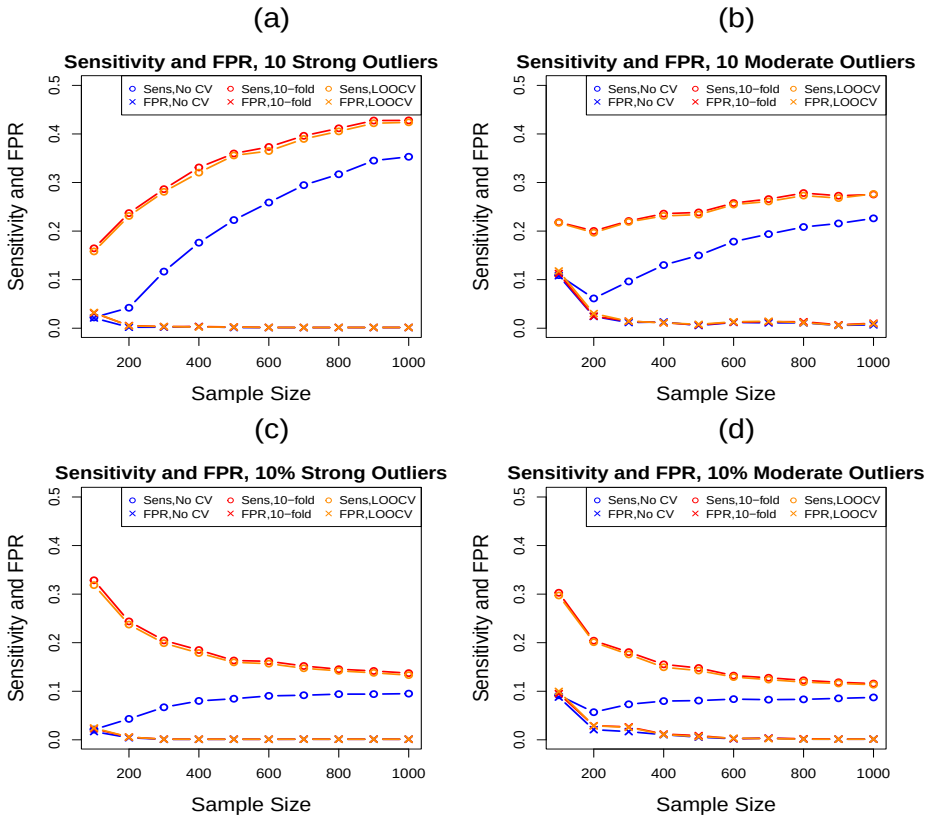

Figure 7: Comparison of the sensitivities (points with ∘) and the false positive rates (points with ×) in detecting outliers using No-CV, 10-fold, and LOOCV Z-residuals.

Fig. 7 shows the sensitivities and FPRs for the four simulation scenarios considered in Fig. 6. The CV Z-residuals display substantially higher sensitivities but nearly identical FPRs when used to detect the true outliers compared to the No-CV Z-residuals in all four scenarios. This comparison results clearly show the advantage of using CV Z-residuals for identifying outliers, despite the slight elevation of type-I error rates observed in SW tests based on these residuals. Interestingly, the sensitivity of 10-fold and LOOCV Z-residuals increases with sample size when the number of outliers is fixed at 10 but decreases and converges to a value of about 0.1 when the percentage of outliers is fixed at 10%.

5 A Real Data Example

This section demonstrates the practical application of CV Z-residuals in identifying outliers within a study on kidney infections (Mcgilchrist and Aisbett, 1991). The dataset comprises records from 38 kidney patients using a portable dialysis machine. It documents the times of the first and second recurrences of kidney infections for these patients. Each patient's survival time is defined as the duration until infection from catheter insertion. The patient records are considered as clusters due to shared frailty, signifying the common effect across patients. Instances where the catheter is removed for reasons other than infection are treated as censored observations, accounting for 24% of the dataset. The dataset encompasses 38 patient clusters, with each patient having exactly two observations, resulting in a total sample size of 76. This dataset is frequently employed to exemplify shared frailty models, and further details can be found in Mcgilchrist and Aisbett (1991).

