Assessing Proportionality of the Cox model in Survival Analysis

Abstract

Survival analysis studies the duration of time until the event of interest occurs. When the occurrence of the event of interest is not observed during the observation time, the data is described as censored. In survival analysis, Cox Proportional Hazards model is used to analyze censored data. The model works under the assumption that the hazard ratios are proportional with respect to time. However, in practice, this assumption is not always met. It is crucial to assess the proportionality because a violation may lead to potential bias in the estimates and undermine the validity of the results.

There are different techniques to test the proportionality assumption using survival curves, time interaction, and goodness-of-fit tests. A simulation study for censored data is carried out to illustrate these techniques in two cases when the proportionality is satisfied and when it is violated. A discussion of the usefulness of these approaches when assessing the appropriateness of the Cox model is represented. Estimates from the Cox models from the simulation study are used to demonstrate how biased the results are when the assumption is not met. A motivating study on the association between serum bilirubin and survival in Primary Biliary Cirrhosis is conducted with the Cox model, followed by a thorough assessment of the proportionality assumption.

Keywords: proportional hazard, Cox model, survival curves, time interaction, time-varying effect, goodness-of-fit

1. Introduction

1.1 Motivation

Motivation for Primary Biliary Cirrhosis (PBC) Primary Biliary Cirrhosis (PBC) is a relatively rare chronic liver disease that mainly affects women (Kaplan, 1996). When someone's immune system attacks the liver, it causes slow, progressive damage to the bile ducts in the liver. Over time, it will lead to fibrosis and cirrhosis of the liver with a lack of proper treatments. Complications of cirrhosis include swelling of the body or even liver cancer. PBC patients often experience common symptoms such as yellowing of the skin (jaundice), fatigue and loss of appetite. Some patients do not have any symptoms at all. In practice, meaningful insights on how biomarkers associate with the hazard of death contribute to better adjustment of personal care and improve treatment regimen. Researchers have studied possible treatments of PBC, which include medications and liver transplantation.

Motivation for survival models in PBC There have been advanced statistical methodologies motivated by the dynamic and complication of diseases such as PBC. One of them is the Cox Proportional Hazards model in Survival Analysis, which is used to study the survival of patients based on failure times and how it is associated with prognostic predictors, after accounting for drop-outs and loss to follow-up. Below is an introduction to survival analysis and the Cox model, where the statistical concepts are explained with detailed examples related to PBC and its common biomarkers - indicators of the progression and stage of diseases.

1.2 Survival Analysis

Survival Analysis consists of statistical procedures to analyze data sets where the outcome variable is the time until an event of interest occurs (Kleinbaum & Klein, 2010). In biomedical research, time can be years, weeks, or days from the initiation of the study until an occurrence of the event of interest, which can be disease incidence, relapse or death. The response variable time is a continuous outcome. The variable event is a binary outcome which takes on value 1 if the event of interest occurs and value 0 otherwise. In an example of Primary Biliary Cirrhosis (PBC), we are interested in studying the association between serum bilirubin and overall survival. The event of interest is failure and the outcome is time until death occurs.

1.3 Censoring

The most distinguishing feature of survival analysis is censoring, which occurs when information on survival time is incomplete (Leung et al. 1997). In other words, the event time is not observed. Possible reasons for censoring are loss to follow-ups, patients withdrawing from the study, or events not taking place during study time. There are different types of censoring:

- Right censoring: event of interest occurs after a certain time point: a PBC patient was alive at the study termination or lost to follow-up during the study

- Left censoring: event of interest occurs before a certain time point: a person was followed up until they became HIV positive. The exact time of their first exposure to the virus is unknown. It might have happened before their first positive test was recorded.
- Interval - censoring: event of interest occurs between a known time interval: an HIV patient tested positive for AIDS. The patient might have developed the disease at some point during their pre-last and last doctor visits.

