EVALUATION OF GOODNESS-OF-FIT TESTS FOR THE COX PROPORTIONAL
HAZARDS MODEL WITH TIME-VARYING COVARIATES

by

Shanshan Hong

Bachelor of Science
Anqing Normal University, 2009

Master of Science
Nanjing Agriculture University, 2012

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2017

Accepted by:

Jiajia Zhang, Director of Thesis

James W Hardin, Reader

Robert Moran, Reader

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ACKNOWLEDGMENTS

This thesis defense would not be possible without the guidance of my committee members, and support from my family.

First and foremost, I would like to express my deepest appreciation to my thesis advisor, Dr. Jiajia Zhang for her continual advice and support thought my study. Without her guidance and persistent assistance, this thesis couldn't be completed. Her advice on both research as well as my career is priceless.

I would like to thank Dr. James W. Hardin and Dr. Robert Moran for serving as my committee members. Thank you for your brilliant suggestions and comments; you made my defense an enjoyable experience.

In addition, I would like to thank Erica Sercy for assistance with the grammar of my thesis.

At last, I would like to thank my beloved family including my parents, my husband and my lovely child. Without your endless love and support, I would not have been able to complete my study in the United States.

ABSTRACT

The proportional hazards (PH) model, proposed by Cox (1972), is one of the most popular survival models for analyzing time-to-event data. To use the PH model properly, one must examine whether the data satisfy the PH assumption. An alternative model should be suggested if the PH assumption is invalid. The main purpose of this thesis is to examine the performance of five existing methods for assessing the PH assumption. Through extensive simulations, the powers of five different existing methods are compared; these methods include the likelihood ratio test, the Schoenfeld residuals test, the scaled Schoenfeld residuals test, Lin et al. (2006) score test, and the martingale-based residuals test. Results from SAS and R show that the power will vary depending on the form of hazard. For the hazard considered here with a clear jump point at which the PH assumption is violated, the power depends on the time to the violation of proportional hazards, the direction and magnitude of the hazard's change, and the censoring rate of the data. Leukemia remission and Stanford heart transplant data were used to illustrate testing of the five methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

As an important tool in statistical data analysis, survival analysis is a major field in which the association is analyzed between the time to an event and potential risk factors. Time is usually defined as the time to an event of interest. The event can be defined as, for example, death or date of diagnosis in public health, machine failure in engineering, or account closing in business. The particular feature of survival data is that time may not be observed completely. That is, some patients may be censored. We can classify censoring into three categories: right censoring, left censoring, and interval censoring. Right censoring means that the participant does not experience the event of interest, which happens due to loss to follow-up, withdrawal from the study of the participant, or the end of the study. That is, the true event time is greater than the time under which the subject was in the study (assuming all subjects eventually experience the event). Left censoring means that the event of interest occurred prior to when the subject enters the study. That is, the true survival time is less than or equal to the first observed time. Interval censoring means that the event had not happened at the previous visit time, but had occurred by the current visit time. That is, the true event time happened within a

known time interval. In this thesis, we focus mainly on data with right censoring. In contrast to traditional regression analysis, survival analysis is conducted to describe the probability that an event does not happen. Thus, we list the definitions of the hazard function and the survival function below.

Let us denote T as the subject's event time. The hazard function, which is denoted by $h(t)$, represents the instantaneous failure rate given that an individual has survived up to time t; this can be written as:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

where Δt denotes an interval of time.

The survival function, denoted as $S(t)$, is used to describe the probability that a person can survive longer than the specified time t. Theoretically, as time t ranges from 0 to infinity, the survival probability decreases from 1 to 0. The relationship between the survival function and the hazard function is expressed as:

$$S(t) = \exp\left[-\int_0^t h(u)d(u)\right]$$
$$h(t) = \frac{dS(t)/dt}{S(t)}$$

Note, $H(t) = \int_0^t h(u)d(u)$ denotes the cumulative hazard risk. There are many parametric distributions that can be used to describe the survival/hazard function, including the exponential distribution, the Weibull distribution, and the lognormal distribution. The maximum likelihood method can be applied to obtain estimates.

Additionally, nonparametric approaches can be used to estimate the survival probability, such as the Kaplan–Meier approach and the Nelson–Aalen approach ([1], [2]).

For a brief introduction of the Kaplan–Meier (KM) method, suppose there are n observations denoted as $t_1, t_2, \cdots, t_n$, and the sorted observation times are denoted as $t_{(i)}$, the estimated survival probability at time t based on the KM approach is given as:

$$\hat{S}_{km}(t) = \prod_{i=1}^{n-1} \widehat{Pr}(T > t_{(i)} | T \geq t_{(i)})$$

The KM estimator can be used to compare survival probability between two different groups based on the log-rank test.

1.1 PH model and PH model assumption

In practice, most researchers want to determine the potential risk factors of an event. Therefore, these potential risk factors should be incorporated into the survival model. The proportional hazards (PH) model, proposed by Cox in 1972 [3], is one of the most popular survival models in practice. Let us assume that X= $(X_1, \cdots X_p)$ is a vector of covariates of interest, and $\beta = (\beta_1, \cdots, \beta_p)$ is a vector of coefficients. The PH model can be written as:

$$h(t, X) = h_0(t) \times e^{\sum_{i=1}^{p} \beta_i X_i}$$

The interpretation of β can be explained by the log hazard ratio. For example, if we want to calculate the hazard ratio of two patients with one variable of interest, Z, and

we have the corresponding data z_1, z_2 and coefficient β', the hazard ratio for the two

patients of the variable Z is:

$$\text{HR} = \frac{h_0(t)\beta' e^{z_1}}{h_0(t)e^{\beta' e^{z_2}}} = e^{\beta'(z_1 - z_2)}$$

If Z is a continuous variable, we can see that with one unit change in Z, the hazard

ratio will change $e^{\beta'}$, and if Z is a categorical variable, if we compare z_1 to z_2, the

hazard ratio is $e^{\beta'}$.

One of the reasons the PH model is popular in literature is because it can be

estimated semi-parametrically. That is, it is not necessary to assume the distribution of

the baseline hazard function, and the coefficient estimates can be obtained through the

partial likelihood function. Assume the observation for the ith subject is denoted as

$t_i, \delta_i, X_{ii} = (x_{i1}, \cdots, x_{ip})$, where δ_i is the censoring indicator, with 1 as event and 0 as

censoring. The partial likelihood is denoted as:

$$\text{L}(\beta) = \prod_{i:\delta i\, =1} \frac{\exp(\beta\, x_{ii})}{\sum_{j\epsilon R(t_{(i)})} \exp(\beta X_{ji})}$$

$R(t_{(i)})$ is the risk set at time $t_{(i)}$, which includes all subjects who are alive at

time $t_{(i)}$.

The first derivative of the logarithm of partial likelihood is:

$$U(\beta) = \frac{\partial l(\beta)}{\partial(\beta)} = \frac{\sum_{i:\delta i=1}[\beta X_{ii} - \log(\sum_{j\epsilon R(t_{(i)})} \exp(\beta X_{ji}))]}{\partial(\beta)} = \sum_{i:\delta i=1}[X_{ii} - \frac{\sum_{j\epsilon R(t_{(i)})} X_{ji}\exp(\beta X_{ji})}{\sum_{j\epsilon R(t_{(i)})} \exp(\beta X_{ji})}]$$

The estimation can be obtained by maximizing the partial likelihood function or

the root of its first derivative. This maximization procedure can be easily implemented

using the Newton–Raphson method, which can be performed in existing packages in

most statistical software. For example, the PH model can be estimated directly in software such as through the "coxph" function in the survival package in R, "phreg" in SAS, and "stcox" in Stata.

Once the estimate of β is obtained, the survival function can be further estimated non-parametrically. The estimated baseline survival function is:

$$[\widehat{S_0}(t)] = \exp\left(-\hat{h}_0(t)\right),$$

where $\hat{h}_0(t)$ is the estimated cumulative baseline hazard function:

$$\hat{h}_0(t) = \sum_{i:ti \,\leq t} \frac{\delta_i}{\sum_{j \epsilon R(t_{(i)})} \exp(\hat{\beta} X_{ji})}$$

The prediction of survival probability for a future patient can be made based on the following:

$$\hat{S}\,(t, X) = [\widehat{S_0}(t)]^{\exp(\Sigma_{i=1}^{p} \hat{\beta}_i X_i)}$$

It is worthwhile to point out that even though there is no specific assumption in PH model estimation, there is an assumption of the model itself. As mentioned above, the PH model can be written in two parts: the time-dependent baseline hazard function and the time-independent regression component. Therefore, the hazard ratio is a time-independent component that leads to the main assumption of the PH model.

1.2 Data examples for which the PH model assumption is not satisfied

In practice, the PH assumption may not be valid ([4], [5]). That is, the hazard ratio is time dependent or the model cannot be written as two distinct parts. Therefore, it is important and necessary to check the PH assumption before estimating and interpreting the results.

First, we write the PH model in terms of the log-log survival function, given by:

$$\log (\text{-log } \hat{S} (t)) = \log\left(- \log(\hat{S}_{km}(t)\right) + \sum_{i=1}^{n} \beta_i X_i$$

For any covariate x, the log-log survival function will be shifted by a constant related to covariate x. Thus, the PH assumption can be graphically checked by plotting time and the non-parametrically estimated survival function (KM estimator). If the log-log survival curves are parallel with regards to covariate x, we can conclude that the covariate x satisfies the PH assumption.

Example I: Data on leukemia remission [6] were obtained from clinical trial data of the drug 6-mercaptopurine (6-MP). The main purpose of this trial was to compare whether the use of 6-MP leads to significant improvement in remission versus the placebo. In total, there were 42 patients with acute leukemia in the data, separated into two groups of 21 patients each. This trial used matched pairs of patients at a given hospital matched on remission status (complete or partial) and random pairs of patients on treatment or placebo. Patients were followed until their leukemia returned or the study

ended (time measured in months). In this dataset, there are five variables: survival time, remission status, sex, log white blood cell count (log WBC), and treatment status. The censoring rate is 28.5%.

We want to use the PH model to quantify the impact of 6-MP on remission. Before modeling the data, the PH assumption should be examined. For the main exposure (6-MP vs placebo), the log-log KM curve of treatment versus time to remission was plotted using SAS (top panel in Figure 1). We notice that the estimated KM curves for the 6-MP and placebo groups are roughly parallel, meaning that the treatment satisfies the PH assumption. Additionally, sex is an important confounder and needs to be justified in the PH model, so the log-log KM curves of sex versus time to remission were also plotted using SAS (bottom panel in Figure 1). The estimated survival curves for males and females cross, indicating that sex does not satisfy the PH assumption.