We fit a shared gamma frailty model with three covariates: covariates: age in years, gender (male or female), and four disease types. Additional specifics are provided in Table S1 in the Supplementary Materials. The fitting process employs the `coxph` function within the `survival` package. Table 1a displays the estimated regression coefficients, along with their corresponding standard errors (SE), p-values, and 95% confidence interval (CI), derived from fitting the shared gamma frailty model using the complete dataset. The results in Table 1a reveal significant associations between the hazard of kidney infection recurrence and two covariates: sex and PKD disease type.

We computed Z-residuals and Cox-Snell (CS) residuals using the No-CV and LOOCV

Table 1: Parameter estimates of three shared gamma frailty models fitted with the kidney infection dataset. The tables (1b) and (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 kidney infection dataset.

Covariate	Estimate	SE	P-value	95% CI
Age	0.003	0.011	0.775	(-0.019, 0.025)
SexMale	1.480	0.358	0.000	(-2.185, -0.781)
DiseaseGN	0.088	0.406	0.829	(-0.709, 0.884)
DiseaseAN	0.351	0.400	0.380	(-0.433, 1.134)
DiseasePKD	-1.430	0.631	0.023	(-2.668, -0.194)
Frailty			0.933	

(b) Excluding cases 42 and 20

Covariate	Estimate	SE	P-value	95% CI
Age	0.007	0.011	0.530	(-0.015, 0.029)
SexMale	2.117	0.400	0.000	(1.333, 2.901)
DiseaseGN	0.359	0.406	0.380	(-0.436, 1.154)
DiseaseAN	0.349	0.407	0.390	(-0.448, 1.147)
DiseasePKD	-0.797	0.638	0.210	(-2.047, 0.453)
Frailty			0.940	

(c) Excluding cases 42, 20, and 15

Covariate	Estimate	SE	P-value	95% CI
Age	0.012	0.011	0.280	(-0.010, 0.034)
SexMale	2.120	0.402	0.000	(1.332, 2.906)
DiseaseGN	0.727	0.415	0.08	(-0.087, 1.540)
DiseaseAN	0.319	0.404	0.430	(-0.473, 1.112)
DiseasePKD	-0.802	0.636	0.210	(-2.049, 0.444)
Frailty			0.940	

methods for the kidney infection dataset. Given the similarity in performance between the 10-fold CV and LOOCV Z-residual methods demonstrated in the simulation studies and

the manageable computational load, we focused on the LOOCV method. Fig. 8 illustrates the residual diagnosis results for the original kidney infection dataset.

The plots in Fig. 8 (a)(b) display scatterplots against the index and QQ plots of the Z-residuals of No-CV Z-residuals. The No-CV Z-residuals predominantly fall within the interval (-3, 3)[1], displaying alignment with the 45° straight line in the QQ plot, all as expected for a random sample from the standard normal. The QQ plot of the No-CV Z-residuals indicates a SW p-value of around 0.70, signifying a well-fitted model to the dataset. Thus, the diagnostic results using No-CV Z-residuals suggest the suitability of the shared frailty model for the dataset, failing to identify any outliers or the inadequacy of the fitted model.

However, analysis of the scatterplot of LOOCV Z-residuals (Fig. 8(e)) reveals that the Z-residuals of cases labelled 20 and 42 exceed 3. These instances are considered outliers for the shared frailty model. The QQ plot of LOOCV Z-residuals displays a noticeable deviation from the 45° straight line, attributed to the considerable Z-residuals of the two identified outliers. The SW p-value of LOOCV Z-residuals is notably small, a value less than 0.01, as evident 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 two cases exhibit excessive Z-residuals, categorized as outliers for this model.

