Research Article

Statistics in Medicine

This is the accepted version of the published paper: https://doi.org/10.1002/sim.8852

Model Diagnostics for Censored Regression via Randomized Survival Probabilities

Longhai Li[*†], **Tingxuan Wu**[†‡], **and Cindy Feng**[‡]

Residuals in normal regression are used to assess a model's goodness-of-fit (GOF) and discover directions for improving the model. However, there is a lack of residuals with a characterized reference distribution for censored regression. In this paper, we propose to diagnose censored regression with normalized randomized survival probabilities (RSP). 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. We prove that RSPs always have the uniform distribution on $(0, 1)$ under the true model with the true generating parameters. Therefore, we can transform RSPs into normally-distributed residuals with the normal quantile function. We call such residuals by normalized RSP (NRSP residuals). We conduct simulation studies to investigate the sizes and powers of statistical tests based on NRSP residuals in detecting the incorrect choice of distribution family and non-linear effect in covariates. Our simulation studies show that, although the GOF tests with NRSP residuals are not as powerful as a traditional GOF test method, a non-linear test based on NRSP residuals has significantly higher power in detecting non-linearity. We also compared these model diagnostics methods with a breast-cancer recurrent-free time dataset. The results show that the NRSP residual diagnostics successfully captures a subtle non-linear relationship in the dataset, which is not detected by the graphical diagnostics with CS residuals and existing GOF tests.

Keywords: goodness-of-fit, residual diagnostics, Cox-Snell residual, quantile residual, model checking

1. Introduction

Model diagnostics is a crucial step in model building to ensure the validity of the statistical inference. Residual analysis is a conventional tool for model checking and diagnostics. Residuals of a model are used to check the overall goodness-of-fit (GOF) of a model, discover the direction for improving the model, and identify outlier observations. Cox-Snell (CS) residuals [1] are widely used for checking survival regression models for failure times. CS residuals are transformed from the survival probabilities with the quantile function of the exponential distribution. When failure times are not censored, and the postulated model is the true model for them, the survival probabilities are uniformly distributed; hence, CS residuals are exponentially distributed. This reference distribution is the basis for model checking with CS residuals. The cumulative hazard function (CHF) of CS residuals is commonly plotted and compared to the $45°$ straight line (unity slope and zero intercept). We can also employ some goodness-of-fit (GOF) tests such as the Kolmogorov-Smirnov (KS) test [2] to check the exponentiality of CS residuals. Unfortunately, when there exist censored failure times, the survival probabilities are no longer uniformly distributed. Correspondingly, CS residuals are no longer exponentially distributed. Indeed, CS residuals are typically randomly censored observations from a distribution. We can still estimate the survival function of CS residuals with KM-like methods [3, 4] that can consider censoring and compare the CHF of censored CS residuals against the $45°$ straight line for model checking. This plot is probably the most commonly used method in practice. Although it is not conducted very often in practice, the agreement of an empirical distribution with

[†] *Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Rd, Saskatoon, SK, S7N5E6, Canada.*
[‡] *School of Public Health, University of Saskatchewan, 104 Clinic Place, Saskatoon, SK, S7N5E5 Canada.*
[*] *Correspondence to: longhai@math.usask.ca*
[§] **List of Abbreviations**: AIC, Alkaike's Information Criterion; AFT, accelerated failure time; CHF, cumulative hazard function; CS, Cox-Snell; GOF, goodness-of-fit; KM, Kaplan-Meier; KS, Kolmogorov-Smirnov; LCKS, Lilliefors-Corrected Kolmogorov-Smirnov; SW, Shapiro-Wilk; SF, Shapiro-Francia; SP, survival probability; USP, unmodified SP; RSP, randomized SP; MSP, modified SP; NRSP, normalized RSP; NMSP, normalized MSP; NUSP, normalized USP.

a reference distribution can also be quantified by some GOF test methods that are extended to handle censored data [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16].

The GOF checking that assesses the agreement of the distribution of residuals with a reference as described above is just the first-line model diagnostics, like checking patients' blood pressure in medical diagnostics. The departure of the distribution of residuals of a flawed model may be too subtle to be detected by GOF checking. More importantly, when a model failure is detected, the GOF test results typically reveal little information about the nature of the model failure, such as non-linear covariate effect, non-constant variances, and lack of independence. For identifying these model discrepancies, we need to conduct more in-depth graphical and quantitative diagnostics. Therefore, the GOF checking is insufficient for practical model diagnostics. We also desire to define residuals that have a known reference distribution under the true model so that we can freely conduct a wide variety of model diagnostics. Some non-random methods have been proposed to modify CS residuals [17], which include the ways of adding CS residuals for censored times with a constant, the martingale residuals [18], the deviance residuals, and possibly others [19]. However, these modified residuals under the true model do not have a unified and characterizable distribution.

This paper proposes using normalized randomized survival probabilities (RSPs) to conduct model diagnostics for censored regression. 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. We prove that RSPs always have the uniform distribution on $(0, 1)$ under the true model. Therefore, we can transform RSPs into normally-distributed residuals with the normal quantile function. We call such residuals by **normalized RSP (NRSP) residuals**. We conduct simulation studies to investigate the sizes and powers of statistical tests based on NRSP residuals in detecting the incorrect choice of distribution family and non-linear effect in covariates. Our simulation studies show that the sizes of model diagnostic tests based on NRSP residuals are very close to nominal. Furthermore, our results show that, although the GOF tests with NRSP residuals are not as powerful as a traditional GOF test method, a non-linearity test based on NRSP residuals has significantly higher powers in detecting the non-linear covariate effects. We also compared these model diagnostic methods with a breast-cancer recurrent-free time dataset. The results show that the model diagnostics with the NRSP residuals successfully captures a non-linear covariate effect in the dataset, which is not detected by the graphical diagnostics with CS residuals and existing GOF tests.

The rest of this paper is organized as follows. Section 2 reviews the existing residuals and diagnostic methods for censored regression. In Section 3, we present the definition of randomized survival probabilities (RSPs), with an illustrative example and proof of the uniformity of RSPs under the true model. In Section 4, we conduct simulation studies to investigate the performances of NRSP residuals. Section 5 presents the results of applying the NRSP residual to a breast cancer recurrence-free time dataset. The article is concluded in Section 6.

2. Review of Existing Model Diagnostics Methods for Censored Regression

In this section, we review some existing model diagnostic methods used in survival analysis. A central concept in these residuals is the survival probability (SP). Suppose T_i^* is the true failure time of the ith individual, which we assume to be a continuous random variable in this article. Let t_i^* denote the realization of T_i^*. In many practical problems, we may not be able to observe t_i^* exactly, but we can observe that T_i^* is greater than a value C_i, which is called right-censoring. The observed failure times are denoted by the pair (T_i, d_i), where $T_i = \min(T_i^*, C_i), d_i = I(T_i^* < C_i)$. Since we will consider only the right-censoring in this article, we will use the "censoring" as a short for the "right-censoring".

Suppose the survival function of T_i^* based on a postulated model is defined as $S_i(t_i^*) = P(T_i^* > t_i^*)$, where the subscript i indicates that the probability depends covariate x_i for the ith individual. Using a simple probability argument, one can prove that the survival probabilities $S_i(T_i^*)$ are uniformly distributed when $S_i(\cdot)$ is the survival function of the true model for T_i^*. SPs can be transformed into random variables with a desired distribution by applying its inverse CDF or survival function. The widely used Cox-Snell (CS) residual is defined as $r_i^c(T_i^*) = -\log(S_i(T_i^*))$, where $-\log(\cdot)$ is the inverse survival function of $\exp(1)$. Therefore, CS residuals are exponentially distributed under the true model. Although it is not used often in practice, one can also define normally-distributed residuals [20]: $r_i^n(T_i^*) = \Phi^{-1}(S_i(T_i^*))$, which we call by **normalized SPs**. Then we can apply a variety of residual diagnostic methods for normal regression to diagnose $S_i(\cdot)$.