Example II: Next we consider Stanford heart transplant data. This study lasted from 1967 to 1974, with the main aim of determining whether the heart transplant program was beneficial for end-stage heart disease patients. Patients were required to not have had a transplant previously and to have no other conditions that may impede post-transplant recovery. During the time waiting for a transplant, patients were monitored for improvement. After transplantation, patients received post-operative care and were followed as long as possible [7].
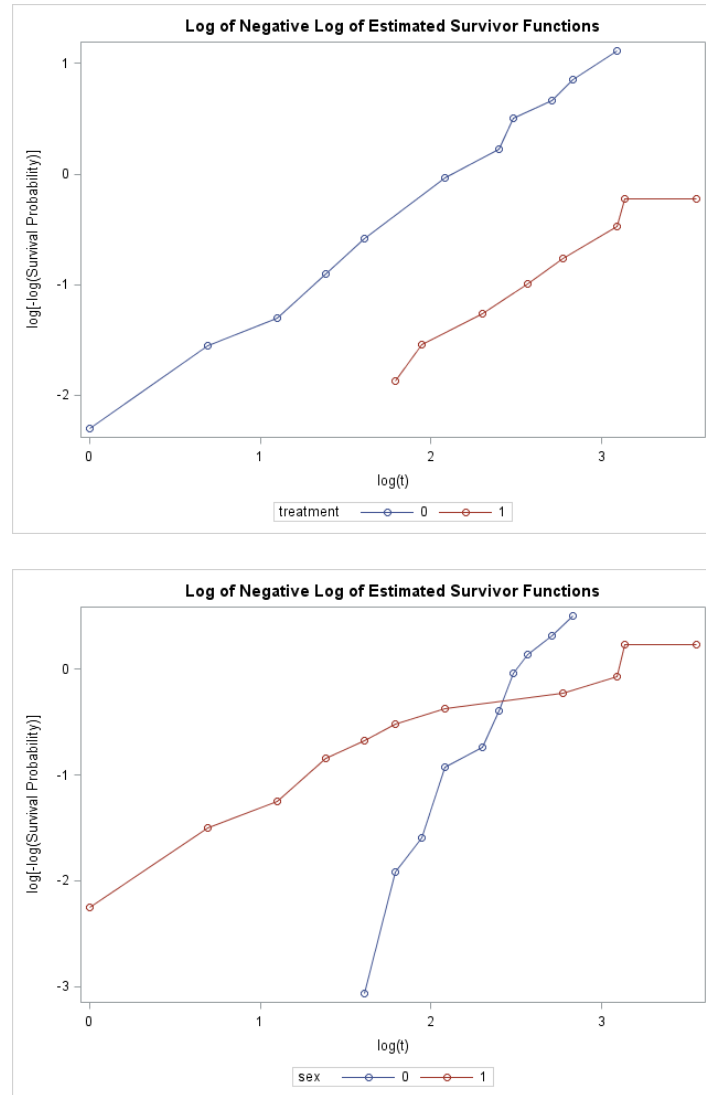
Figure 1.1 Log-log KM curve of leukemia remission data for treatment (top) and sex (bottom)

According to Crowley and Hu [8], the data contain 103 patients, 69 of whom received a heart transplant, and the censoring rate is 27%. The dataset also contains the following variables: birth date, age at acceptance into the program, date of transplant, waiting time, the age at transplant, and previous surgery.

8

As transplant status is the main variable of interest in the dataset, we first must check if it satisfies the PH assumption, shown below using log-log survival curves.
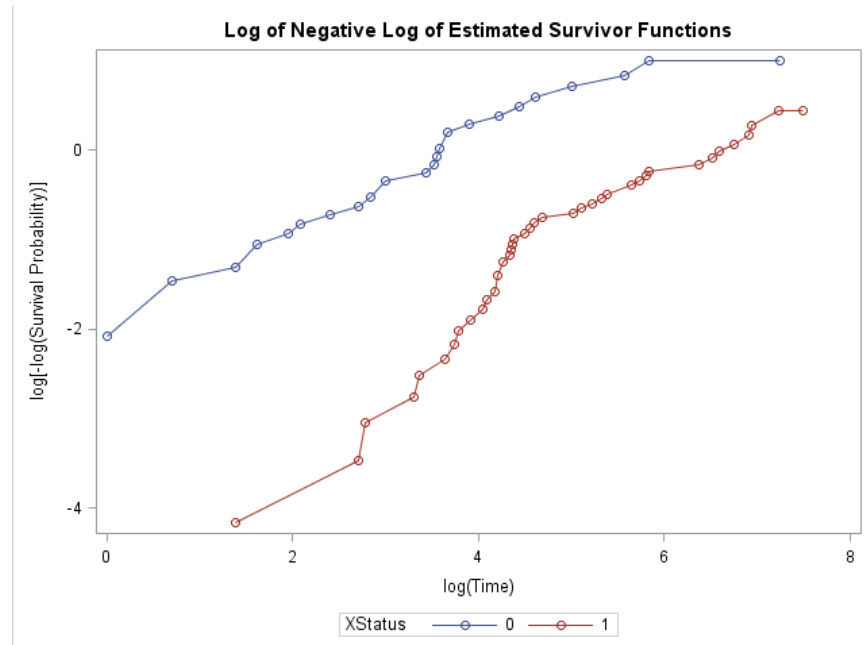
Figure 1.2 Log-log KM curve of Stanford heart transplant data

According to Figure 1.2, the two curves don't intersect but are instead parallel. Form this graph, we think the transplant effect is constant over time.

1.3 Outline of thesis

In the examples described above, graphical tools are used to evaluate the PH assumption. However, sometimes it is hard to decide "how parallel is parallel." A formal test with a corresponding p-value is easier to use in practice.

The main purpose of this thesis is to compare five different existing statistical methods for checking the PH assumption with a valid testing procedure and P-value. We will implement all methods in both SAS and R and then compare the power of each test under different settings in which the PH assumption is violated.

In Chapter 2, we introduce the five exisiting methods: the likelihood ratio test, the Schoenfeld residuals test, the scaled Schoenfeld residuals test, the Lin et al. (2006) score test [9], and the martingale-based residuals test. We assess these methods using R and SAS software. Extensive simulation studies were conducted and associated results are presented in Chapter 3. In section 3.1, we first introduce simulation settings under the piecewise exponential distribution. In section 3.2, we present results for the power of each test under simulation settings. We also compare the results from SAS and R. In Chapter 4, we apply the five methods to the leukemia remission data and Stanford heart transplant data. Finally, we finish with a discussion and conclusions in Chapter 5.

CHAPTER 2 PH MODEL ASSUMPTION CHECKING METHODS

Although a PH model with a time-dependent variable is a challenge, many methods have been proposed to deal with the situation. For example, Therneau and Grambsch [10] advised plotting a smooth curve based on Schoenfeld residuals. Fleming and Harrington [11] suggested a procedure that involves a function of time as a time-dependent variable in the PH model.

A time-dependent variable, whose values may differ over time, can be written as X(t) and can be classified as an internal time-dependent covariate, for example, blood pressure, or an external time-dependent covariate, like air pollution. In this situation, the PH model may still be used, but such data don't satisfy the PH assumption. We call the model an extended-Cox model and the variable a time-dependent variable. The general formula is shown below:

$$h(t,X(t)) = h_0(t) \ \exp[\sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \varphi_j X_j(t)]$$

In the extended model, the exponential portion contains two components: time-independent covariates (X_i) and time-dependent covariates ($X_j(t)$). This model is very useful when the covariate changes over time, and it also can be used to check the PH assumption.

The partial likelihood function of a time-dependent variable is:

$$L(\varphi)=\prod_k \frac{\prod_{j\epsilon H_j}\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))}{\prod_l^{m-1}[\sum_{j:Y_{j\geq t_k}}\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))-\frac{l}{m}\sum_{j\epsilon H_k}\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))}$$

Here t_k is the unique time, H_k is subject j with $Y_j=t_k$ and $\delta i=1$, $m_k=|H_k|$, and $X_{jj}(t)=(X_{j1}(t),\ X_{j2}(t)\cdots X_{jp2}(t))$ are the realized values of the covariates for subject j.

The first derivative of the logarithm of the partial likelihood is:

$$U(\varphi)=\sum_k(\sum_{j\epsilon H_k} X_{jj}-\sum_{l=0}^{m-1}\frac{[\sum_{j:Y_{j\geq t_k}}X_{jj}(t))*\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))-\frac{l}{m}\sum_{j\epsilon H_k}X_{jj}(t))*\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))}{[\sum_{j:Y_{j\geq t_k}}\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))-\frac{l}{m}\sum_{j\epsilon H_k}\exp(\sum_{j=1}^{n2}\varphi_j X_{jj}(t))}$$

The time-dependent coefficient model can be implemented in SAS [12] and R [13] through building a time-dependent covariate. SAS Institute Inc (2010) provides survival estimators in "proc phreg." The survival [15] package in R provides the estimates using the "coxph" function. We also can check PH model assumption using SAS and R via the methods described below.

In the following example, we consider a simple extended-Cox model that contains a fixed binary variable, X_1, and a time-dependent variable, $X_2(t)$. We assume X_2 may violate the PH assumption. In this case,

$$h(t,X(t))= h_0(t)\ \exp(\beta_1 X_1 + \beta_2 X_2(t)).$$

Although X_2 is a time-dependent variable, it can change over time. The hazard ratio is constant conditional on time t [15]. Testing the PH model assumption is equal to testing the coefficient β_2.

2.1 Likelihood ratio test

In the PH model, the likelihood ratio test can be used to compare the goodness of fit of the null and alternative models. The likelihood ratio can be interpreted as how many times more likely the data are under one model than the other. The two models are nested. The test shows the difference between the log of the ratio of the two likelihoods.

To test the PH assumption, Cox (1972) suggested adding a time-dependent variable, $X_2(t)$, in the model. The form of this variable can be the product of a function of time, g(t), and the variable of interest. Here we consider g(t)=t, with the following models:

$$L_{PH\ model} \text{ is: } h(t,X(t))= h_0(t)\ \exp(\beta_1 X_1 + \beta_2 X_2)$$

$$L_{ext.Cox\ model} \text{ is: } h(t, X(t))= h_0(t)\ \exp(\beta_1 X_1 + \beta_2 (X_2 * t))$$

The test statistics are:

$$LR = -2\log(L_{PH\ model})\text{-}(-2\log(L_{ext.Cox\ model}) \sim X_p^2$$

The test statistics follow a chi-squared distribution with p degrees of freedom, where p is equal to $df_{ext.Cox\ model}\text{-}df_{PH\ model}$.

In SAS, we could use the function "phreg" to fit the null and alternative models. In R, we use the function "coxph" to fit the null models and the "coxph" with "tt" function fit the alternative model [16]. The code is as follows:

fit1<-coxph(Surv(time,event)~x1+x2,data= Leukemia)

fit2<-coxph(Surv(time,event)~x1+x2+tt(x2),data=leukemia,tt = function(x, t, ...) x * t)

We can calculate the test statistics by extracting the value of the log likelihood from SAS and R and then getting the p-value.