Compared to all the raw infection times portrayed in Fig. S4 in the Supplementary Materials, the infection time of case 42 ranks highest among all but doesn't particularly stand out, while the infection time of case 20 sits near the median of all infection times and doesn't appear as an outlier at all. This observation highlights the distinction between outliers concerning raw observations and those concerning a fitted model. Z-residuals are a monotone transformation of the tail (or survival) probabilities of the conditional distribution of failure time given covariates (eqn. (4)). Therefore, the identification of outliers based on Z-residuals has factored in covariate effects. However, identifying outliers based solely on raw failure times neglects covariate effects; in other words, it is grounded in a model with only the intercept term.

There exists randomness in the Z-residuals of censored observations, meaning that dif-

[1] 99.73% of observations from the standard normal should be within (-3,3)

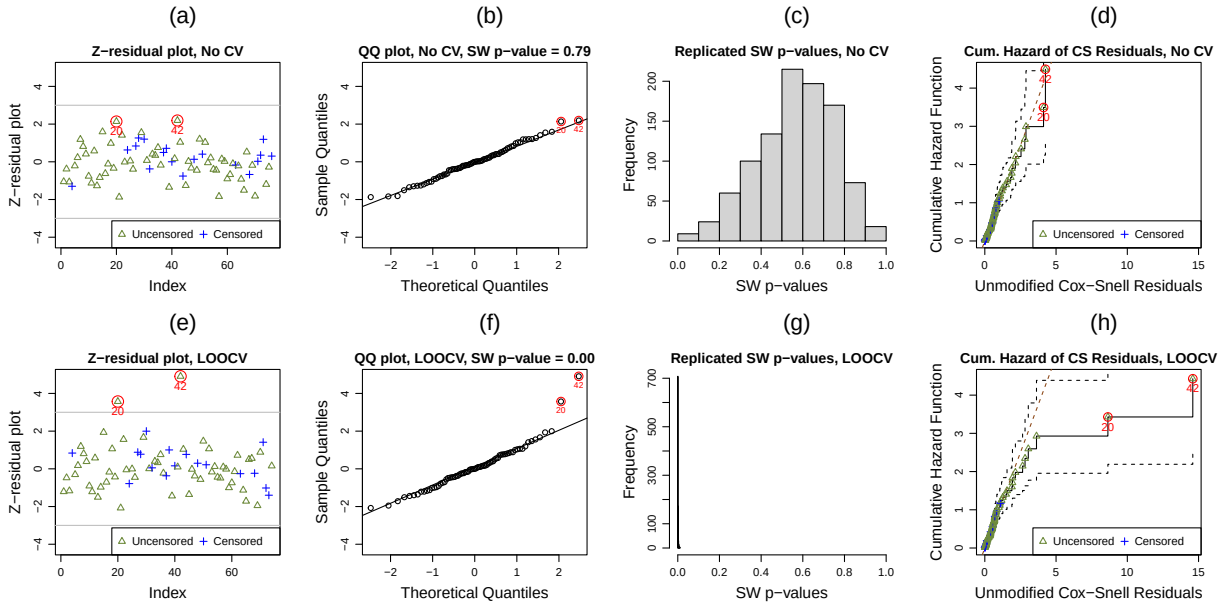

Figure 8: Scatterplots and QQ plots of No-CV and LOOCV Z-residuals of the fitted shared frailty models based on the original kidney infection dataset. The third column presents the histograms of 1000 replicated SW p-values of Z-residuals. The fourth column shows the CS residuals computed with the No-CV and LOOCV methods.

ferent sets of Z-residuals can be generated for the same dataset using distinct random numbers. Thus, to test the robustness of the previously conducted diagnosis, we replicated a large number of realizations of Z-residuals. Fig. 8 (c)(g) exhibit the histograms of 1000 SW test p-values, each derived from a set of No-CV or LOOCV Z-residuals. More than 95% of the SW p-values for No-CV Z-residuals surpass 0.05, whereas 100% of the SW p-values for LOOCV Z-residuals fall below 0.05. Note that the histogram of replicated SW p-values based on LOOCV Z-residuals concentrates highly in a tiny area near 0, perhaps all less than 0.01. This consistency across numerous replications confirms that the evaluation of the misspecification of the shared frailty model is not incidental to a specific set of LOOCV Z-residuals but a recurring conclusion supported by extensive Z-residual replications.