Censoring is also classified based on the information it carries: informative and non-informative. The former indicates that reasons for censoring are related to the prognosis of the patients while the latter means that censored data does not provide any information on the health condition of the patients. When censoring is non-informative, the censoring distribution and event time distribution are independent. Thus, subjects remaining in the study are representative of censored ones not being observed in the study. In survival analysis, censoring is often assumed to be independent, random and non-informative (Kleinbaum & Klein, 2010).

id	time	status	event
1	1.095	dead	1
2	14.152	alive	0
3	2.771	dead	1
4	5.271	dead	1
5	4.121	transplanted	0

Table 1: Data Layout

Table 1 displays the information on PBC data. *Id* is the unique identification number of each patient. *time* is the age in years until the patient died or had a liver transplant. If a patient was alive, their time was 14 years, which was the study time. *event* is 1 if the patient died and is 0 for alive/transplanted or censored.

1.4 Survival Function

The survival function $S(t)$ is the probability that a person experiences an event after time *t*. In other words, $S(t)$ gives probability of a random variable time *T* exceeding a time point *t*.

$$S(t) = Pr(T > t) = 1 - F(t) = 1 - Pr(T \leq t)$$

It is a non-increasing function as *t* ranges from 0 to 1 with $t = 0$ corresponding to $S(t) = 1$. That is, at the beginning of the study, no subjects have experienced the event yet so their probability of surviving past time $t = 0$ is 1. On the other hand, if the study had no end point, eventually every subject would die. Thus, the probability of surviving past time $t = \infty$ is 0 theoretically. In practice, since the study time is finite and not all subjects experience the event, the graph of the

survival function looks like a step function and does not reach 0 at the end of the study. Survival function can be calculated with Kaplan-Meier method and used to directly describe survival.

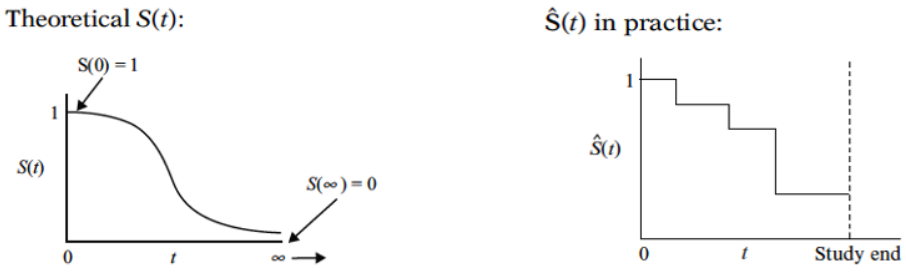

Figure 1: Survival Functions (These pictures are from Survival Analysis by Kleinbaum and Klein, 2010). The theoretical survival function is a smooth curve while the obtained survival function is a step function.

1.5 Hazard Function

In contrast to the survival function which focuses on surviving, the hazard function focuses on failing (having the event). The hazard function for an event is defined

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t+\Delta t \mid T \geq t)}{\Delta t}$$

It is a non-negative function that describes the instantaneous rate for occurrence of an event over the time interval $[t,\ t + \Delta t)$, on the condition that the individual has survived up to time *t*. In other words, it gives the instantaneous potential for having the event of interest at time *t*, given survival up to time *t* (Kleinbaum & Klein, 2010). Each subject has a unique hazard function.

The hazard function can be constant with respect to time and decreasing or increasing over time. The function $h(t)$ can be used to identify the model form. The relationship between survival function and hazard function of the Cox PH is

$$S(t) = exp\,(-\,(H(t))) = exp\,(-\,(\int_0^t h(s)\,ds))$$

or

$$H(t) = -\,log(S(t))$$

where $S(t)$ is the probability of survival up to time *t* and $H(t)$ is the cumulative hazard function. $H(t)$ is the sum of risks from time 0 to time *t*:

$$H(t) = \int_0^t h(s)\, ds$$

1.6 The Cox Proportional Hazards Model

Since the standard statistical models do not account for censoring, it motivates an advanced statistical model to handle censored data - the Cox Proportional Hazard (PH) model. The hazard function of the Cox PH model is given by Cox (Cox, 1972):

$$h(t, X) = h_o(t)\, exp\,(\sum_{i=1}^{p} \beta_i X_i)$$

$$h(t, X) = h_o(t)\, exp\,(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

where $h_o(t)$ is the unspecified baseline hazard function. β is a vector of coefficients and $X$is a vector of predictor variables. The Cox PH model works under the assumption that all patients share the same baseline hazard function $h_o(t)$ which only depends on time t, while the exponential part $exp\,(\sum_{i=1}^{p} \beta_i X_i)$ does not.