2.2 Schoenfeld residuals test

Schoenfeld residuals are partial residuals that can be used to check the PH assumption. Schoenfeld residuals are obtained by substituting the partial likelihood estimator of the coefficient. The Schoenfeld residual for the ith subject on the pth covariate is:

$$\hat{r}_{ip} = \delta_i(x_{ip} - \hat{\bar{x}}_{w_{i,p}}) \qquad \hat{\bar{x}}_{wi,p} = \frac{\sum_{j \in R(ti)} x_{jp} e^{x'_j \beta_p}}{\sum_{j \in R(ti)} e^{x'_j \beta_p}}$$

The result is the covariate value of a person who gets the event at time t minus the expected value of the covariate and is not defined for censored individuals. Kumar [17] found that Schoenfeld residuals against time should be scattered randomly around zero, so testing time-dependent covariates is equivalent to testing a non-zero slope in a generalized linear regression of the Schoenfeld residuals on functions of time [18]. A non-zero slope is an indication of a violation of the PH assumption.

In R, we can get Schoenfeld residuals from the "coxph" function by using the "residual" option, and the sample code is:

14

fit<-coxph(Surv(time,event)~x1+x2,data= Leukemia)

rs<-residuals(fit,type="schoenfeld")

We then perform a regression model of Schoenfeld residuals and ranked time to get the p-value of the slope. A p-value less than 0.05 is considered to indicate evidence for a non-zero slope. In SAS, the functions "phreg" and "glm" are used in this test.

2.3 Scaled Schoenfeld residuals test

Grambsch and Therneau (1994) [19] suggested using scaled Schoenfeld residuals to test time-dependent variables. Scaled Schoenfeld residuals are:

$$\hat{r}_{ip}^* = [\widehat{Var}(\hat{r}_{i1},\hat{r}_{i2,}\cdots\hat{r}_{ip})]^{-1}(\hat{r}_{ip})$$

Grambsch and Therneau also suggested that $E(\hat{r}_{ip}^*) + \hat{\beta}_p \approx \beta_p(t_i)$, so when assume that PH assumption is valid, $\hat{\beta}_p= \beta_p(t_i)$, then $E(\hat{r}_{ip}^*) = 0$ which means that scaled Shoenfeld residuals are a random walk across the time scale. So the null hypothesis for the test is that the regression coefficient between the scaled Schoenfeld residuals and a function of time should be zero. Once the scaled Schoenfeld residuals are created, the p-value of the correlation between the scaled Schoenfeld residuals and ranked time can be obtained.

In R, the "cox.zph" function can be used to test the PH assumption. The sample code is:

fit<-coxph(Surv(time,event)~x1+x2,data= Leukemia)

zph<-cox.zph(fit,transform="rank")

The transformation of time can be "KM," "rank," or "identity". We used "rank" in our analysis. In SAS, we need to use "Phreg" to first get Schoenfeld residuals. Then we calculate the scaled Schoenfeld residuals [20] and perform the regression model of scaled Schoenfeld residuals against the ranked time variable.

2.4 Lin et al. (2006) score test

For all the tests described above, we need to specify the function of the time variable; however, Lin et al. (2006) propose "score-type" tests for the proportional hazards assumption in the PH model by using a natural smoothing spline representation of the corresponding non-parametric functions of time or covariate. In this test, the function of time doesn't need to be defined. They consider the alternative model as:

$$h(t|x_1, x_2) = h_0(t)exp\{X_1\beta + X_2\gamma(t)\},$$

where $h_0(t)$ is baseline hazard, and X_2 is a scalar covariate of interest. In our model, we consider X_2 as a time-dependent variable, $\gamma(.)$, as an arbitrary smooth function of time.

$\gamma(.)$ is infinite-dimensional, according to Lin et al., and they consider the smoothing spline $\gamma(t)$ as:

$$\gamma(t) = \sum_{k=1}^{m} \delta_k \varphi_k(t) + \sum_{l=1}^{r} a_l R(t, t_l^0),$$

where m$\geq$ 1is an integer, $t^0 = (t_1^0, \cdots, t_r^0)$T is the (r×1) vector of orderd, and r is the total number events. Assume $0 < t_1^0 < \cdots < t_r^0 < 1$. Where δ_k and a_l are

constant, $\{\varphi_k(t)\}_{k=1}^m$ is a basis for the space of (m-1)th-order polynomials, and

$R(t,s)=\int_0^1 (t-u)_+^{m-1}(s-u)_+^{m-1}/\{(m-1)!\}^2 du$, where $x_+=x$ if $x>0$ and 0

otherwise. $\delta=((\delta_{1,}\cdots\delta_{m,})^T$, $a=((a_{1,}a_{2.}\cdots a_{r,})^T$,

where γ is the corresponding vector of $\gamma(t)$ evaluated at each element of t^0.

$$\gamma = H\delta + \sigma a$$

$$H=\begin{bmatrix} \varphi_1(t_1^0) & \cdots & \varphi_m(t_1^0) \\ \vdots & \ddots & \vdots \\ \varphi_m(t_r^0) & \cdots & \varphi_m(t_r^0) \end{bmatrix} \qquad \sigma=\begin{bmatrix} R(t_1^0,t_1^0) & \cdots & R(t_1^0,t_r^0) \\ \vdots & \ddots & \vdots \\ R(t_r^0,t_1^0) & \cdots & R(t_r^0,t_r^0) \end{bmatrix}$$

In our analysis, we consider m=1, then R(t,s)=min(t,s)

$$H = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{r\times 1}, \qquad \sigma = \begin{bmatrix} t_1^0 & t_1^0 & t_1^0 & \cdots & t_1^0 \\ t_1^0 & t_2^0 & t_2^0 & \cdots & t_2^0 \\ t_1^0 & t_2^0 & t_3^0 & \cdots & t_3^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^0 & t_2^0 & t_3^0 & \cdots & t_r^0 \end{bmatrix}$$

The test statistic is T= U_τ { $\widehat{\beta}$, γ ($\widehat{\delta}$,0)}/K~ $\chi^2_{v,1-a}$ where

K=tr$\{(wvw'\sigma)^2\}$/tr(wvw'σ), v= tr$\{(wvw'\sigma)\}^2$/tr$\{(wvw'\sigma)^2\}$

$$U_\tau\{\widehat{\beta}, \gamma(\widehat{\delta},0)\}=\hat{r}_{ip}{}' * \sigma * \hat{r}_{ip}$$

$\hat{r}_{ip}$ is the Schoenfeld residual corresponding to the covariate X_2.

$V = \begin{bmatrix} V_{\beta\beta}^* & V_{\beta\gamma}^* \\ V_{\gamma\beta}^* & V_{\gamma\gamma}^* \end{bmatrix}$ is the inverse matrix of variance-covariance matrix of the null

model.

$$V_{\gamma\gamma} = \frac{1}{r} * V_{\gamma\gamma}^* * diag(r)$$

$$V_{\gamma\beta} = V_{\gamma\beta}^* * H$$

$$V_{\beta\gamma} = V_{\gamma\beta}^* * H'$$

$$wvw' = V_{\gamma\gamma} - (V_{\gamma\beta}, V_{\gamma\gamma}H) \begin{bmatrix} V_{\beta\beta}^* & V_{\beta\gamma}H \\ H^T V_{\gamma\beta} & H^T V_{\gamma\gamma}H \end{bmatrix}^{-1} \begin{pmatrix} V_{\beta\gamma} \\ H^T V_{\gamma\gamma} \end{pmatrix}$$

The advantage of this test is that the test statistic is easy to calculate. We only need to fit the null model, which is easily implemented in SAS via "phreg" and in R via the function "coxph." In this case, the test can eliminate the possibility of misspecification of the time variables. Based on the test statistic, Schoenfeld residuals from the null model and corresponding time values are needed. In R, the matrix calculation can be performed easily, and in SAS, we can do the test using the "proc iml" function.

2.5 Martingale-based residuals test

Barlow and Prentice (1993) proposed martingale-based residuals to check the PH model assumption. The martingale residuals can be interpreted as the differences at time t between the observed and expected numbers of events for the ith subject [21].

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_{ii}(u) \exp(\hat{\beta}' X_i) \, d\hat{h}_0(u)$$

Here $\widehat{M}_i(t)$ is shorthand for $\widehat{M}_i(\infty)$, $N_i(t)$ is the number of events for the ith subject experienced over time t, and $\hat{h}_0$ is the cumulative baseline hazard.

Plotting the cumulative martingale residuals against follow-up time may check the violation of the PH assumption. However, these residual plots are not easily interpreted. Lin (1993) suggested diagnosing the Cox model by grouping the cumulative martingale residuals with respect to follow-up time.

In R, we could use the package "gof" (Klaus, 2015) and code for "cumres.coxph" to check the PH model assumption for right-censored data using the code：

fit.cox <- coxph(Surv(time,event) ~ z1 + z2, data= Leukemia)

system.time(data.gof <- cumres(fit.cox,R=1000))

In this package, the Kolmogorov-Smirnov test statistics of the different coordinates of the observed score process are calculated.

In SAS, the "assess" option under the "phreg" function can check the Cox regression model. The "assess" statement creates the plot, and the "resample" option calculates the p-value of a Kolmogorov-type supremum test based on 1,000 simulated residuals patterns.

For all methods described above, a significant p-value (p<0.05) implies the variable of interest is a time-dependent variable, which is not constant over time, so the PH assumption is violated. In the following simulations, we perform the above tests to check the PH model assumption and calculate power under a number of different settings.

CHAPTER 3 SIMULATION DESIGN AND RESULTS

We designed a simulation setting to compare the powers of the five methods of checking PH model assumption. All methods were performed in both R and SAS.

3.1 Data generation

In our simulation setting, we generate time-to-event data using an extension of the two-piece exponential method ([22], [23], [24]). We built the data under a simple condition: a binary fixed variable X_1 with coefficient β_1, which is constant over time, and a binary time-dependent variable, $X_2(t)$. We use $Y_1 \sim \text{Exp}(1)$ as the random time at which the time-dependent variable switches on, and $Y_2 \sim \text{Exp}(1)$ as the random failure time. T_0 is the time after Y_1 at which the PH assumption is violated and the hazard ratio changes. For the variable X_2, β_2 is the coefficient before time Y_1+T_0, and β_3 is the coefficient after time Y_1+T_0. So the cut-off time of variable X_2 is Y_1+T_0, and we generate a piecewise exponential random time variable T as the following rate:

$$
\lambda = \begin{cases}
r_0 = \exp(\beta_1 X_1) & \text{if } Y_2 < Y_1 \\
r_1(t) = \exp(\beta_1 X_1 + \beta_2 X_{2(t)}) & \text{if } Y_1 \leq Y_2 < Y_1 + t_0 \\
r_2(t) = \exp(\beta_1 X_1 + \beta_3 X_{2(t)}) & \text{if } Y_1 + t_0 \leq Y_2
\end{cases}
$$

Here we consider X1 to be a binary time-independent variable with the distribution X1~binom (0.5) and X2(t) to be a binary time-dependent variable with the distribution X2(t)~binom(0.5).