To further validate the above diagnoses and illustrate the effect of CV in residual diagnostics, we compute CS residuals using both No-CV and LOOCV methods. The CHFs of these residuals are depicted in Fig. 8 (d)(h). The CHF of No-CV CS residuals closely aligns with the 45° straight line, indicating a well-fitted model for the dataset. Conversely,

the CHF of the LOOCV CS residuals deviates from the 45° straight line in the upper tail, indicating inadequacy of the fitted model. The agreement between the CS residuals' model adequacy check and the Z-residual-based diagnosis is notable. However, the diagnosis with Z-residuals offers more information on the nature of the discrepancy in the inadequate model, detecting outliers and providing a quantitative measure of the statistical significance of the model's departure.

Finally, after considering the removal of the two outliers (cases 42 and 20) from the original kidney infection dataset and re-fitting the shared gamma frailty model, the results in Table 1b display that the covariate DiseasePKD is no longer statistically significant at a 5% level. The discrepancy between Table 1a and 1b emphasizes how parameter estimation and inference can be greatly influenced by including outliers, underscoring the importance of model diagnosis and outlier detection in practical data analysis.

Fig. S5 in the Supplementary Materials displays the residual diagnosis results after excluding these two outliers, indicating that the refitted model is reasonably good for the dataset without cases 42 and 20. However, it's observed that case 15 has a Z-residual marginally greater than 3. However, this may not raise substantial concern, as most of the SW p-values of LOOCV Z-residuals exceed 0.05. The model is refitted after further removing case 15, as detailed in Table 1c. The Z-residual diagnosis, depicted in Fig. S6 in the Supplementary Materials, does not indicate any inadequacy in the model fitted with the three cases removed nor identify an outlier for the model.

6 Conclusions and Discussions

Residual diagnosis plays a crucial role in validating the adequacy of fitted models (Cook and Weisberg, 1982; Collett, 2015a). However, the conventional method of using the same dataset to fit a model and diagnose its performance can potentially introduce optimistic bias. Introducing techniques such as CV can address this concern by using subsets of the data for model fitting and evaluation, thereby potentially mitigating the issues related to the double use of the entire dataset and providing a more reliable assessment of the model's performance (Marshall and Spiegelhalter, 2007). In this paper, we introduced CV methods to compute Z-residuals for shared frailty models. Through our comprehensive comparison

between the traditional No-CV Z-residuals (Li et al., 2021; Wu et al., 2024+b) and the CV Z-residuals, we demonstrated that residual diagnosis without CV tends to exhibit a conservative bias. In contrast, the CV methods notably enhanced the sensitivity and accuracy of the SW-test with Z-residuals, significantly improving the detection of model inadequacy and the identification of outliers.

Our simulations identified a potential challenge for CV Z-residuals: the slight elevation of type-I error rates in SW tests. The primary reason behind the inflated model rejection rates may be attributed to the notable difference between the training dataset (used to fit the model) and the testing dataset (used to assess model performance). In smaller sample sizes, this difference can significantly affect the model's generalization and performance assessment. This finding emphasizes the importance of cautious interpretation and thorough evaluation, as model assessment in smaller sample sizes could be less reliable and potentially prone to inflated rates of model rejection. In our opinion, to tackle this issue, improvements in frailty estimation techniques within shared frailty models are essential. One possible approach involves enhancing frailty estimation algorithms by employing stronger penalization or Bayesian methods for obtaining a better estimate of the survival function (Huang et al., 2023; Bürkner et al., 2024). An alternative path to address the potential elevation in type-I error rates involves refining the methods used for computing CV Z-residuals or for conducting SW tests. The objective here is to yield Z-residuals that are less stringent in model rejections. Introducing a strategy to marginalize the frailties (Liu et al., 2017) during the computation of randomized survival probabilities could offer a viable solution. An intriguing avenue for future research lies in a detailed comparison between methods for comparing the performance of methods for computing Z-residuals with and without frailty marginalization. This exploration could shed light on each approach's advantages and potential limitations.