If T_i is censored, the **unmodified survival probability (USP)**, $S_i(T_i)$, is larger than $S_i(T_i^*)$ since $T_i < T_i^*$. Thus, when there are censored observations, the distribution of $S_i(T_i)$ is no longer uniformly distributed under the true model. The non-uniformity in USPs causes the difficulty of performing residual diagnostics. The **unmodified CS residuals**, $r_i^c(T_i) = -\log(S_i(T_i))$, and **normalized unmodified SPs (NUSP)**, $r_i^n(T_i)$, can be treated as univariate data with censoring if we ignore x_i. In practice, the most widely used diagnostic tool is to apply KM methods [3] to $\{(r_i^c(T_i), d_i)|i = 1, \ldots, n\}$ to get an estimate of the CHF of CS residuals. Under the true model, the CHF of CS residuals is expected to be close to the 45° straight line. In addition to the graphical checking, we also desire a quantitative measure of the GOF. This

problem becomes challenging due to censoring. However, some methods have been developed for checking the GOF of univariate data with censoring. Shapiro-Wilk (SW) and Shapiro-Francia (SF) normality tests [21, 22] have been extended to singly censored data [5, 6, 7]. Although censoring times C_i for T_i^* may not be random, unmodified SPs and their transformations are typically randomly censored due to *the randomness in covariates*. The chi-squared test and some normality tests have been extended to randomly censored data; see [8, 9, 10, 11, 12, 13, 14] among others. The function `gofTestCensored` in R package `EnvStats` [15, 16] provides an SF test for multiply censored data. The method used in `gofTestCensored` is a generalization of the method for extending SW and SF tests for singly censored data. The key idea of these extensions is to measure the product-moment correlation between the uncensored observations and the corresponding standard normal quantiles with the linking probabilities (called plotting positions) estimated with KM-like methods [3, 4]. The details are given in the manual page of the function `gofTestCensored`, which also contains a detailed discussion of SW and SF tests.

The graphical and quantitative methods for comparing the distribution of residuals to a reference distribution are useful in detecting the lack of model fit. Many model mis-specifications may be captured by these methods. However, examining the distribution of residuals alone is only the first-line model diagnostics. An analogy in medical diagnostics is that we check the blood pressure or temperature of patients for detecting diseases, which is not sufficiently powerful for the identification of potential diseases. For example, these methods ignore the covariates. Thus, they may fail to detect the model mis-specification in linking T_i^* with x_i. More importantly, the GOF test results typically cannot reveal the nature of model mis-specification, especially that related to x_i, such as non-linearity, lack of independence, non-constant variances, and many others. For conducting such in-depth diagnostics, we desire to define residuals that have a known reference distribution under the true model. A number of methods have been proposed to modify USPs or their transformations. A commonly used method is to shrink the USPs of the censored failure times:

$$S_i'(T_i, d_i, \eta) = \begin{cases} S_i(T_i), & \text{if } T_i \text{ is uncensored, i.e., } d_i = 1, \\ \eta\, S_i(T_i), & \text{if } T_i \text{ is censored, i.e., } d_i = 0, \end{cases} \tag{1}$$

where $\eta \in (0, 1)$. We call the shrunken SPs by **modified SPs (MSPs)**. Transforming the MSPs with the inverse survival of $\exp(1)$ results in the modified CS residuals with a constant $\Delta = -\log(\eta)$ added to the CS residuals of censored failure times, given by $r_i^{c\prime}(T_i, d_i, \Delta) = -\log(S_i'(T_i, d_i, \eta))) = r_i^c(T_i) + \Delta(1 - d_i)$. We can similarly define **normalized MSPs (NMSP)** [20]: $r_i^{\text{NMSP}}(T_i, d_i, \eta) = \Phi^{-1}(S_i'(T_i, d_i, \eta))$. There are many different choices for the shrinkage factor η or Δ in the literature based on different choices of conditional means of SPs or their transformations given $T_i^* > T_i$; $\Delta = 1$ and $\Delta = \log(2)$ are often chosen; see [17, 19, 20]. Other residuals, for example the martingale residuals $r_i^M(T_i, d_i) = d_i - r_i^c(T_i)$ and the deviance residuals, and residual-based diagnostic tools have also been proposed for diagnosing censored regression; see [17, 18, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31] and the references therein. Although many residuals by modifying the USPs or their transformations exist, a common drawback for these modified residuals is that their distributions under the true model are very complicated due to censoring, thus, they cannot be characterized clearly with a known distribution or probability table. This distribution depends on the distribution of censoring times C_i, which varies for different datasets. Therefore, there is a lack of reference distributions for us to conduct model diagnostics with these residuals.

3. Normalized Randomized Survival Probabilities

3.1. Definition of Randomized Survival Probabilities

We will propose to diagnose censored regression with normalized randomized survival probabilities (RSPs). The randomized survival probability (**RSP**) for T_i is defined as:

$$S_i^R(T_i, d_i, U_i) = \begin{cases} S_i(T_i), & \text{if } T_i \text{ is uncensored, i.e., } d_i = 1, \\ U_i\, S_i(T_i), & \text{if } T_i \text{ is censored, i.e., } d_i = 0, \end{cases} \tag{2}$$

where U_i is a uniform random number on $(0, 1)$, and $S_i(\cdot)$ is the postulated survival function for T_i^* given x_i. From the definition, we see that the fixed shrinkage factor η in MSPs is replaced by a random number $U_i \in (0, 1)$. In other words, $S_i^R(T_i, d_i, U_i)$ is a random number between 0 and $S_i(T_i)$ when T_i is censored. We will show that the randomized SP is uniformly distributed on $(0, 1)$ given x_i under the true model. Therefore, we can transform them into residuals with any desired distribution. We prefer to transforming them with the normal quantile:

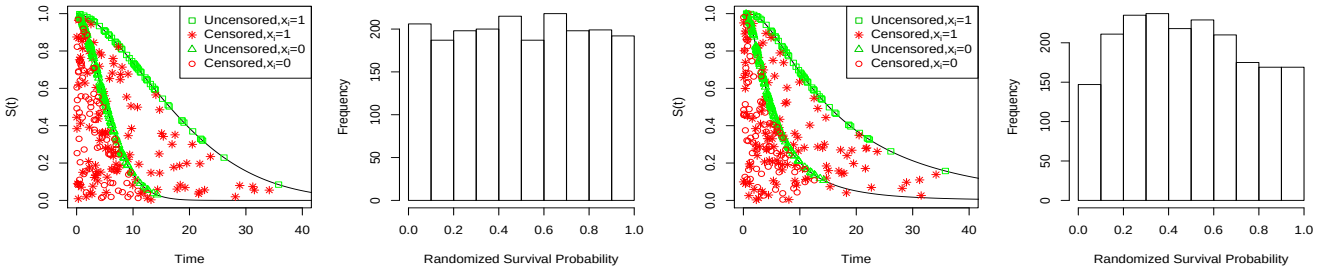
$$r_i^{\text{NRSP}}(T_i, d_i, U_i) = \Phi^{-1}(S_i^R(T_i, d_i, U_i)). \tag{3}$$

We name the residuals in (3) as normalized randomized SP (**NRSP**) residuals. Due to the normality of NRSP residuals under the true model, we can conduct model diagnostics with NRSP residuals for censored data in the same way as

conducting model diagnostics for a normal regression. There are a few advantages of transforming RSPs into NRSPs. First, the diagnostic methods for checking normal 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 plotting RSPs. However, the presence of such extreme SPs, even very few, is indicative of model mis-specification. Normal transformation can highlight such extreme RSPs.