For the exponential density function $f_T(t) = \lambda e^{-\lambda t}$, we consider that the baseline hazard of the PH model is 1. Based on above formula, we classify all subjects into three conditions: $Y_2 < Y_1$ means the subjects failed before X_2 (t) was switched on, $Y_1 \leq Y_2 < Y_1 + t_0$ means the subjects failed between when $X_2(t)$ was switched on and before the hazard change, and $Y_1 + t_0 \leq Y_2$ means the subjects failed after $X_2(t)$ was switched on and after the PH assumption was violated. Based on the above conditions, we generate time T as:

$$
T = \begin{cases}
\dfrac{Y_1}{r_0} & \text{if } Y_2 < Y_1 \\[2ex]
Y_1 + \dfrac{Y_2 - Y_1}{r_1(t)} & \text{if } Y_1 \leq Y_2 < Y_1 + t_0 r_1(t) \\[2ex]
Y_1 + t_0 + \dfrac{Y_2 - Y_1 - t_0 r_1(t)}{r_2(t)} & \text{if } Y_1 + t_0 r_1(t) \leq Y_2
\end{cases}
$$

After generating time T, we create the event indicated, assuming all subjects failed on the study first. We then generate random right-censoring time by using a uniform distribution, applying censoring time by replacing time T, and updating the event indicated.

In our simulation studies, we consider that the PH model assumption was violated in three ways: the time-dependent variable effect disappeared ($\beta_3=0$), reversed ($\beta_3=-\beta_2$), or increased ($\beta_3=2\beta_2$) at time Y_1+t_0. An increasing effect means β_3 is doubled in magnitude compared to β_2, representing a strong effect. We use random t_0 with a uniform

(0,2) distribution, as well as fixed t_0. The sample size is 500. The simulation setting is

presented in Table 3.1:

Table 3.1 Simulation setting

β_2	0.1	0.5	1	1.75	2.5	3.5	4.5	
	-0.1	-0.5	-1	-1.75	-2	-2.5	-3	-3.5
β_3	0	$-\beta_2$	$2\beta_2$					
t_0 (years)	~Unif(0,2)	0.05	0.5	1				
n	500							
censoring	0%	20%						

We consider a very large range of pre-change time-dependent variable

coefficients from -3.5 to 5. For disappeared and reversed effects, β_2 has the same value.

For increasing effects, we consider the negative value of β_2 to be -1.75, as the power is

equal to 1. Finally, the sample size is fixed at 500 with 100 repetitions.

3.2 Simulation results

Figure 3.1 and Figure 3.2 have the same simulation setting, where the sample size

is 500 with random t_0~uniform (0, 2) and the censoring rate is 0%. Figure 3.1 was

generated using R, and Figure 3.2 was generated using SAS.
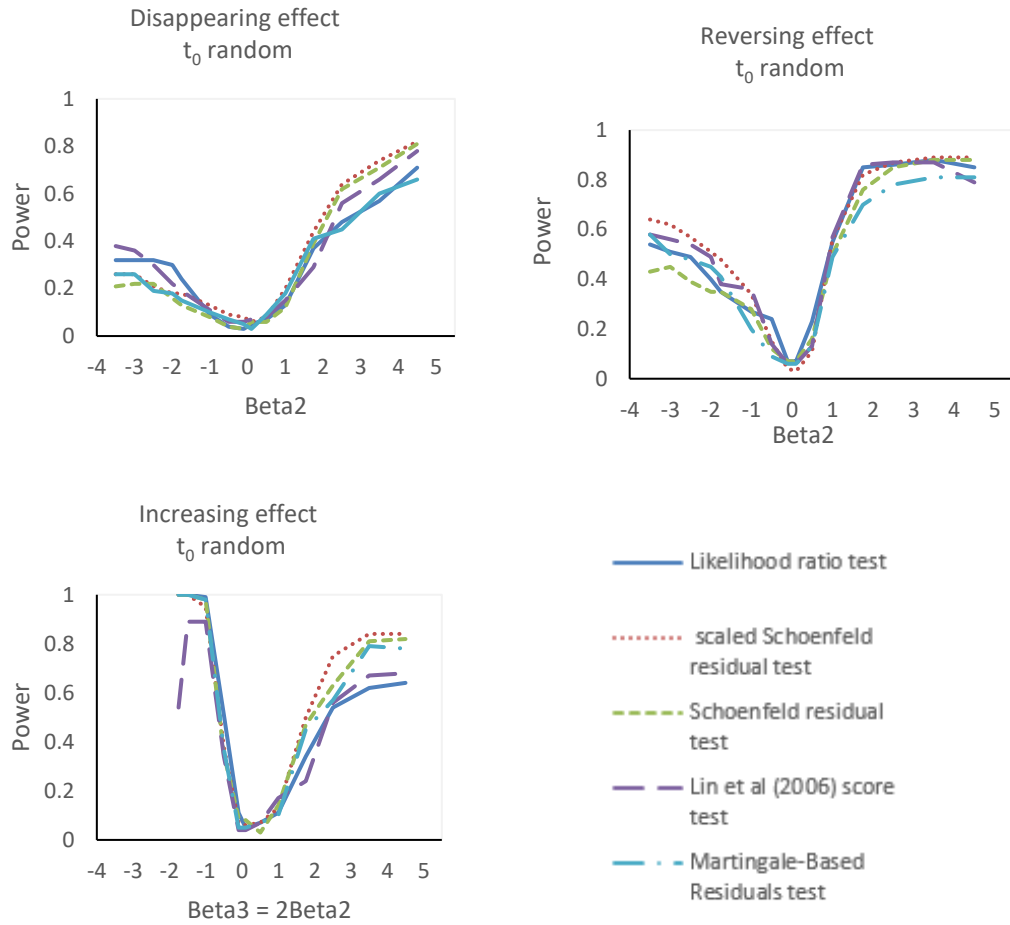
Figure 3.1 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0,0% censoring performed in R

For three different situations, all tests have the lowest power when β_2 is near 0, and the powers are all 90% or below when β_2 is positive. In the disappearing effect, the powers of all tests are less than 35% when β_2 is negative, and the Schoenfeld residuals test and scaled Schoenfeld residuals test have the highest power when β_2 is positive. In the reversing effect, the powers of all the tests are below 60% when β_2 is negative. For

positive values of β_2, except for the martingale residuals test, the powers of all other

tests are over 80% when $\beta_2 \geq 2.5$. When β_2 is between 2 and 3, the power peaks. For
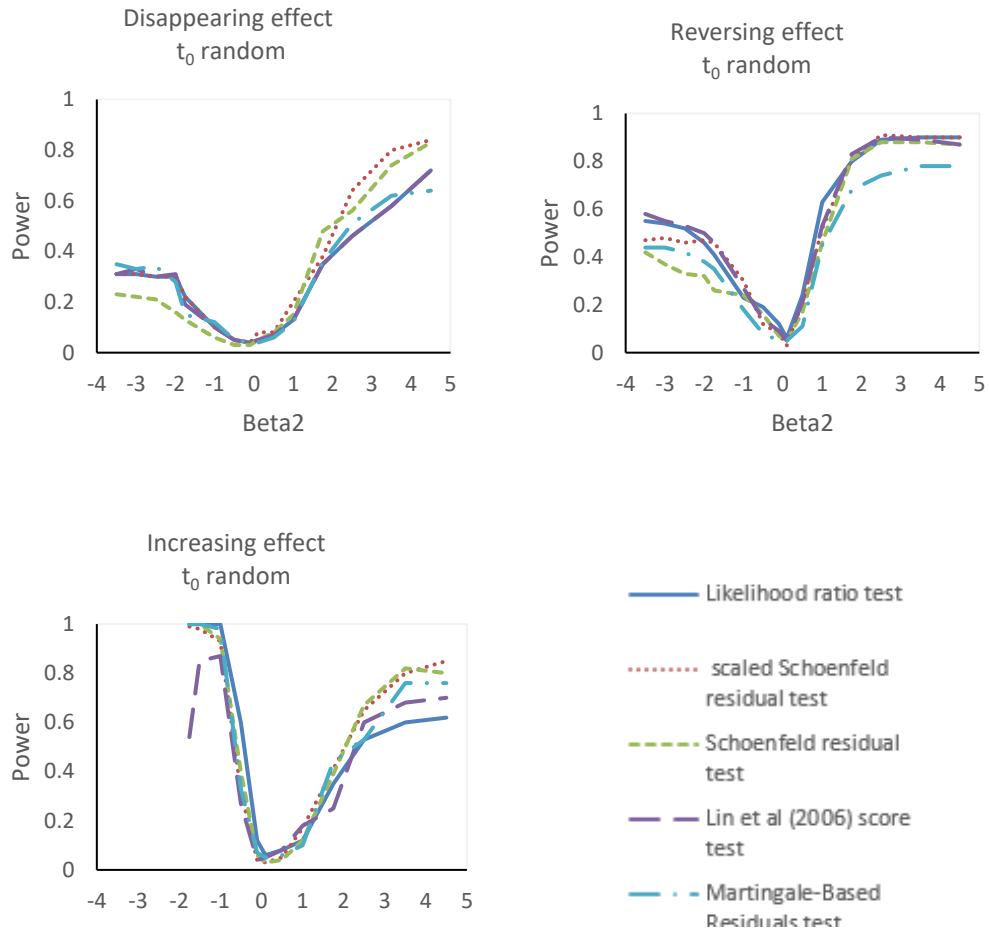
Figure 3.2 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 0% censoring performed in SAS

an increasing effect, all tests had high power for extreme values and when $\beta_2 \leq -1$

except the Lin et al. (2006) score test, and the power of the other tests can reach 1. But

the power of the Lin et al. (2006) score test decreases when β_2 is 1.75. The results from

R and SAS are similar for all tests.