Despite the challenges encountered with CV Z-residuals in scenarios involving small sample sizes, it is noteworthy that the benefits and substantial advantages of utilizing CV Z-residuals become more pronounced and valuable as the sample size increases. As the sample size grows, the CV Z-residuals display increased robustness in capturing intricate patterns associated with non-linear covariate relationships within the model, thus providing

a more nuanced and precise evaluation of how various covariates influence the outcome. Moreover, the utility of CV Z-residuals in outlier detection becomes more pronounced with larger sample sizes. The increased data volume allows for a more comprehensive assessment, enabling the identification of potential outliers with higher precision and confidence.

Availability of R code and Datasets

We have included an R function for computing CV Z-residuals for the output of the `coxph` function in the `survival` R package in the Supplementary Materials of this paper. We have also collected this R function into an R package called `Zresidual`, which can be downloaded and installed directly from GitHub via this link: https://github.com/tiw150/Zresidual. For further details on using the function and the `Zresidual` package, please refer to our demonstration available on this webpage: https://tiw150.github.io/CV_Zresidual_demo.html.

This above GitHub repository also includes an array of R code snippets for conducting simulation studies and real data analysis. Additionally, the datasets employed in these studies are provided in this repository.

Supplementary Materials

Additional figures and tables are available online via this link: http://....

Acknowledgements

The authors thank the editor, the associate editor, and the referees for helpful comments. The authors gratefully acknowledge the financial supports from the Natural Sciences and Engineering Research Council of Canada Discovery Grants (individual) program to Dr. Feng and Dr. Li.

Conflict of Interests Statement

No potential conflict of interest was reported by the authors.

References

Balan, T. A. and Putter, H. (2020), "A tutorial on frailty models," *Statistical Methods in Medical Research*, 29, 3424–3454.

Bürkner, P.-C., Gabry, J., Weber, S., Johnson, A., Modrak, M., Badr, H. S., Weber, F., Vehtari, A., Ben-Shachar, M. S., Rabel, H., Mills, S. C., Wild, S., and Popov, V. (2024), "brms: Bayesian Regression Models using 'Stan'," .

Collett, D. (2015a), *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC.

— (2015b), *Modelling survival data in medical research, third edition*, Taylor & Francis Group.

Cook, R. D. (1986), "Assessment of Local Influence," *Journal of the Royal Statistical Society. Series B (Methodological)*, 48, 133–169.

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

Cox, D. R. and Snell, E. J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society. Series B, Methodological*, 30, 248–275.

Duchateau, L. and Janssen, P. (2008), *The Frailty Model*, Statistics for Biology and Health, New York: Springer Verlag.

Gelman, A., Hwang, J., and Vehtari, A. (2014), "Understanding Predictive Information Criteria for Bayesian Models," *Statistics and Computing*, 24, 997–1016.

Gorfine, M., Zucker, D. M., and Hsu, L. (2006), "Prospective Survival Analysis with a General Semiparametric Shared Frailty Model: A Pseudo Full Likelihood Approach," *Biometrika*, 93, 735–741.

Grambsch, P. M. and Therneau, T. M. (1994), "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515–526.

Ha, I. D., Lee, Y., and Song, J. (2001), "Hierarchical likelihood approach for frailty models," *Biometrika*, 88, 233–233.

Hanagal, D. (2015), "Modeling survival data using frailty models," *Statistical methods in medical research*, 24, 936–936.

Henderson, R. (2001), "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, 2533–2534.

Hirsch, K. and Wienke, A. (2011), "Software for semiparametric shared gamma and log-normal frailty models: An overview," *Computer methods and programs in biomedicine*, 107, 582–597.