3.2. *Illustration of the Uniformity of RSP*

We generate 2000 failure times, T_i^*, as follows: $\log(T_i^*) = 2 + x_i + \epsilon_i$, where ϵ_i is generated from the extreme-value distribution with a shape parameter γ set as 1.8, and x_i is generated as a Bernoulli (0.5). This model is called Weibull accelerated failure time (Weibull AFT) model [17, 32, 33]. Figure 1a depicts 400 RSPs along with the two fitted survival curves when Weibull model is fitted to the dataset. For the uncensored times, the survival probabilities are calculated with the survival functions and for each censored time T_i, the survival probability evaluated at observed T_i is replaced by a random number between 0 and $S_i(T_i)$. The histogram in Fig. 1b shows clearly that the RSPs are uniformly distributed on $(0, 1)$. We also fitted log-normal model as a wrong model for this dataset with results shown by Figure 1c and 1d. We see that due to the mis-specified distribution family, the RSPs are not uniformly distributed.



(a) RSPs, True Model **(b) Histogram of RSPs, True Model** **(c) RSPs, Wrong Model** **(d) Histogram of RSPs, Wrong Model**

Figure 1. Illustration of the uniformity of RSPs. An animated display of this figure with multiple simulated datasets is shown in the URL given in Section B.

3.3. *Proof of the Uniformity of RSP*

We first rewrite $S_i^R(T_i, d_i, U_i)$ as a function of (T_i^*, C_i, U_i) as follows:

$$S_i^{R'}(T_i^*, C_i, U_i) = \begin{cases} S_i(T_i^*), & \text{if } T_i^* \le C_i \\ U_i\, S_i(C_i), & \text{if } T_i^* > C_i. \end{cases} \tag{4}$$

We will show that the conditional distribution of $S_i^{R'}(T_i^*, C_i, U_i)$ given $C_i = c$ is uniform on $(0, 1)$. *To proceed, we assume that T_i^* and C_i are independent, that is, censoring times are non-informative to the original failure times.* Based on this assumption, the distribution of $S_i(T_i^*)$ is untacted given $C_i = c$. Hence, given $T_i^* \le c$, the RSP is equal to $S_i(T_i^*)$, and it is uniformly distributed on $(S_i(c), 1)$. And, given $T_i^* > c$, the RSP is equal to $U_i\, S_i(c)$, which is uniformly distributed on $(0, S_i(c))$ due to the uniformity of U_i. In addition, the probability of $T_i^* > c$ given $C_i = c$ is $S_i(c)$. With $\lambda(B)$ denoting the length of an interval B on $(0, 1)$, we can derive the following equations:

$$\begin{aligned} & P(S_i^{R'}(T_i^*, C_i, U_i) \in B \mid C_i = c) && (5) \\ = \;& P(S_i(T_i^*) \in B \mid C_i = c, T_i^* \le c) \times P(T_i^* \le c) + P(U_i S_i(c) \in B \mid C_i = c, T_i^* > c) \times P(T_i^* > c) && (6) \\ = \;& \lambda(B \cap (S_i(c), 1)) + \lambda(B \cap (0, S_i(c))) = \lambda(B) && (7) \end{aligned}$$

Since the conditional distribution of $S_i^{R'}(T_i^*, C_i, U_i)$ given $C_i = c$ is uniform on $(0, 1)$, the marginal distribution of $S_i^{R'}(T_i^*, C_i, U_i)$ is uniform on $(0, 1)$ too after the C_i is marginalized away by applying the total probability rule again. The proof that the randomized SPs are uniformly distributed on $(0, 1)$ is completed.

It is worth pointing out that we have proved that the RSP given x_i is uniformly distributed on $(0, 1)$. Therefore, the marginal distribution of RSPs is also uniformly distributed on $(0, 1)$. The marginal uniformity is used in GOF tests, and the conditional uniformity of RSPs can be used to check model assumptions in linking T_i^* with x_i.

3.4. *Model Diagnostics Based on NRSP Residuals*

NRSP residuals can be used in testing a model's GOF, for which we recommend SW or SF normality tests. In the proof given in Section 3.3, we assume that the postulated model $S_i(\cdot)$ is the true model for T_i^*. In practice, the $S_i(\cdot)$ needs to

be estimated with sample data. When the same dataset, $\{(T_i, d_i)|i = 1, \ldots, n\}$, is used to estimate the model parameters and used to calculate residuals for model checking, there might be a conservatism (bias) problem due to the double use of the dataset. NRSP residuals may be more concentrated around 0 than exactly distributed as $N(0, 1)$. However, this conservatism is very small when the sample size n is much larger than the number of parameters. For GOF tests, our simulation results show that the SW and SF normality tests applied to NRSP residuals are more resistant to the conservatism than the KS test; see a dedicated investigation in Section A.1 with simulation studies. However, when a very complex model (for example, with many covariates) is fitted to a small number of failure times, it may be necessary to apply cross-validation methods to compute NRSP residuals. For example, in leave-one-out cross-validation, an observation is held out; then, the model parameters are estimated with the remaining observations; finally, the estimated parameters are used to calculate the residual for the held-out observation.

When a model is correctly specified, the conditional distribution of NRSPs given x_i is approximately standard normal and is homogeneous for varying x_i and the linear predictors (fitted values). Therefore, most model diagnostic tools for normal regression can be applied to NRSPs for diagnosing censored regression. In particular, a scatterplot of NRSP residuals versus each x_i can be used to check whether the linear assumption with x_i is appropriate or not. For qualitatively testing the non-linearity, we apply the F-test in ANOVA for testing the equality of means of NRSPs in the k groups that are formed by cutting fitted values with equally-spaced intervals.

3.5. A P-value Upper Bound for Assessing Replicated NRSP GOF Test p-values

A difficulty in conducting statistical tests with NRSP residuals is the randomness in the test p-values. Given a fitted model, we can generate multiple sets of NRSP residuals and obtain replicated test p-values. We can choose to randomly report one of them and draw a histogram of the replicated test p-values to assess the model fit. However, the randomness may still be undesirable. In this section, we describe a theoretically non-random p-value upper bound based on a formula about the distribution of order statistics of correlated random variables. Suppose $d_1, \ldots, d_J$ are possibly correlated random variables with a common survival function $S(\cdot)$. Let $d_{(r)}$ be the rth order statistics. It is shown [34, 35] that:

$$P(d_{(r)} > t) \leq \min\left(1, S(t)\frac{J}{J - r + 1}\right), \text{for } r = 1, \ldots, J. \tag{8}$$

This formula has been used to give an upper bound for Bayesian model checking with pivotal discrepancy measures calculated with posterior samples [36, 37], which are correlated random variables. Here we apply this upper bound to a different scenario. Suppose $p_1, \ldots, p_J$ are replicated NRSP test p-values for a fitted model with the same dataset. We have shown that each p_j is uniformly distributed on $(0, 1)$ under the true model. However, $p_1, \ldots, p_J$ are correlated because they use the same dataset. Applying the formula (8) to $d_i = -p_i$, we obtain the following inequality for the rth order statistics $p_{(r)}$:

$$P(p_{(r)} < t) \leq \min\left(1, t\frac{J}{r}\right). \tag{9}$$

Based on (9), a p-value upper bound for observed (simulated) rth statistics $p^{\text{obs}}_{(r)}$ is given by $\min\left(1, p^{\text{obs}}_{(r)}\frac{J}{r}\right)$. To avoid the selection of r, we report the minimal upper bound for $r = 1, \ldots, J$, denoted by p_{min}:

$$p_{\text{min}} = \min_{r=1,\ldots,J} \min\left(1, p^{\text{obs}}_{(r)}\frac{J}{r}\right). \tag{10}$$

The p_{min} is rather conservative for assessing model GOF 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 [37]) than the conventional 0.05 for exact p-values.

4. Simulation Studies

4.1. Detection of Mis-specified Distribution Family

In this simulation setting, we generated original failure times from a Weibull AFT model with $\log(T^*_i) = 2 + x_i + \epsilon_i$, with ϵ_i generated from the extreme-value distribution with a shape parameter of 2, and censoring times C_i were generated from $\exp(\theta)$. We set four different values of θ to obtain four different censoring rates (c): $0\%, 20\%, 50\%$, and 80%. We generated datasets with varying sample size n ranging from 100 to 800. We then examined the performances of various

residuals after we fitted the true model (Weibull AFT) and a wrong model (log-normal AFT) with the same linear link function, $\log(T_i^*) = \beta_0 + \beta_1 x_i + \epsilon_i$, to these datasets.