We also perform the test with fixed t_0 and consider the disappearing effect. The

sample size is 500, with 0% censoring. All tests were performed in R and SAS with 100

repetitions.
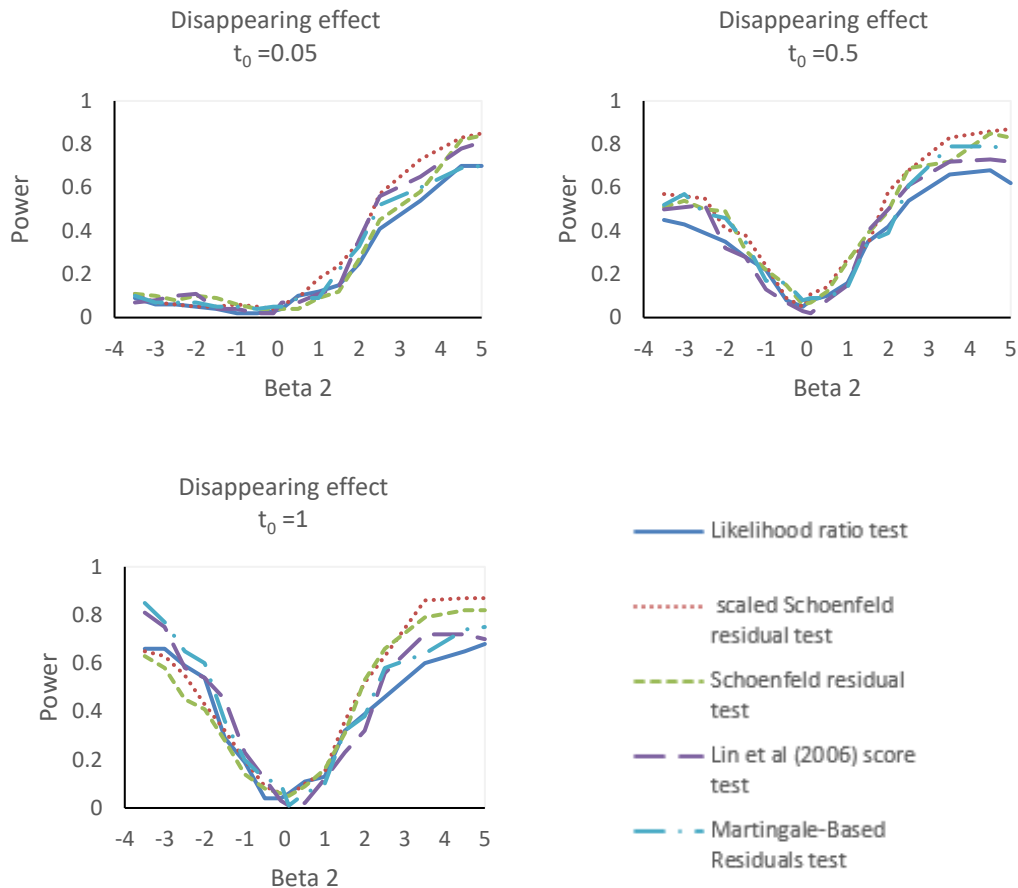
Figure 3.3 Power in the scenarios of disappearing effect, by fixed t_0, 0% censoring,
performed in R

Figure 3.3 and Figure 3.4 show the power of all tests with fixed t_0. When t_0 is 0.05

years, the powers of all tests are very low for negative β_2 and increase as β_2 increases for

positive values of β_2. For t_0 of 6 months, the powers of all tests for positive values of β_2

can reach a similar maximum power as when t_0 is 0.05 and show a similar pattern as when

the lowest t_0 is used. When β_2 is negative, the powers of all tests are 50% or below, which

is much higher than the power of all tests when t_0 is 0.05. When t_0 is 1 year, the powers of

all tests are the highest between three different t_0 when β_2 is negative.
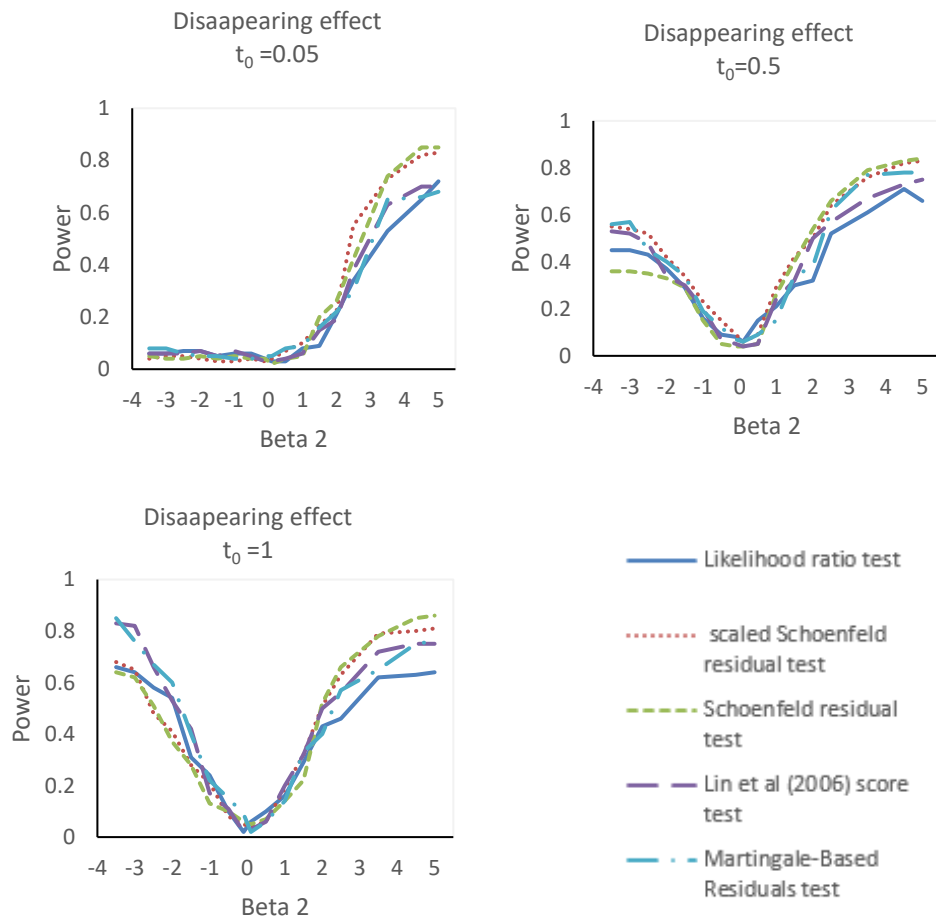
Figure 3.4 Power in the scenarios of disappearing effect, by fixed t_0, 0% censoring, performed in SAS

When $\beta_2 \leq -2.5$, the powers of all tests are over 60%. The highest power with a

positive β_2 doesn't differ much between t_0 at 0.05 and t_0 at 0.5. For all tests, the

Schoenfeld residuals and scaled Schoenfeld residuals tests have the highest power for

positive β_2. The likelihood ratio test has the lowest power. The results from SAS and R

are similar.

We also consider a sample size of 500 and 20% censoring, with random t_0.
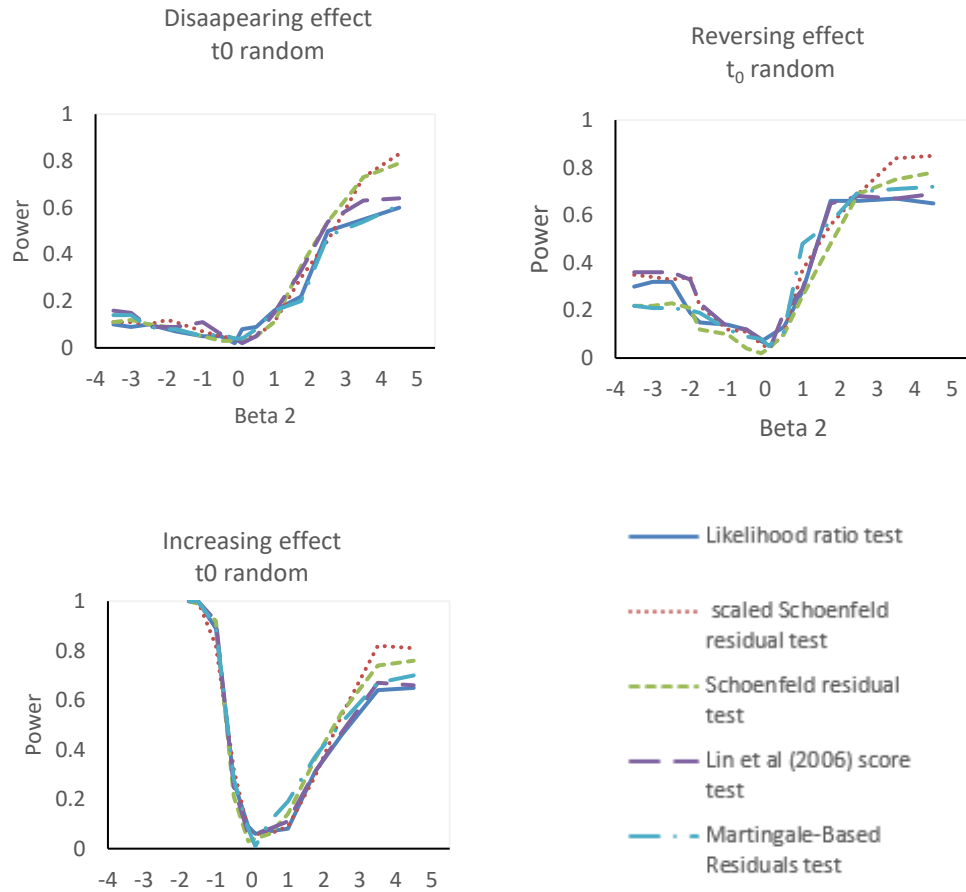
Figure 3.5 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 20% censoring performed in R

Figure 3.5 and Figure 3.6Figure 3.6 show the power results for three different

scenarios with random t_0 ~ uniform(0,2). The sample size is 500 and the censoring rate is

20%. The powers for all tests with positive β_2 under the three different scenarios are

slightly lower than the powers using a 0% censoring rate. With the disappearing scenario, the powers of all tests with a 20% censoring rate are much lower than the powers of all tests with a 0% censoring rate for negative β_2, and all powers are 15% or below. With the reversing effect, the powers of all tests with negative β_2 are higher than those with the disappearing effect.
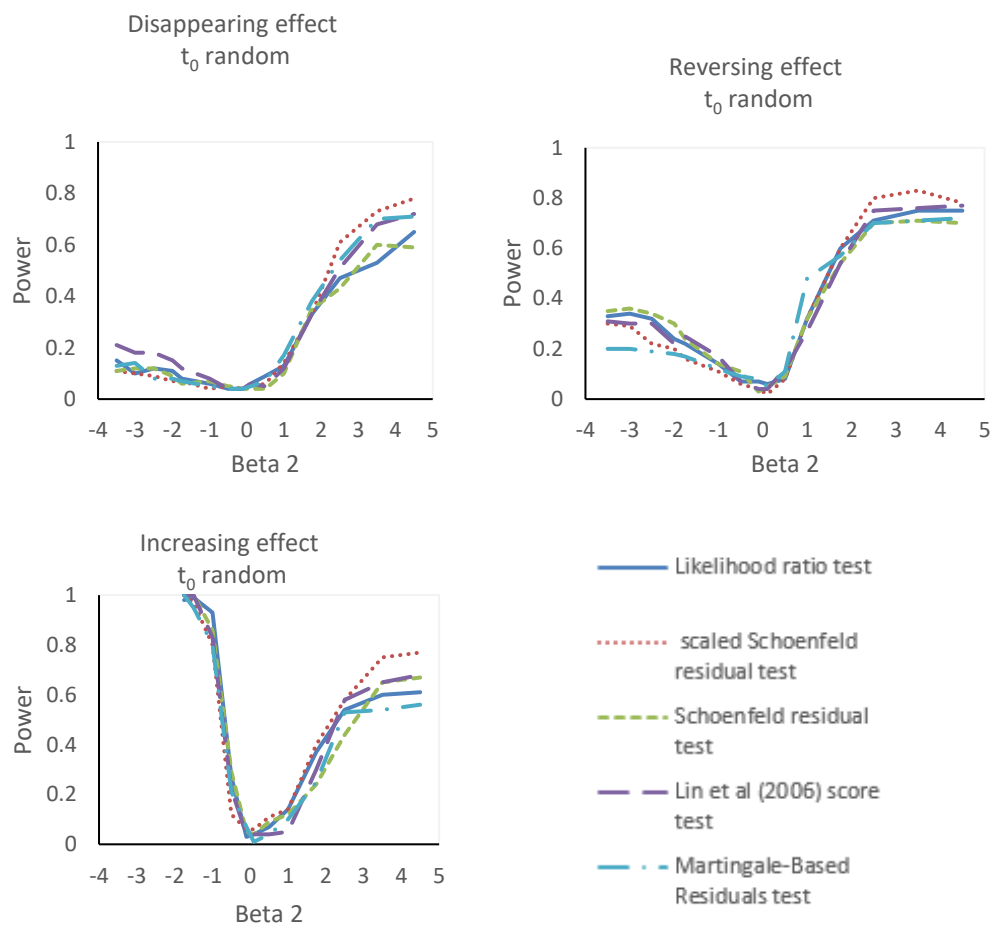
Figure 3.6 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 20% censoring performed in SAS

Compared to the simulation setting with 0% censoring rate, the powers of all tests are still lower than the powers with a 0% censoring rate. With the increasing effect, the powers of all tests are over 90% when $\beta_2 \leq 1$, and the maximum power can reach 1. Schoenfeld residuals have the highest power. The results from SAS and R are similar.