Hougaard, P. (1995), "Frailty Models for Survival Data," *Lifetime Data Anal*, 1, 255–273.

Huang, X., Xu, J., and Zhou, Y. (2023), "Efficient algorithms for survival data with multiple outcomes using the frailty model," *Statistical Methods in Medical Research*, 32, 118–132, publisher: SAGE Publications Ltd STM.

Johnson, N. L. and Kotz, S. (1977), "Distributions in statistics: Continuous univariate distributions," *Advances in Mathematics*, 26, 327–327.

Karagrigoriou, A. (2011), "Frailty Models in Survival Analysis," *Journal of Applied Statistics*, 38, 2988–2989.

Li, L., Feng, C. X., and Qiu, S. (2017), "Estimating Cross-Validatory Predictive p-Values with Integrated Importance Sampling for Disease Mapping Models," *Statistics in Medicine*, 36, 2220–2236.

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2015), "Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC," *Statistics and computing*, 26, 881–897.

Li, L., Wu, T., and Feng, C. (2021), "Model diagnostics for censored regression via randomized survival probabilities," *Statistics in medicine*, 40, 1482–1497.

Lin, D. Y. (2007a), "On the Breslow estimator," *Lifetime data analysis*, 13, 471–480.

— (2007b), "On the Breslow Estimator," *Lifetime Data Anal*, 13, 471–480.

Liu, X., Pawitan, Y., and Clements, M. S. (2017), "Generalized survival models for correlated time-to-event data," *Statistics in medicine*, 36, 4743–4762.

Marshall, E. C. and Spiegelhalter, D. J. (2003), "Approximate cross-validatory predictive checks in disease mapping models," *Statistics in medicine*, 22, 1649–1660.

— (2007), "Identifying outliers in Bayesian hierarchical models: a simulation-based approach," *Bayesian analysis*, 2, 409–444.

McCullagh, P. (Peter), . (1989), *Generalized linear models*, Monographs on statistics and applied probability (Series) ; 37, London ; New York: Chapman and Hall, 2nd ed.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall / CRC.

Mcgilchrist, C. A. and Aisbett, C. W. (1991), "Regression with Frailty in Survival Analysis," *Biometrics*, 47, 461–466.

Pierce, D. A. and Schafer, D. W. (1986), "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 977–986.

Piironen, J. and Vehtari, A. (2017), "Comparison of Bayesian Predictive Methods for Model Selection," *Statistics and Computing*, 27, 711–735.

Ripatti, S. and Palmgren, J. (2000), "Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood," *Biometrics*, 56, 1016–1022.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., code), S. S. F. D., Multiclass), M. D. H. . T., and paired test CI), Z. B. D. (2023), "pROC: Display and Analyze ROC Curves," .

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

Smith, A. L., Zheng, T., and Gelman, A. (2022), "Prediction Scoring of Data-Driven Discoveries for Reproducible Research," *Statistics and Computing*, 33, 11.

Therneau, T. M. (2000), *Modeling survival data : extending the Cox model*, Statistics for biology and health, New York: Springer.

Therneau, T. M. and Grambsch, P. M. (2013), *Modeling Survival Data: Extending the Cox Model*, Springer Science & Business Media.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990), "Martingale-Based Residuals for Survival Models," *Biometrika*, 77, 147–160.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979), "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography*, 16, 439–454.

Vehtari, A., Gelman, A., and Gabry, J. (2017), "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC," *Statistics and Computing*, 27, 1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024), "Pareto Smoothed Importance Sampling," *Journal of Machine Learning Research*, 25, 1–58.

Wu, T., Feng, C., and Li, L. (2024+a), "A Comparison of Estimation Methods for Shared Gamma Frailty Models," *Statistics in Biosciences*, to appear.

Wu, T., Li, L., and Feng, C. (2024+b), "Z-residual diagnostic tool for assessing covariate functional form in shared frailty models," *Journal of Applied Statistics*, to appear.