We first look at the performances of NRSP residuals on a single dataset with the sample size $n = 800$ and $\theta = 0.08$, which induce a censoring rate $c \approx 50\%$. Fig. 2 displays the NRSP residuals against the index and their normal QQ plots under the true and the wrong models. Under the true model, the NRSP residuals are randomly scattered without exhibiting any pattern. They are mostly within the interval (-3, 3). The QQ plot of the NRSP residuals aligns nearly perfectly with the 45° straight line. Under the wrong model, the NRSP residuals are skewed to the right, and the QQ plot also indicates the deviation of NRSP residuals from the normal. Note that large NRSP residuals correspond to small failure times because of the descent of survival function. The scatterplot and QQ plot of NRSP residuals under the wrong log-normal model indicate that the true model for the dataset has more probability on the left than the fitted model (log-normal). The corresponding residual and QQ plots of the NMSP and deviance residuals are given in Figure S3. We see that the distributions of the NMSP and deviance residuals under the true model are far from the standard normal. From looking at the CHF of the wrong model (log-normal), we can draw the same conclusion regarding the problem of the model. However, we notice that the estimated cumulative hazard curve under the true model is not very straight, although it seems to be within the confidence bands. An animated display of Fig. 2 with multiple simulated datasets is shown in the URL given in Section B.

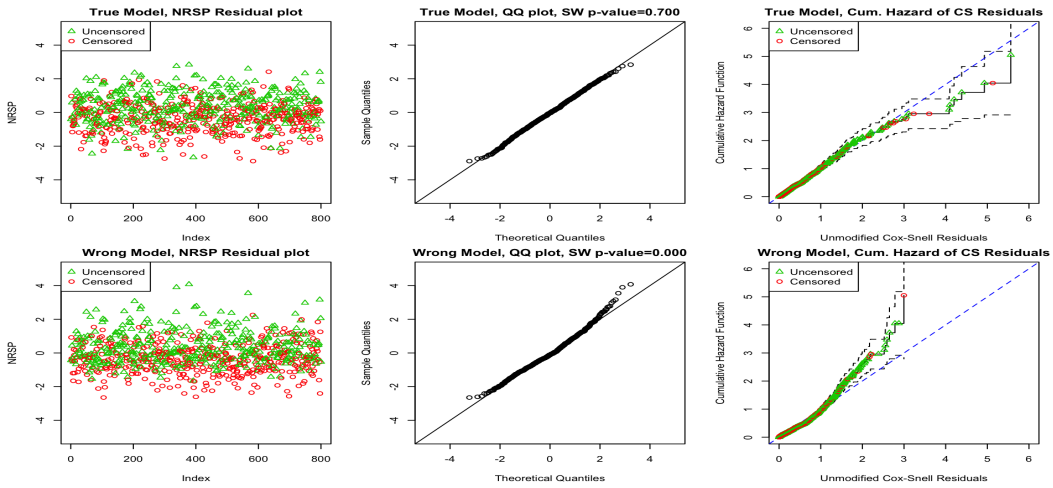


Figure 2. Performance of using the NRSP residuals as a graphical tool for detecting mis-specification of distribution family. The dataset has a sample size n = 800 and a censoring rate $c \approx 50\%$. The true model is a Weibull AFT, and the wrong model is a log-normal AFT. Note that large NRSP residuals correspond to small failure times. An animated display of this figure with multiple simulated datasets is shown in the URL given in Section B.

For demonstrating the performances of NRSP residuals in discriminating good and bad models with replicated simulated datasets, we evaluated the performances of a set of GOF tests applied to NRSP and other residuals with simulated datasets. The GOF test methods are given in the 2nd row of Table 1. The names of GOF test methods are denoted by "R-T" with "R" denoting residual name and "T" denoting test method. In particular, NUSP-CSF is the method that an extension of SF normality test is applied to NUSP residuals, which is implemented with `gofTestCensored` in R package `EnvStats` [15, 16]. We generated 2000 datasets with the Weibull AFT model for each combination of a sample size n and a censoring rate c, controlled by the mean of exponential censoring times. Using 1000 datasets generated under each scenario, we estimated the probabilities of model rejections when cutting GOF test p-values with 0.05. Table 1 displays the percentages of model rejections of each GOF testing method under the true model and the wrong log-normal model. We see that the false-positive rates of NRSP-SW and NRSP-SF are close to the nominal level 0.05 for all scenarios (well-calibrated) and have good powers (true-positive rates) in detecting the incorrect choice of distribution family. NUSP-CSF has higher true-positive rates and lower false-positive rates than NRSP-SW and NRSP-SF. That is, NUSP-CSF is more discriminative than NRSP-SW and NRSP-SF. However, NRSP residuals plots can show more details about how a model misfits a dataset; for example, which observations are not accommodated by the model (i.e., outlier or divergent observations).

Table 1 also shows that the performances of other GOF tests are not satisfactory. Because NMSP and deviance residuals are not normally distributed when $c > 0$, NMSP-SW and Dev-SW have very high false-positive rates when $c > 0$. The true-positive rates of the NMSP-SW and Dev-SW methods are very high. However, the high true-positive rates of NMSP-SW and Dev-SW do not imply that they are discriminative because they have very high false-positive rates (nearly 100% when c is large). NRSP-KS has low false-positive and true positive rates because the KS test is more affected by the double use of data in estimating parameters and calculating residuals than SW and SF tests; see a more dedicated study in appended Section A.1. In appended Fig. S4, we provide the histograms of test p-values of NRSP-SW, NMSP-SW and Dev-SW when $n = 800$ and $c \approx 50\%$ for showing more details of their performances.

Table 1. Comparison of the percentages of model rejections of various GOF tests. A model is rejected when the test p-value is smaller than 0.05. Note that we use a random NRSP test p-value rather than the $p_{\min}$. The response variable is simulated from a Weibull AFT model with varying sample size and censoring rate. The wrong model is a log-normal AFT model with the same linear link function as the true model.

		Under the true model						Under the wrong model					
n	$100c$	NRSP-SW	NRSP-SF	NUSP-CSF	NRSP-KS	NMSP-SW	Dev-SW	NRSP-SW	NRSP-SF	NUSP-CSF	NRSP-KS	NMSP-SW	Dev-SW
100	0	4.40	4.95	4.95	0.05	4.40	4.20	94.10	93.55	93.55	10.05	94.10	93.25
200	0	3.35	3.70	3.70	0.00	3.35	3.10	99.85	99.75	99.75	42.05	99.85	99.85
400	0	4.40	4.55	4.55	0.00	4.40	4.60	100.00	100.00	100.00	90.30	100.00	100.00
800	0	5.10	5.15	5.15	0.00	5.10	5.60	100.00	100.00	100.00	99.95	100.00	100.00
100	20	4.70	4.45	4.25	0.10	32.93	12.06	77.88	78.58	85.19	4.90	98.55	84.83
200	20	4.75	5.25	4.60	0.20	60.43	21.71	97.50	97.40	99.10	20.01	99.95	99.15
400	20	4.80	4.25	3.90	0.00	89.65	38.55	100.00	100.00	100.00	61.75	100.00	100.00
800	20	4.45	4.45	4.65	0.05	100.00	71.45	100.00	100.00	100.00	96.90	100.00	100.00
100	50	4.13	4.58	2.82	0.35	99.90	89.28	44.34	49.57	60.34	2.37	100.00	98.74
200	50	4.37	4.37	2.26	0.75	100.00	99.55	75.25	78.16	90.16	7.93	100.00	100.00
400	50	4.94	4.94	2.47	0.55	100.00	100.00	96.92	97.43	99.65	19.93	100.00	100.00
800	50	4.87	4.67	2.16	0.40	100.00	100.00	100.00	100.00	100.00	55.49	100.00	100.00
100	80	4.41	4.31	1.85	2.35	100.00	100.00	11.62	15.37	22.33	2.15	100.00	100.00
200	80	4.31	4.81	2.06	1.95	100.00	100.00	23.01	29.22	45.96	3.46	100.00	100.00
400	80	4.71	4.31	1.10	2.20	100.00	100.00	46.97	52.33	75.71	3.46	100.00	100.00
800	80	5.27	5.17	0.80	2.26	100.00	100.00	77.17	80.98	96.54	6.12	100.00	100.00