CHAPTER 4 REAL DATA ANALYSIS

Because it is very common that the effect of the variable of interest on the risk of disease may not be constant over time, it is important to check the PH assumption before using the PH model to interpret the association between risk factors and the disease of interest.

4.1 Leukemia remission data

The dataset from Freireich et al. (1963) contains matched-pairs clinical trial data. To compare the remission-free probability in the treatment and placebo groups, patients were followed until their leukemia relapsed or until the end of the study. There were 21 patients in each group. Five variables were collected for each patient: survival time (1 to 35 weeks), status (1 as event and 0 as censoring), sex (1 as male and 0 as female), log white blood cell count (log WBC), and treatment status (1 as treatment and 0 as placebo). We would like to investigate the treatment effects by adjusting for sex. The null PH model is:

$$h(t,X)= h_0(t) \ \exp(\beta_1 X_1 + \beta_2 X_2),$$

where X_1 is treatment status, and X_2 is sex. To use this model, we need to check whether both treatment and sex satisfy the PH assumption. The test based on the graphical tool in chapter one already indicates that treatment satisfies the PH assumption, and sex cannot satisfy the PH assumption. We conducted the test discussed earlier to further check whether sex satisfies the PH assumption using SAS and R.

The following tables list the p-values of tests for checking the PH assumption for the PH model with treatment and sex:

Table 4.1 P-values of five tests for treatment

Test	p-value (R)	p-value (SAS)
Likelihood ratio test	0.994	0.916
Schoenfeld residuals test	0.915	0.918
Scaled Schoenfeld residuals test	0.204	0.231
Lin et al. (2006) score test	0.684	0.702
Martingale-based residuals	0.362	0.337

The p-values of all five tests are greater than 0.05, and we conclude that the treatment variable satisfies the PH assumption.

Table 4.2 P-values of five tests for sex

Test	p-value (R)	p-value (SAS)
Likelihood ratio test	0.0017	0.0029
Schoenfeld residuals test	0.0005	0.009
Scaled Schoenfeld residuals test	0.0007	0.0012
Lin et al. (2006) score test	0.0015	0.0025
Martingale-based residuals	0	0.002

The p-values of all five tests are less than 0.05, and we conclude that the variable

sex cannot satisfy the PH assumption. That is, in practice, this variable cannot be adjusted

directly in the PH model.

4.2 Stanford heart transplant data

The Stanford heart transplant study began in 1967 and ended in 1974. There were

103 patients in this study, of which 69 received a heart transplant and 75 died during

follow-up. Crowley and Hu (1977) used the time-dependent PH model to analyze the

effect of heart transplant on patients' survival using this data set. However, the patients'

transplant status changes with time. Let us assume that a patient's transplant status is

denoted by 1 for transplant and 0 for no transplant. That is, for patients who didn't

receive a transplant, the transplant statuses are 0 at all the observed times, whereas for the

patients who did receive a transplant, their transplant statuses changed from 0 to 1 at the

transplant time and stayed at 1 after the transplant time. Therefore, Crowley and Hu

considered transplant status as a time-dependent variable in the PH model and concluded

that patients with heart disease can benefit from heart transplant. Therefore, we focus on

the model presented by Crowley and Hu as the following:

$$h(t,X(t)) = h_0(t) \ \exp(\beta_1 X_1(t) + \beta_2 X_2(t)),$$

where X_1 (t) is the patient's age at transplant (if the patient didn't have a

transplant, the variable is 0, otherwise the variable is the patient's age). X_2 (t) is

transplant status. Crowley and Hu calculated estimates of $\hat{\beta}_1$=0.057 and $\hat{\beta}_2$=-2.67. We obtained similar results, as shown below.

Table 4.3 Analysis of maximum likelihood estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
$\hat{\beta}_1$	1	0.05698	0.02245	6.4392	0.0112	1.059
$\hat{\beta}_2$	1	-2.72431	1.13932	5.7177	0.0168	0.066

We used a simulation study to assess the power of the five methods under similar conditions as the Stanford heart transplant data. Aitkin (1983) [25] used the piecewise method to re-analyze the Stanford heart transplant data and found that the hazard for patients after transplant declined after 60 days. The reversing effect is similar to the post-transplant condition after 60 days described by Aitkin. Therefore, we consider β_2 between 0 and 3, the cut-off point as 60 days, sample sizes of 100 with 25% censoring, and 1,000 replicates.

The following figures illustrate the simulation results of the data. We consider the hazard reversed after 60 days, and 67% patients received a transplant.
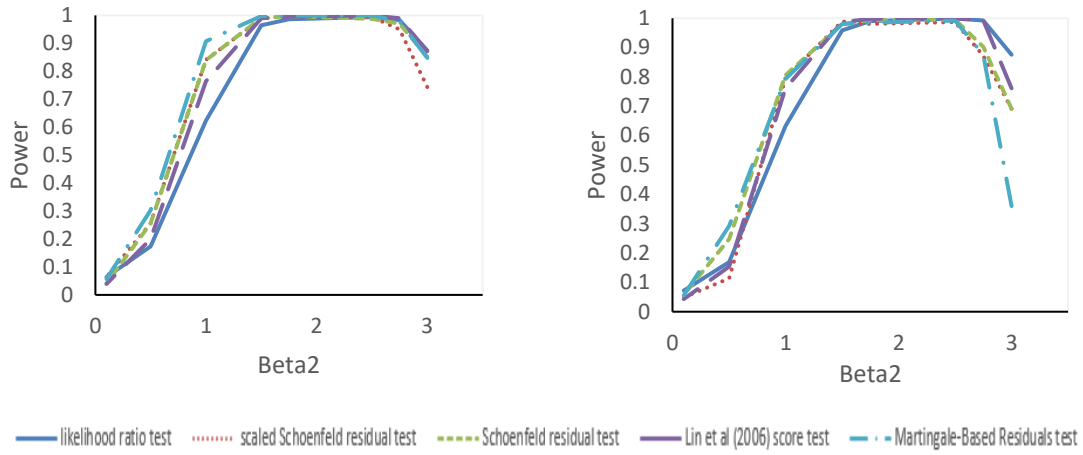
Figure 4.1 Simulation results for Stanford heart transplant data performed in R (left) and SAS (right)

For fixed t_0 at 60 days, the highest power achieved was when β_2 was between 1 and 2, and the power declined when $\beta_2 \geq 2.5$. The maximum powers for all tests are over 90%. Among the five tests, the likelihood ratio test has the lowest power. Therefore, all tests are powerful if the post-transplant hazard of death decreases after 60 days.

CHAPTER 5 DISCUSSION

In this thesis, we investigate five different methods to check the PH assumption, which was violated in three situations. When the hazard decreases with time, the power of the tests depends on the time at which the hazard changed and the scale of the hazard ratio change at this point. When the hazard increases with time before the change time, more patients are expected to have the event of interest, and thus the power increases. If there is a long time period before the hazard change, patients have a chance to survive a long time before the hazard change. For the reversing effect, the powers of all tests are greater than the powers in the disappearing effect because the magnitude of the hazard change in the reversing effect is larger than the hazard change in the disappearing effect. For the increasing effect, the magnitude of the hazard change is the same as for the disappearing effect, but the direction is different. By comparing the results from increasing effect to disappearing effect, we can conclude that the power of each test also dependent on the direction of the hazard change.

We conclude that the power of each test for checking the PH assumption in the presence of a time-dependent variable depends on the form of the hazard function. Specifically, for the function we considered in this thesis, the hazard function is a jump

function with constant hazards before and after a certain time point. The power is affected by the direction of the change, the magnitude of the hazard change, the time at which the hazard function changes, and the censoring rate of data.

For the specific simulation setting we considered, among the five methods, the most powerful tests were the Schoenfeld residuals test and the scale Schoenfeld residuals test. In most cases we considered, the likelihood ratio test had the lowest power. In the likelihood ratio test, the direct interaction of the time-dependent variable with time is considered, whereas in the Schoenfeld residuals test and the scale Schoenfeld residuals test, the interaction is reflected in the time-dependent variable and ranked time. The ranked time may be more sensitive to reflecting the time-dependent property. The score test (Lin et al. (2006)) doesn't need to specify the form of time, and the power is comparable to that of the Schoenfeld residuals test and the scale Schoenfeld residuals test and does not change much when the hazard function changes.

We investigated the power of each test in SAS and R and found only slight differences in the results between the two platforms.

When the variable of interest doesn't satisfy the PH assumption, there are several solutions. First, we could use the stratified Cox procedure for a signal predictor. The main idea here is to split the variable of interest into subgroups on the basis of a categorical variable [26]. Second, we could use a time-dependent PH model [5], which

uses interaction with a function of time for the covariate for which the assumption is not satisfied.

In my thesis, I used a piecewise exponential method to generate time-to-event data. In the future, we could generate survival data by using other distributions, such as the Weibull distribution, and examine the performance of the five methods for checking the PH assumption.

REFERENCES

[1] Nelson, Wayne. "Hazard plotting for incomplete failure data (Multiply censored data plotting on various type hazard papers for engineering information on time to failure distribution)." *Journal of Quality Technology* 1 (1969): 27-52.

[2] Nelson, Wayne. "Theory and applications of hazard plotting for censored failure data." *Technometrics* 14.4 (1972): 945-966.

[3] Cox, David R. "Regression models and life-tables." *Breakthroughs in statistics*. Springer New York, 1992. 527-541.

[4] Altman, Douglas G., and Bianca L. De Stavola. "Practical problems in fitting a proportional hazards model to data with udated measurements of the covariates." *Statistics in medicine* 13.4 (1994): 301-341.

[5] Fisher, Lloyd D., and Danyu Y. Lin. "Time-dependent covariates in the Cox proportional-hazards regression model." *Annual review of public health* 20.1 (1999): 145-157.

[6] Freireich, Emil J., et al. "The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy." *Blood* 21.6 (1963): 699-716.

[7] Clark, David A., et al. "Cardiac transplantation in man." *Ann. Intern. Med* 75 (1971): 15-21.

[8] Crowley, John, and Marie Hu. "Covariance analysis of heart transplant survival data." *Journal of the American Statistical Association* 72.357 (1977): 27-36.

[9] Lin, Jiang, Daowen Zhang, and Marie Davidian. "Smoothing Spline-Based Score Tests for Proportional Hazards Models." *Biometrics* 62.3 (2006): 803-812.

[10] Grambsch, Patricia M., and Terry M. Therneau. "Proportional hazards tests and diagnostics based on weighted residuals." *Biometrika* (1994): 515-526.

[11] Fleming, Thomas R., and David P. Harrington. *Counting processes and survival analysis*. Vol. 169. John Wiley & Sons, 2011.

[12] Allison, Ian. "Organizational factors shaping software process improvement in small-medium sized software teams: A multi-case analysis." *Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on the*. IEEE, 2010.

[13] Pinheiro, J., et al. "Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-105." *Vienna, Austria* (2012).

[14] Therneau, Terry. "A package for survival analysis in S. R package version 2.37-4." *See http://CRAN. R-project. org/package= survival* (2014).

[15] Hosmer Jr, David W., and Stanley Lemeshow. "Applied survival analysis: Regression modelling of time to event data (1999)." (1999): 561-562.

[16] Therneau, Terry, Cindy Crowson, and Elizabeth Atkinson. "Using time dependent covariates and time dependent coefficients in the cox model. "*Survival Vignettes* (2017).

[17] Kumar, Dhananjay, and Bengt Klefsjö. "Proportional hazards model: a review." *Reliability Engineering & System Safety* 44.2 (1994): 177-188.

[18] Schoenfeld, David. "Partial residuals for the proportional hazards regression model." *Biometrika* (1982): 239-241.

[19] Grambsch, Patricia M., and Terry M. Therneau. "Proportional hazards tests and diagnostics based on weighted residuals." *Biometrika* (1994): 515-526.

[20] Chen, Xiao. "Score test of proportionality assumption for cox models. "*Statistical Consulting Group UCLA* (2008).

[21] Lin, Danyu Y., Lee-Jen Wei, and Zhiliang Ying. "Checking the Cox model with cumulative sums of martingale-based residuals." *Biometrika* 80.3 (1993): 557-572. [22] Zhou, Mai. "Understanding the Cox regression models with time-change covariates." *The American Statistician* 55.2 (2001): 153-155.

[23] Leemis, Lawrence M., Li-Hsing Shih, and Kurt Reynertson. "Variate generation for accelerated life and proportional hazards models with time dependent covariates." *Statistics & probability letters* 10.4 (1990): 335-339.

[24] Grant, Shannon, Ying Qing Chen, and Susanne May. "Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates." *Lifetime data analysis* 20.3 (2014): 355-368.

[25] Aitkin, Murray, Nan Laird, and Brian Francis. "A reanalysis of the Stanford heart transplant data." *Journal of the American Statistical Association*78.382 (1983): 264-274.

[26] Hosmer, David W., Stanley Lemeshow, and Susanne May. "*Applied survival analysis.*" (2011).

APPENDIX A SIMULATION RESULTS

Table A.1 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 0% censoring performed in R. $\beta 3=0$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.32	0.26	0.21	0.38	0.26
-3	0.32	0.26	0.22	0.36	0.26
-2.5	0.32	0.21	0.22	0.3	0.19
-2	0.3	0.18	0.16	0.22	0.18
-1.75	0.24	0.18	0.13	0.19	0.15
-1	0.09	0.13	0.08	0.11	0.1
-0.5	0.04	0.09	0.04	0.06	0.07
-0.1	0.03	0.08	0.03	0.06	0.05
0.1	0.05	0.06	0.06	0.07	0.03
0.5	0.07	0.08	0.06	0.08	0.09
1	0.13	0.2	0.12	0.15	0.18
1.75	0.37	0.44	0.39	0.29	0.41
2.5	0.48	0.64	0.62	0.56	0.45
3.5	0.57	0.74	0.71	0.66	0.6
4.5	0.71	0.82	0.81	0.78	0.66

Table A.2 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 0% censoring performed in R. $\beta_3=-\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.54	0.64	0.43	0.58	0.58
-3	0.51	0.62	0.45	0.56	0.5
-2.5	0.49	0.57	0.39	0.54	0.48
-2	0.4	0.51	0.35	0.49	0.45
-1.5	0.35	0.48	0.35	0.38	0.41
-1	0.27	0.34	0.28	0.36	0.2
-0.5	0.24	0.15	0.12	0.14	0.09
-0.1	0.07	0.04	0.07	0.06	0.06
0.1	0.07	0.03	0.07	0.06	0.06
0.5	0.23	0.11	0.16	0.13	0.13
1	0.55	0.56	0.5	0.57	0.49
1.75	0.85	0.82	0.76	0.86	0.7
2.5	0.86	0.87	0.85	0.87	0.78
3.5	0.88	0.89	0.88	0.87	0.81
4.5	0.85	0.89	0.88	0.82	0.81

Table A.3 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 0% censoring performed in R. $\beta_3=2\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-1.75	1	1	1	0.54	1
-1.5	1	1	1	0.89	1
-1	0.99	0.95	0.98	0.89	0.98
-0.5	0.51	0.38	0.35	0.34	0.37
-0.1	0.11	0.07	0.09	0.04	0.05
0.1	0.05	0.06	0.08	0.04	0.05
0.5	0.07	0.07	0.03	0.07	0.07
1	0.11	0.13	0.14	0.17	0.1

1.75	0.34	0.5	0.47	0.24	0.45
2.5	0.54	0.75	0.63	0.56	0.57
3.5	0.62	0.84	0.81	0.67	0.79
4.5	0.64	0.84	0.82	0.68	0.78

Table A.4 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 0% censoring performed in SAS. $\beta_3=0$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.31	0.31	0.23	0.31	0.35
-3	0.31	0.31	0.22	0.33	0.33
-2.5	0.3	0.3	0.21	0.3	0.34
-2	0.3	0.3	0.16	0.31	0.28
-1.75	0.22	0.22	0.13	0.19	0.15
-1	0.1	0.1	0.06	0.1	0.12
-0.5	0.05	0.05	0.03	0.05	0.05
-0.1	0.04	0.04	0.03	0.04	0.04
0.1	0.05	0.08	0.05	0.05	0.04
0.5	0.07	0.08	0.08	0.07	0.06
1	0.13	0.2	0.15	0.13	0.12
1.75	0.35	0.38	0.48	0.35	0.36
2.5	0.46	0.64	0.56	0.46	0.51
3.5	0.58	0.8	0.74	0.58	0.62
4.5	0.72	0.84	0.83	0.72	0.64

Table A.5 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 0% censoring performed in SAS. $\beta_3=-\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.55	0.47	0.42	0.58	0.44
-3	0.54	0.48	0.37	0.55	0.44
-2.5	0.52	0.46	0.33	0.53	0.42
-2	0.46	0.47	0.32	0.5	0.38

-1.5	0.41	0.46	0.26	0.46	0.35
-1	0.23	0.3	0.24	0.27	0.18
-0.5	0.19	0.12	0.16	0.14	0.08
-0.1	0.12	0.08	0.07	0.09	0.05
0.1	0.06	0.03	0.05	0.05	0.05
0.5	0.24	0.21	0.17	0.21	0.11
1	0.63	0.52	0.46	0.53	0.45
1.75	0.8	0.81	0.81	0.83	0.68
2.5	0.89	0.91	0.88	0.9	0.74
3.5	0.9	0.9	0.88	0.89	0.78
4.5	0.9	0.9	0.87	0.87	0.78

Table A.6 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 0% censoring performed in SAS. $\beta_3=2\beta_2$.

	Power				
β_2	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-1.75	1	0.99	1	0.54	1
-1.5	1	0.98	1	0.85	1
-1	1	0.93	0.94	0.87	0.98
-0.5	0.6	0.32	0.4	0.27	0.33
-0.1	0.12	0.04	0.06	0.04	0.07
0.1	0.06	0.03	0.03	0.05	0.05
0.5	0.08	0.05	0.04	0.08	0.06
1	0.12	0.17	0.12	0.18	0.1
1.75	0.35	0.41	0.39	0.25	0.44
2.5	0.53	0.65	0.67	0.6	0.53
3.5	0.6	0.8	0.82	0.68	0.76
4.5	0.62	0.84	0.8	0.7	0.76

Table A.7 Power in the scenarios of disappearing effect, by fixed t0, 0% censoring, performed in R. $\beta_3=0$.

	Power				
β_2	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test

-3.5	0.09	0.11	0.11	0.07	0.1
-3	0.06	0.07	0.1	0.08	0.07
-2.5	0.06	0.06	0.08	0.1	0.07
-2	0.05	0.05	0.1	0.11	0.07
-1.5	0.04	0.05	0.09	0.04	0.05
-1	0.02	0.06	0.06	0.04	0.05
-0.5	0.02	0.05	0.04	0.02	0.04
-0.1	0.03	0.04	0.04	0.02	0.05
0.1	0.04	0.06	0.04	0.07	0.05
0.5	0.1	0.09	0.04	0.07	0.09
1	0.12	0.18	0.09	0.11	0.09
1.5	0.15	0.24	0.12	0.14	0.22
2	0.25	0.35	0.27	0.36	0.33
2.5	0.41	0.57	0.45	0.56	0.52
3.5	0.54	0.73	0.58	0.65	0.6
4.5	0.7	0.83	0.82	0.78	0.69

Table A.8 Power in the scenarios of disappearing effect, by fixed t0, 0% censoring, performed in R. $\beta_3 = -\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.45	0.57	0.51	0.5	0.52
-3	0.43	0.56	0.54	0.51	0.57
-2.5	0.39	0.55	0.5	0.52	0.48
-2	0.35	0.41	0.49	0.32	0.46
-1.5	0.28	0.38	0.31	0.28	0.35
-1	0.22	0.24	0.22	0.13	0.17
-0.5	0.08	0.09	0.15	0.07	0.16
-0.1	0.05	0.05	0.06	0.03	0.08
0.1	0.08	0.11	0.07	0.02	0.09
0.5	0.1	0.14	0.12	0.08	0.09
1	0.16	0.27	0.26	0.15	0.14
1.5	0.35	0.34	0.39	0.4	0.35
2	0.42	0.58	0.49	0.5	0.39
2.5	0.54	0.68	0.69	0.61	0.61
3.5	0.66	0.83	0.72	0.72	0.79