4.2. Detection of Non-linear Covariate Effect

In this section, we demonstrate the effectiveness of the NRSP residuals in detecting non-linear covariate effects. The response variable is simulated from a Weibull AFT regression model with a non-linear link function: $\log(T_i^*) = 2 + 5\sin(2x_i) + \epsilon_i$. The covariate x_i was generated uniformly on $(0, 3\pi/2)$. The shape parameter of the Weibull distribution was set as 1.8. The censoring times C_i were generated from $\exp(\theta)$ with θ varied for obtaining different censoring rates. We considered fitting a Weibull AFT model assuming $\log(T_i^*) = \beta_0 + \beta_1 x_i + \epsilon_i$ as a wrong model, and fitting a Weibull AFT model assuming $\log(T_i^*) = \beta_0 + \beta_1 \sin(2x_i) + \epsilon_i$ as the true model.

We first look at the performances of NRSP residuals on a single dataset with the sample size $n = 800$ and $c \approx 50\%$. Figure 3 displays the NRSP residuals against the covariate x_i and their normal QQ plots. Under the true model, the residuals are mostly bounded between -3 and 3 as standard normal deviates without a visible pattern; the QQ plot aligns well with the 45° straight line. Under the wrong model, a non-linear pattern in the NRSP residual scatterplot is obvious, and the QQ plot deviates from the 45° straight line. The CHF of CS residuals under the wrong model aligns well with the 45° straight line. Therefore, for this example, the visual inspection of the CHF of CS residuals fails to detect the non-linearity in the dataset. The scatterplots and QQ plots of the NMSP and deviance residuals are given in Figure S5. There are non-linear patterns in their scatterplots under the true and wrong models; hence, we cannot use the non-linear patterns to distinguish good and bad models.

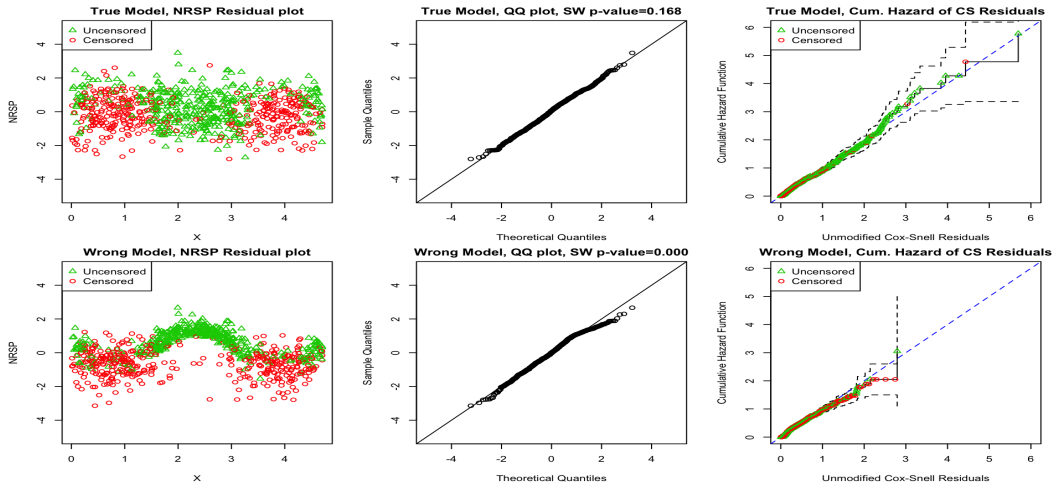


Figure 3. Performance of the NRSP residuals as a graphical tool for detecting non-linear effect in covariate. The dataset has a sample size n = 800 and a censoring rate $c \approx 50\%$. The true model is a Weibull AFT model $\log(T_i^*) = \beta_0 + \beta_1 \sin(2x_i) + \epsilon_i$ and the wrong model is a Weibull AFT model $\log(T_i^*) = \beta_0 + \beta_1 x_i + \epsilon_i$. Note that large NRSP residuals correspond to small failure times. An animated display of this figure with multiple simulated datasets is shown in the URL given in Section B.

We used simulated datasets to evaluate the performances of a set of statistical tests applied to NRSP and other residuals. GOF tests only compare the distribution of residuals as univariate data with a hypothetical distribution; hence, they may not detect implausible assumptions in linking T_i^* and x_i. Therefore, we also investigated a non-linearity test with NRSP residuals, denoted by **NRSP-AOV**. In this test, we first divide NRSP residuals into $k = 10$ groups by cutting the fitted values into equally-spaced intervals; then, we apply the F-test in ANOVA to test whether the means of NRSP residuals are equal in the k groups. We generated 2000 datasets for each combination of a sample size n and a censoring rate c from the true model for estimating the percentages of model rejections (i.e., test p-values < 0.05) of each test method. The results are shown in Table 2. NRSP-SW and NRSP-SF methods have false-positive rates close to the nominal level 0.05 for all scenarios and good power in detecting non-linearity. The performances of NMSP-SW, Dev-SW, and NRSP-KS are not satisfactory, as described in Section 4.1. We see that NUSP-CSF is more discriminative than NRSP-SW and NRSP-SF. However, the NRSP residuals enable us to conduct non-linearity diagnostics in addition to the GOF checking. Table 2 shows that the NRSP-AOV can detect the non-linearity with very high powers (nearly 100%), significantly higher than those of the GOF tests, including NUSP-CSF.

Table 2. Comparison of the percentages of model rejections of various statistical tests. A model is rejected when the test p-value is smaller than 0.05. Note that we use a random NRSP test p-value rather than the $p_{\min}$. The response variable is simulated from a Weibull AFT model $\log(T_i^*) = 2 + 5\sin(2x_i) + \epsilon_i$ with varying sample size and censoring rate. We consider fitting a Weibull AFT regression model $\log(T_i^*) = \beta_0 + \beta_1 x_i + \epsilon_i$ as a wrong model.

		Under the true model							Under the wrong model						
n	$100c$	NRSP-SW	NRSP-SF	NUSP-CSF	NRSP-AOV	NRSP-KS	NMSP-SW	Dev-SW	NRSP-SW	NRSP-SF	NUSP-CSF	**NRSP-AOV**	NRSP-KS	NMSP-SW	Dev-SW
100	0	4.90	4.65	4.65	3.00	0.00	4.90	4.80	62.30	43.85	43.85	**100.00**	0.40	62.30	63.85
200	0	3.75	4.60	4.60	3.75	0.00	3.75	3.75	95.00	89.85	89.85	**100.00**	5.90	95.00	95.75
400	0	4.50	4.15	4.15	3.60	0.00	4.50	4.20	99.95	99.95	99.95	**100.00**	45.80	99.95	99.95
800	0	4.45	4.50	4.50	3.05	0.05	4.45	5.10	100.00	100.00	100.00	**100.00**	98.40	100.00	100.00
100	20	4.90	5.10	4.80	3.50	0.10	30.35	12.40	50.55	34.60	50.70	**100.00**	0.20	78.15	80.00
200	20	5.45	5.15	5.35	4.35	0.00	55.20	20.25	88.05	80.05	91.55	**100.00**	2.00	98.85	98.85
400	20	5.00	5.25	4.45	3.30	0.05	88.00	37.90	99.75	99.60	99.90	**100.00**	25.20	100.00	100.00
800	20	5.35	5.60	5.00	2.60	0.20	99.60	68.80	100.00	100.00	100.00	**100.00**	89.25	100.00	100.00
100	50	4.45	4.95	3.25	3.75	0.35	99.95	94.25	40.35	26.55	56.95	**100.00**	0.30	99.80	99.80
200	50	5.70	6.30	2.65	3.40	0.55	100.00	99.75	82.35	72.05	92.80	**100.00**	0.60	100.00	100.00
400	50	5.45	5.15	1.60	3.20	0.85	100.00	100.00	99.50	99.05	99.85	**100.00**	7.95	100.00	100.00
800	50	4.55	4.10	1.35	3.60	0.65	100.00	100.00	100.00	100.00	100.00	**100.00**	55.75	100.00	100.00
100	80	4.46	4.67	1.73	2.64	1.98	100.00	100.00	8.52	5.33	21.36	**92.03**	1.12	100.00	100.00
200	80	4.28	4.58	1.83	3.16	2.04	100.00	100.00	24.49	14.77	47.00	**99.90**	2.49	100.00	100.00
400	80	5.07	5.23	0.72	4.20	3.02	100.00	100.00	59.92	46.44	83.24	**100.00**	2.20	100.00	100.00
800	80	4.90	5.01	0.46	3.41	2.17	100.00	100.00	93.86	89.73	98.61	**100.00**	4.44	100.00	100.00