4.5	0.68	0.86	0.85	0.73	0.79

Table A.9 Power in the scenarios of disappearing effect, by fixed t0, 0% censoring, performed in R. $\beta_3=2\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.66	0.65	0.63	0.81	0.85
-3	0.66	0.63	0.58	0.75	0.77
-2.5	0.59	0.55	0.45	0.58	0.65
-2	0.54	0.43	0.41	0.54	0.6
-1.5	0.29	0.32	0.28	0.45	0.36
-1	0.19	0.2	0.14	0.23	0.2
-0.5	0.04	0.09	0.08	0.12	0.12
-0.1	0.04	0.06	0.06	0.03	0.1
0.1	0.06	0.05	0.05	0.01	0.01
0.5	0.11	0.1	0.09	0.02	0.06
1	0.13	0.15	0.16	0.12	0.1
1.5	0.32	0.36	0.31	0.23	0.32
2	0.39	0.52	0.53	0.32	0.38
2.5	0.46	0.63	0.66	0.56	0.58
3.5	0.6	0.86	0.79	0.72	0.64
4.5	0.65	0.87	0.82	0.72	0.74

Table A.10 Power in the scenarios of disappearing effect, by t0, 0% censoring, performed in SAS. $\beta_3=0$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.06	0.04	0.05	0.06	0.08
-3	0.06	0.05	0.04	0.06	0.08
-2.5	0.07	0.05	0.04	0.06	0.06
-2	0.07	0.04	0.05	0.07	0.05
-1.5	0.05	0.03	0.04	0.05	0.05
-1	0.06	0.03	0.05	0.07	0.04

-0.5	0.06	0.04	0.04	0.05	0.04
-0.1	0.04	0.03	0.04	0.03	0.05
0.1	0.03	0.04	0.02	0.03	0.05
0.5	0.03	0.07	0.04	0.04	0.08
1	0.08	0.1	0.05	0.06	0.09
1.5	0.09	0.17	0.2	0.15	0.16
2	0.21	0.22	0.26	0.19	0.22
2.5	0.34	0.55	0.42	0.38	0.3
3.5	0.53	0.73	0.74	0.63	0.65
4.5	0.65	0.82	0.85	0.7	0.66

Table A.11 Power in the scenarios of disappearing effect, by t_0, 0% censoring, performed in SAS. $\beta_3 = -\beta_2$.

	Power				
β_2	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.45	0.55	0.36	0.53	0.56
-3	0.45	0.54	0.36	0.52	0.57
-2.5	0.43	0.52	0.35	0.48	0.45
-2	0.37	0.42	0.33	0.34	0.4
-1.5	0.29	0.34	0.29	0.3	0.34
-1	0.16	0.23	0.15	0.2	0.19
-0.5	0.09	0.15	0.05	0.07	0.12
-0.1	0.08	0.09	0.04	0.05	0.07
0.1	0.06	0.06	0.04	0.04	0.06
0.5	0.15	0.09	0.05	0.05	0.09
1	0.21	0.29	0.26	0.24	0.15
1.5	0.3	0.42	0.4	0.32	0.34
2	0.32	0.5	0.54	0.5	0.38
2.5	0.52	0.64	0.66	0.57	0.62
3.5	0.61	0.76	0.79	0.67	0.77
4.5	0.71	0.82	0.83	0.73	0.78

Table A.12 Power in the scenarios of disappearing effect, by t_0, 0% censoring, performed in SAS. $\beta_3 = 2\beta_2$.

β_2	Power

	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.66	0.68	0.64	0.83	0.85
-3	0.64	0.65	0.62	0.82	0.76
-2.5	0.58	0.48	0.51	0.66	0.67
-2	0.54	0.41	0.37	0.53	0.6
-1.5	0.31	0.28	0.28	0.42	0.4
-1	0.24	0.21	0.13	0.17	0.22
-0.5	0.1	0.08	0.1	0.11	0.15
-0.1	0.02	0.05	0.06	0.02	0.09
0.1	0.06	0.04	0.05	0.03	0.02
0.5	0.1	0.06	0.07	0.06	0.06
1	0.16	0.19	0.14	0.2	0.14
1.5	0.29	0.32	0.22	0.32	0.33
2	0.43	0.51	0.52	0.5	0.4
2.5	0.46	0.63	0.66	0.56	0.57
3.5	0.62	0.79	0.78	0.72	0.65
4.5	0.63	0.8	0.85	0.75	0.75

Table A.13 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 20% censoring performed in R. $\beta_3=0$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.15	0.11	0.11	0.21	0.13
-3	0.1	0.1	0.12	0.18	0.14
-2.5	0.12	0.09	0.12	0.18	0.08
-2	0.11	0.07	0.09	0.15	0.08
-1.75	0.08	0.07	0.06	0.12	0.07
-1	0.06	0.04	0.07	0.08	0.05
-0.5	0.04	0.05	0.05	0.04	0.04
-0.1	0.04	0.04	0.04	0.04	0.04
0.1	0.06	0.05	0.04	0.06	0.05
0.5	0.09	0.06	0.04	0.05	0.07
1	0.13	0.14	0.1	0.12	0.17
1.75	0.33	0.33	0.35	0.33	0.38

2.5	0.47	0.61	0.43	0.51	0.54
3.5	0.53	0.73	0.6	0.68	0.7
4.5	0.65	0.78	0.59	0.72	0.71
4.75	0.65	0.8	0.6	0.75	0.72

Table A.14 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 20% censoring performed in R. $\beta_3=-\beta_2$.

	Power				
β_2	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.33	0.3	0.35	0.31	0.2
-3	0.34	0.29	0.36	0.3	0.2
-2.5	0.32	0.22	0.34	0.3	0.19
-2	0.24	0.2	0.3	0.22	0.18
-1.5	0.22	0.16	0.24	0.25	0.17
-1	0.14	0.11	0.14	0.17	0.12
-0.5	0.07	0.06	0.11	0.07	0.09
-0.1	0.07	0.04	0.03	0.04	0.08
0.1	0.06	0.02	0.05	0.04	0.05
0.5	0.08	0.08	0.09	0.12	0.11
1	0.32	0.32	0.32	0.27	0.48
1.75	0.6	0.6	0.54	0.54	0.57
2.5	0.71	0.8	0.7	0.75	0.7
3.5	0.75	0.83	0.71	0.76	0.71
4.5	0.75	0.78	0.7	0.77	0.72

Table A.15 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 20% censoring performed in R. $\beta_3=2\beta_2$.

	Power				
β_2	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-1.75	0.98	1	1	1	1
-1.5	0.99	0.97	1	1	0.95
-1	0.93	0.8	0.86	0.83	0.8

β₂					
-0.5	0.27	0.12	0.29	0.22	0.21
-0.1	0.03	0.06	0.05	0.04	0.07
0.1	0.04	0.06	0.04	0.04	0.01
0.5	0.07	0.11	0.09	0.04	0.04
1	0.14	0.14	0.12	0.05	0.1
1.75	0.37	0.4	0.24	0.3	0.25
2.5	0.54	0.58	0.44	0.58	0.53
3.5	0.6	0.75	0.65	0.65	0.54
4.5	0.61	0.77	0.67	0.68	0.56

Table A.16 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 20% censoring performed in SAS. $\beta_3=0$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-3.5	0.1	0.11	0.11	0.16	0.14
-3	0.09	0.11	0.12	0.15	0.14
-2.5	0.1	0.1	0.1	0.09	0.09
-2	0.08	0.12	0.09	0.09	0.08
-1.75	0.07	0.11	0.08	0.09	0.08
-1	0.05	0.07	0.05	0.11	0.05
-0.5	0.05	0.05	0.03	0.06	0.06
-0.1	0.02	0.04	0.03	0.04	0.04
0.1	0.08	0.04	0.03	0.02	0.04
0.5	0.09	0.05	0.05	0.05	0.08
1	0.16	0.11	0.11	0.15	0.16
1.75	0.22	0.3	0.35	0.33	0.2
2.5	0.5	0.46	0.54	0.54	0.48
3.5	0.55	0.73	0.73	0.63	0.54
4.5	0.6	0.83	0.79	0.64	0.61
4.75	0.63	0.84	0.81	0.65	0.63

Table A.17 Power in the scenarios of disappearing, increasing, and reversing effect, with random t_0, 20% censoring performed in SAS. $\beta_3=-\beta_2$.

β_2	Power				
	Likelihood	Scaled	Schoenfeld	Lin et al.	Martingale-base

	ratio test	Schoenfeld residuals test	residuals test	(2006) score test	d residuals test
-3.5	0.3	0.35	0.22	0.36	0.22
-3	0.32	0.34	0.22	0.36	0.21
-2.5	0.32	0.33	0.23	0.36	0.21
-2	0.19	0.34	0.21	0.33	0.2
-1.5	0.15	0.22	0.12	0.23	0.19
-1	0.14	0.12	0.1	0.13	0.12
-0.5	0.12	0.11	0.04	0.12	0.09
-0.1	0.07	0.06	0.02	0.08	0.08
0.1	0.09	0.04	0.04	0.03	0.05
0.5	0.13	0.1	0.1	0.19	0.11
1	0.28	0.37	0.26	0.29	0.48
1.75	0.66	0.56	0.48	0.65	0.57
2.5	0.66	0.69	0.69	0.68	0.67
3.5	0.67	0.84	0.75	0.67	0.71
4.5	0.65	0.85	0.78	0.69	0.72

Table A.18 Power in the scenarios of disappearing, increasing, and reversing effect, with random t0, 20% censoring performed in SAS. $\beta_3=2\beta_2$.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
-1.75	1	1	1	1	1
-1.5	1	1	0.99	1	1
-1	0.89	0.82	0.92	0.92	0.9
-0.5	0.27	0.32	0.21	0.25	0.28
-0.1	0.09	0.07	0.03	0.06	0.08
0.1	0.06	0.06	0.04	0.06	0.01
0.5	0.07	0.06	0.06	0.08	0.11
1	0.08	0.09	0.14	0.11	0.24
1.75	0.31	0.29	0.36	0.31	0.41
2.5	0.46	0.54	0.55	0.47	0.51
3.5	0.64	0.82	0.74	0.67	0.67
4.5	0.65	0.81	0.76	0.66	0.7

Table A.19 Simulation results for Stanford heart transplant data performed in R.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
3	0.874	0.744	0.853	0.872	0.848
2.75	0.983	0.95	0.97	0.99	0.978
2.5	0.994	0.994	0.988	1	1
1.75	0.987	0.998	0.997	0.996	0.999
1.5	0.965	0.989	0.991	0.99	0.998
1	0.623	0.841	0.84	0.764	0.908
0.5	0.174	0.256	0.256	0.2	0.303
0.1	0.064	0.064	0.041	0.039	0.057

Table A.20 Simulation results for Stanford heart transplant data performed in SAS.

β_2	Power				
	Likelihood ratio test	Scaled Schoenfeld residuals test	Schoenfeld residuals test	Lin et al. (2006) score test	Martingale-based residuals test
3	0.876	0.692	0.691	0.761	0.36
2.75	0.993	0.872	0.903	0.994	0.876
2.5	0.999	0.987	0.995	1	0.992
1.75	0.99	0.981	0.994	0.998	0.985
1.5	0.959	0.989	0.981	0.987	0.98
1	0.633	0.796	0.806	0.761	0.793
0.5	0.169	0.113	0.248	0.154	0.292
0.1	0.072	0.049	0.058	0.043	0.056