We also evaluated the performances of NRSP model diagnostics when we reject models with $p_{\min} \leq 0.05$; see the appended Table S1. As expected, the powers of NRSP tests with $p_{\min} \leq 0.05$ become smaller than those of NRSP tests with random p-values because $p_{\min}$ is a p-value upper bound. However, the powers of NRSP-AOV with $p_{\min} \leq 0.05$ are still near 100% for this example, although $p_{\min}$ is generally conservative.

5. A Real Data Example

This section will demonstrate the power of the model diagnostics with NRSP residuals using a recurrence-free times dataset of breast cancer patients [38, 39]. A cohort study of breast cancer in a large number of hospitals was carried out by the German Breast Cancer Study Group to compare three cycles of chemotherapy with six cycles and also to investigate the effect of additional hormonal treatment consisting of a daily dose of 30 mg of tamoxifen over two years. The patients in the study had primary histologically proven non-metastatic node-positive breast cancer who had been treated with mastectomy. The dataset is consolidated from 41 centres with a total of 686 patients. The censoring rate is 56.5%. The response variable of interest is the recurrence-free time, which is the time from entry to the study until a recurrence of cancer or death. We consider the following covariates: the tamoxifen treatment indicator, patient age, menopausal status, size and grade of the tumour, number of positive lymph nodes, progesterone and estrogen receptor status. More descriptions of these variables can be found from [38] and Table S2 in the appendix.

We fitted Weibull, log-logistic and log-normal AFT models with the available variables to the recurrence-free failure times. Table S3 (in the appendix) shows the estimated regression coefficients, the corresponding standard errors and p-values for the covariate effects from fitting the three AFT models. The third column of Figure 4 displays the estimated CHFs of CS residuals of the three models. All of the three curves appear to align well with the 45° straight line. The confidence bands estimated with at least one uncensored observation contain the 45° straight line. However, we will show that all of the three models misfit the data with model diagnostics based on NRSP and NUSP residuals.

We calculated the NRSP residuals of the three AFT models. The first and second columns of Figure 4 present the scatterplots and QQ plots of the NRSP residuals versus the index for each model. For the Weibull and log-logistic models, their NRSP residuals skew to the left; the QQ plots of the NRSP residuals also deviate from the 45° straight line in the upper tail. These observations suggest that a more appropriate model for the dataset should assign more probability to the right than Weibull and log-logistic models (note that small NRSP residuals correspond to large T_i^*). In contrast, for the log-normal model, the NRSP residuals are mostly between -3 and 3 and do not exhibit a visible pattern; the QQ plot of NRSP residuals aligns well with the 45° straight line.

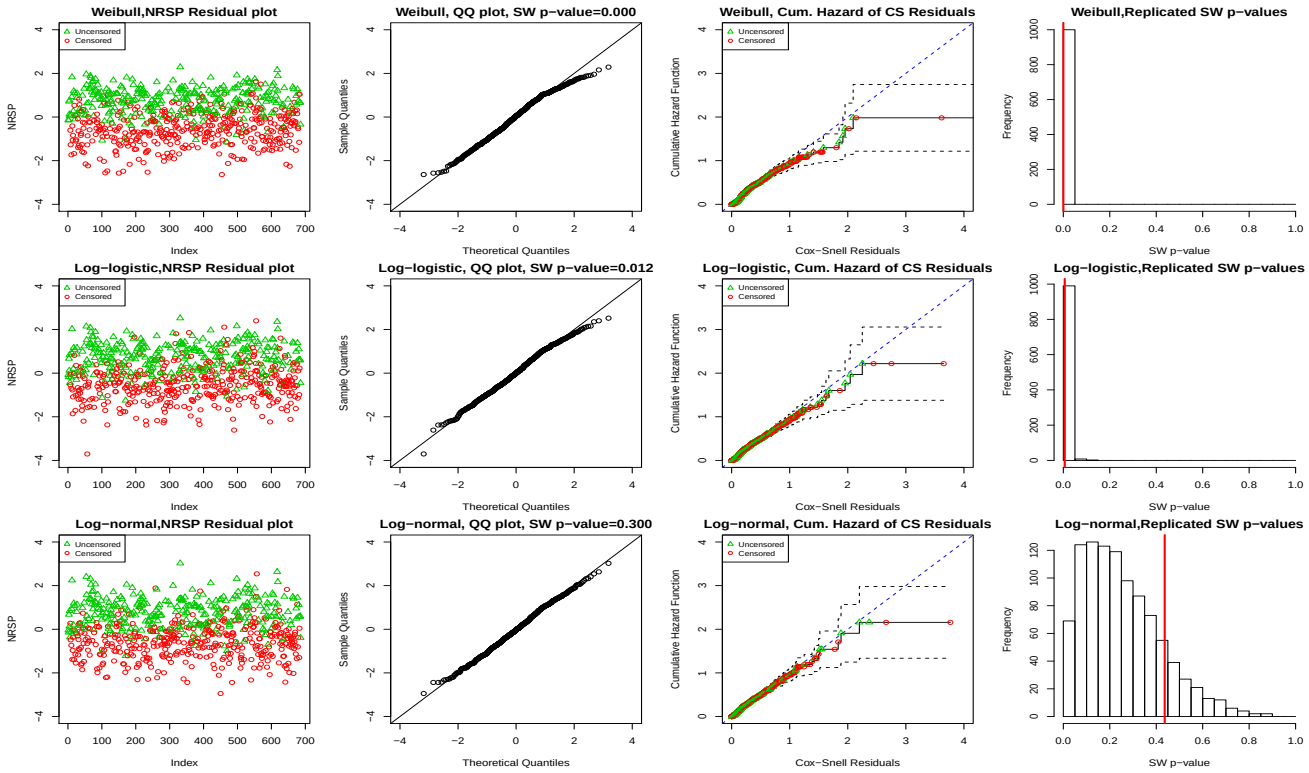


Figure 4. NRSP residuals of the Weibull, log-logistic, and log-normal AFT models fitted to the breast cancer patients dataset. The last column presents the histograms of 1000 replicated NRSP-SW p-values of each model. The vertical red lines indicate $p_{\min}$ calculated with the 1000 replicated NRSP p-values . An animated display of this figure for replicated NRSP residuals except the last column is shown in the URL given in Section B.

We further conducted GOF tests for the three models. The NUSP-CSF p-values for the three models are given in the 2nd row of Table 3, which shows that the Weibull and log-logistic models do not fit the data well, and the log-normal model appears a good fit. One difficulty in applying NRSP test methods is the fluctuation in test p-values due to the randomness in generating NRSP residuals. One way to remedy is to generate multiple sets of NRSP residuals. We then look at the histograms of the replicated test p-values and calculate the p-value upper bound $p_{\min}$ as described in Section 3.5. We generated 1000 realizations of the NRSP residuals for this dataset, and consequently, we obtained 1000 replicated NRSP-SW p-values for each model. We show an animated display of the scatterplot and QQ plots of the replicated NRSP residuals in the URL given in Section B. From the animation, we see that we can identify the misfits of Weibull and log-logistic models in most sets of replicated residuals. The fourth column of Figure 4 displays the histograms of 1000 replicated NRSP-SW test p-values. The $p_{\min}$ and the percentages of replicated NRSP-SW and NRSP-SF p-values being < 0.05 for each model are given in Table 3. The $p_{\min}$ values of the Weibull and log-logistic models are also significantly smaller than 0.05. The small $p_{\min}$ values provide strong evidence that the Weibull and log-logistic models do not fit the dataset well, but the log-normal seems a good model for the dataset with these GOF tests.

We also calculated the NMSP and deviance residuals for the three models. Figure S7 and S8 in the appendix display the residual plots and the QQ plots of the NMSP and deviance residuals. We see that all of these residuals deviate from a

Table 3. Model diagnostic test p-values or $p_{\min}$ for NRSP tests and AIC values of different models for the breast cancer data. The numbers in brackets for NRSP tests are the percentages of replicated NRSP test p-values being ≤ 0.05.

Model	Weibull	log-logistic	log-normal	log-normal with log(nodes)
AIC	5181	5153	5139	5121
NUSP-CSF p-values	1.69e-5	1.94e-3	0.133	0.172
NRSP-SW $p_{\min}$	2.43e-05 (100.00)	6.01e-03 (99.00)	4.36e-01 (6.90)	5.29e-01 (5.70)
NRSP-SF $p_{\min}$	9.50e-05 (100.00)	1.52e-02 (97.00)	4.85e-01 (5.80)	6.37e-01 (3.50)
NRSP-AOV $p_{\min}$	1.97e-02 (60.40)	7.21e-02 (46.00)	**5.22e-02 (52.40)**	9.99e-01 (0.50)

normal distribution due to censoring. Therefore, the NMSP and deviance residuals fail to distinguish the GOFs of these three models to this dataset.

We also compared the model checking results based on the NRSP residuals to the model comparison results based on AICs. First, let us clarify the difference between model *checking* and model *comparison*. AIC is a measure of the out-of-sample predictive performance of a model. Even when none of the models in a set fit a dataset well, AIC will always choose one as the best model. Therefore, the model comparison alone is insufficient for model evaluation. We still need to conduct model diagnostics for the best model chosen by AIC or other information criteria. On the other hand, the model with better AIC is believed to fit the dataset better. Therefore, it is still meaningful to compare the model comparison results based on AIC with the model diagnostics results. Table 3 displays the AIC values of the three fitted AFT models. We see that the log-normal model has a lower AIC than the Weibull and log-logistic models.

The GOF tests with NRSP residuals and NUSP-CSF report fairly large p-values for the log-normal model. Despite these GOF test results, we further checked whether the linear assumption is plausible with the NRSP-AOV method (*described in Section 4.2*). We calculated 1000 replicated NRSP-AOV p-values. The top-left plot of Figure 5 shows the histogram of these replicated NRSP-AOV p-values for the log-normal model. The percentage of NRSP-AOV p-values < 0.05 is about 50%, and the $p_{\min}$ is only slightly larger than 0.05. Therefore, we have fair evidence to suspect that there is non-linearity. We drew the NRSP residuals against each covariate and linear predictor, one of which is shown in the top-right plot of Fig. 5. This plot shows that large "nodes" values appear to have small NRSP residuals (not symmetric about 0). We tried a logarithm transformation for nodes and fitted a log-normal model with log(nodes) and other variables. This model is labelled as "log-normal with log(nodes)" in Table 3 and Fig. 5. We see that the NRSP residuals of this model are more homogeneous along with the variable "nodes", as shown in the bottom-right plot of Fig. 5. The replicated NRSP-AOV p-values (bottom-left plot of Fig. 5) are mostly larger than 0.05 with $p_{\min}$ near 1. Furthermore, the AIC value of this model, as shown in the last column of Table 3, supports that the log transformation results in a better model. This above analysis clearly shows that the non-linearity diagnostics with NRSP residuals successfully detect the non-linear effect of "nodes" in this dataset. However, the effect is too subtle to be detected by the above GOF tests, including the NUSP-CSF test.

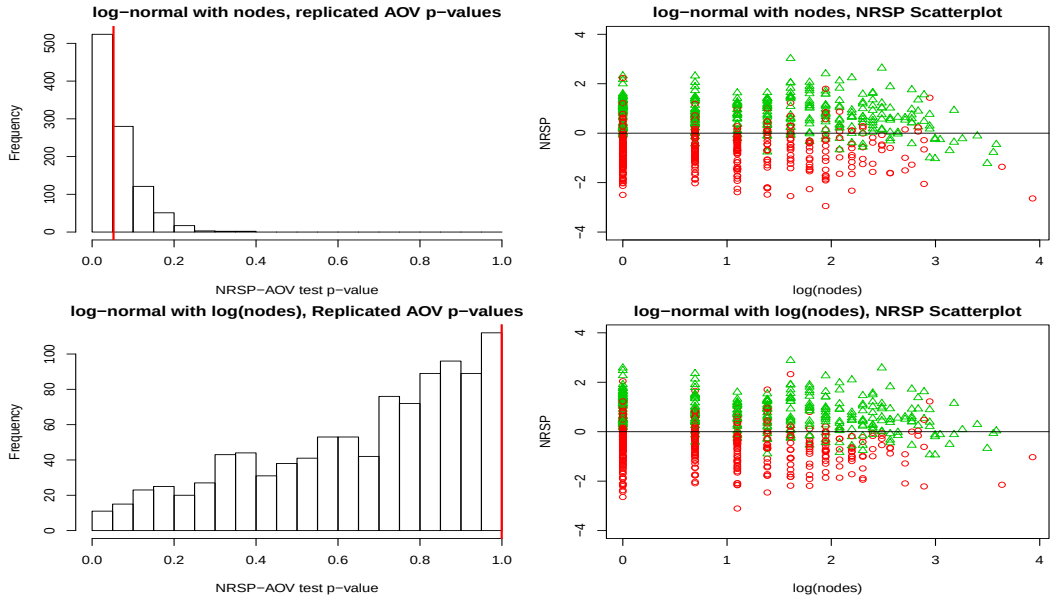


Figure 5. NRSP non-linearity residual diagnosis for the breast cancer data. The red vertical lines in the histograms of replicated NRSP-AOV p-values show the $p_{\min}$ values.

6. Conclusions and Discussions

This paper has proposed using randomized survival probabilities (RSPs) to conduct model diagnostics for censored regression. We have proved that RSPs always have the uniform distribution on $(0, 1)$ under the true model. Consequently, NRSP residuals are approximately distributed with $N(0, 1)$ under the true model. With this unified reference distribution for NRSP residuals, we can conduct a wide variety of residual diagnostics for censored regression. Our simulation studies show that, although the GOF tests with NRSP residuals are not as powerful as a traditional GOF test method, a non-linearity test with NRSP residuals has significantly higher power in detecting non-linearity. The real data analysis shows that the NRSP residual diagnostics successfully captures a subtle non-linear relationship in the dataset, which is not detected by the graphical diagnostics with the CS residuals and existing GOF tests.

The NRSP residual for one-sided censored regression, as described in this article, can be easily extended to interval-censored regression by drawing a random number between the two survival probabilities calculated on the two bounds of each interval. The NRSP residual for interval-censored regression can be regarded as an extension of the randomized quantile residual [40] for count regression if we consider a count observation as an observation of a continuous variable censored by fixed integer intervals.

To overcome the randomness in NRSP testing p-values, we need to devise valid methods to obtain non-random GOF test p-values based on NRSP residuals. In this article, we describe a method for obtaining a p-value upper bound $p_{\min}$. $p_{\min}$ is informative when the model departure is clear. However, it is generally conservative. We desire to have a more powerful and non-random summary of the replicated NRSP test p-values. Our preliminary results (not shown in this article) show that averaging replicated NRSP test p-values can boost the discriminative power of NRSP-SW and NRSP-SF methods to the power of NUSP-CSF, as measured by ROC curves. However, the averages of replicated NRSP-SW or NRSP-SF p-values are no longer uniformly distributed. We believe that to characterize the distribution of the average of replicated NRSP testing p-values or to obtain a rule of thumb is an interesting topic.

This article shows that a simple non-linearity test by applying ANOVA to NRSP residuals has superior power than GOF tests in detecting non-linearity. We expect that many other specific model mis-specification tests that target a particular model discrepancy have higher powers than GOF tests. For example, statistical tests for checking proportional hazard assumption in Cox regression seem to be demanded very often; see [24] among others. We believe that developing specific quantitative and graphical diagnostic tools based on NRSP residuals will be fruitful because of the explicit characterization of the approximate standard normal distribution of NRSPs under the true model.

A. Additional Figures and Tables

Additional figures, tables, and simulation results can be found from `https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fsim.8852&file=sim8852-sup-0001-supinfo.pdf`.

B. R Functions, Demonstration Examples, and Data Availability

R functions for computing NRSP residuals for `survreg` and `coxph` objects with demonstration examples and the dataset used in this paper are available here: `https://longhaisk.github.io/software/NRSP/`.

Acknowledgement

We much appreciate the anonymous referees of *Statistics in Medicine* for their useful comments to improve this paper.

References

1. Cox DR, Snell EJ. A General Definition of Residuals. *Journal of the Royal Statistical Society. Series B (Methodological)* 1968: 248–275.
2. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 1951; 46(253): 68-78. doi: 10.1080/01621459.1951.10500769
3. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American statistical association* 1958; 53(282): 457–481.
4. Hirsch RM, Stedinger JR. Plotting Positions for Historical Floods and Their Precision. *Water resources research* 1987; 23(4): 715–727.
5. Verrill S, Johnson RA. The Asymptotic Equivalence of Some Modified Shapiro-Wilk Statistics–Complete and Censored Sample Cases. *The Annals of Statistics* 1987; 15(1): 413–419.
6. Verrill S, Johnson RA. Tables and Large-Sample Distribution Theory for Censored-Data Correlation Statistics for Testing Normality. *Journal of the American statistical Association* 1988; 83(404): 1192–1197.
7. Royston P. A Toolkit for Testing for Non-Normality in Complete and Censored Samples. *Journal of the Royal Statistical Society: Series D (The Statistician)* 1993; 42(1): 37–43.
8. Koziol JA, Green SB. A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika* 1976; 63(3): 465–474. doi: 10.1093/biomet/63.3.465

9. Csorgo S, Horvath L. On the Koziol-Green Model for Random Censorship. *Biometrika* 1981; 68(2): 391–401. doi: 10.2307/2335584
10. Chen CH. A Correlation Goodness-of-Fit Test for Randomly Censored Data. *Biometrika* 1984; 71(2): 315–322. doi: 10.1093/biomet/71.2.315
11. Akritas MG. Pearson-Type Goodness-of-Fit Tests: The Univariate Case. *Journal of the American Statistical Association* 1988; 83(401): 222–230.
12. Hollander M, Pena EA. A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data. *Journal of the American Statistical Association* 1992; 87(418): 458–463.
13. Cao J, Moosman A, Johnson VE. A Bayesian Chi-Squared Goodness-of-Fit Test for Censored Data Models. *Biometrics* 2010; 66(2): 426–434. doi: 10.1111/j.1541-0420.2009.01294.x
14. Kim N. Tests Based on EDF Statistics for Randomly Censored Normal Distributions When Parameters Are Unknown. *Communications for Statistical Applications and Methods* 2019; 26(5): 431–443. doi: 10.29220/CSAM.2019.26.5.431
15. Millard SP. EnvStats: Package for Environmental Statistics, Including US EPA Guidance. 2018.
16. Steven P. Millard . *EnvStats: An R Package for Environmental Statistics*. New York, NY: Springer. 2nd ed. 2013.. ed. 2013.
17. Collett D. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC . 2015.
18. Therneau TM, Grambsch PM, Fleming TR. Martingale-Based Residuals for Survival Models. *Biometrika* 1990; 77(1): 147-160. doi: 10.1093/biomet/77.1.147
19. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media . 2013.
20. Nardi A, Schemper M. New Residuals for Cox Regression and Their Application to Outlier Screening. *Biometrics* 1999; 55(2): 523-529. doi: 10.1111/j.0006-341X.1999.00523.x
21. Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 1965; 52(3/4): 591–611.
22. Shapiro SS, Francia RS. An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association* 1972; 67(337): 215–216.
23. Peng Y, Taylor JMG. Residual-Based Model Diagnosis Methods for Mixture Cure Models. *Biometrics* 2017; 73(2): 495-505. doi: 10.1111/biom.12582
24. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* 1994; 81(3): 515–526.
25. Keleş S, Segal MR. Residual-Based Tree-Structured Survival Analysis. *Statistics in Medicine* 2002; 21(2): 313-326. doi: 10.1002/sim.981
26. Farrington CP. Residuals for Proportional Hazards Models with Interval-Censored Survival Data. *Biometrics* 2000; 56(2): 473-482. doi: 10.1111/j.0006-341X.2000.00473.x
27. Davison AC, Gigli A. Deviance Residuals and Normal Scores Plots. *Biometrika* 1989; 76(2): 211-221. doi: 10.1093/biomet/76.2.211
28. Lin DY, Wei LJ, Ying Z. Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. *Biometrika* 1993; 80(3): 557-572. doi: 10.2307/2337177
29. Law M, Jackson D. Residual Plots for Linear Regression Models with Censored Outcome Data: A Refined Method for Visualizing Residual Uncertainty. *Communications in Statistics - Simulation and Computation* 2017; 46(4): 3159–3171. doi: 10.1080/03610918.2015.1076470
30. Shepherd BE, Li C, Liu Q. Probability-Scale Residuals for Continuous, Discrete, and Censored Data. *The Canadian journal of statistics = Revue canadienne de statistique* 2016; 44(4): 463–479. doi: 10.1002/cjs.11302
31. Hillis SL. Residual Plots for the Censored Data Linear Regression Model. *Statistics in Medicine* 1995; 14(18): 2023–2036. doi: 10.1002/sim.4780141808
32. George B, Seals S, Aban I. Survival Analysis and Regression Models. *Journal of Nuclear Cardiology* 2014; 21(4): 686-694. doi: 10.1007/s12350-014-9908-2
33. Carroll KJ. On the Use and Utility of the Weibull Model in the Analysis of Survival Data. *Controlled Clinical Trials* 2003; 24(6): 682 - 701. doi: https://doi.org/10.1016/S0197-2456(03)00072-2
34. Caraux G, Gascuel O. Bounds on Distribution Functions of Order Statistics for Dependent Variates. *Statistics & probability letters* 1992; 14(2): 103–105.
35. Rychlik T. Stochastically Extremal Distributions of Order Statistics for Dependent Samples. *Statistics & probability letters* 1992; 13(5): 337–341.
36. Johnson VE. Bayesian Model Assessment Using Pivotal Quantities. *Bayesian Analysis* 2007; 2(4): 719–733. doi: 10.1214/07-BA229
37. Yuan Y, Johnson VE. Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models. *Biometrics* 2012; 68(1): 156–164. doi: 10.1111/j.1541-0420.2011.01668.x
38. Schumacher M, Bastert G, Bojar H, et al. Randomized 2 x 2 Trial Evaluating Hormonal Treatment and the Duration of Chemotherapy in Node-Positive Breast Cancer Patients. German Breast Cancer Study Group.. *Journal of Clinical Oncology* 1994; 12(10): 2086-2093. doi: 10.1200/JCO.1994.12.10.2086
39. Sauerbrei W, Royston P. Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1999; 162(1): 71-94. doi: 10.1111/1467-985X.00122
40. Dunn PK, Smyth GK. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* 1996; 5(3): 236–244.