The covariates X in this Cox model are called time-independent covariates. The exponential part $exp\,(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$ is a constant, time-independent exponential function of linear regression of covariates unique to each subject. The Cox model measures multiplicative effects of the covariates on the hazard for an event. The interpretation is straightforward: one unit increase in X_i associates with a multiplicative change in the hazard by $exp\,(\beta_i)$. In a PBC example, if β_i denotes the coefficient of serum bilirubin $exp\,(\beta_i)$ = 1.05, we can say that each unit increase in the level of serum bilirubin corresponds to a 5% increase in the hazard of death.

From the hazard function, the Hazard-Ratio is constructed

$$HR = \frac{h(t, X^*)}{h(t, X)} = exp\,(\sum_{i=1}^{p} \beta_i (X_i - X_i^*))$$

where X^* is the set of covariates for subject i and $X$is the set of covariates for subject $j \neq i$. The Hazard Ratio (HR) is constant with respect to time because it does not involve time t. If X is categorical covariate sex in the PBC data set and HR = $\frac{h(t, male)}{h(t, female)}$ = 0.52, the interpretation is female patients have 48% lower hazard of death than males. In the case of a continuous covariate age and the HR = 0.08, it means that one year increase in age is associated with an 8% increase in the hazard of death.

In general, if the HR is bigger than 1, the hazard for the first group $h(t, X^*)$ is multiplied by $exp\,(\sum_{i=1}^{p} \beta_i(X_i - X_i^*))$ compared to the hazard for the second group $h(t, X)$. If the HR is smaller than 1, the opposite is true. Since HR does not involve time *t*, the hazard for one subject is proportional for any other subject at any time. This explains the proportionality assumption.

2. Proportionality Assumption Assessment

In practice, not all Cox PH models satisfy the proportionality assumption. Thus, it is important to assess this assumption because a violation might result in potential bias in the estimation. There are three common ways to assess the proportional hazards assumption. Those approaches are (Kleinbaum & Klein, 2010):

- Graphics with survival curves
- Including time-by-covariate interaction term
- Goodness-of-fit test

2.1 Graphics with survival curves

The first approach compares the log-log survival curves and compares the observed versus expected survival curves. The relationship between survival function and instantaneous risk function of Cox PH is

$$S(t) = exp\,(-\,(H(t))) = exp\,(-\,(\int_0^t h(s)\,ds))$$

The log-log transformation of the estimated survival curve is

$$log\,-\,log\,(S) = ln\,(-\,ln\,(S))$$

The difference between log-log survival curves of two different subjects is

$$ln\,(-\,ln\,S(t, X_1))\,-\,ln\,(-\,ln\,S(t, X_2))$$

$$= log(h_o(t))\,+\,(\sum_{i=1}^{p} \beta_i X_{1i})\,-\,log(h_o(t))\,-\,(\sum_{i=1}^{p} \beta_i X_{2i})$$

$$= exp\,(\sum_{i=1}^{p} \beta_i(X_{1i} - X_{2i}))$$

The difference between the two estimated log-log survival curves is linear. Thus, the two curves are approximately parallel if X are time-independent. Otherwise, the assumption is violated. To

assess PH assumption for continuous variables, one can stratify these variables into different categories and perform the graphical test.

Kaplan-Meier survival curves can also be used to detect any suspicious non-proportionality. The Kaplan-Meier (KM) survival probability at failure time t is estimated using the Law of Total Probability

$$S(t) = \underset{t}{S(t-1)}\ Pr(T^* > t \mid T^* > t-1)$$
$$= \prod_{i=1} Pr(T^* > t \mid T^* > t-1)$$

where $Pr(T^* > t \mid T^* > t-1) = \frac{r_i - d_i}{d_i}$ such that r_i denotes the number of subjects at risk at a unique time event t_i and d_i denotes the number of events t_i. The number of subjects at risk r_i include the subjects who have not yet experienced the event at time t or who are censored by time t. In Kaplan-Meier curves, the tick marks indicate censoring.

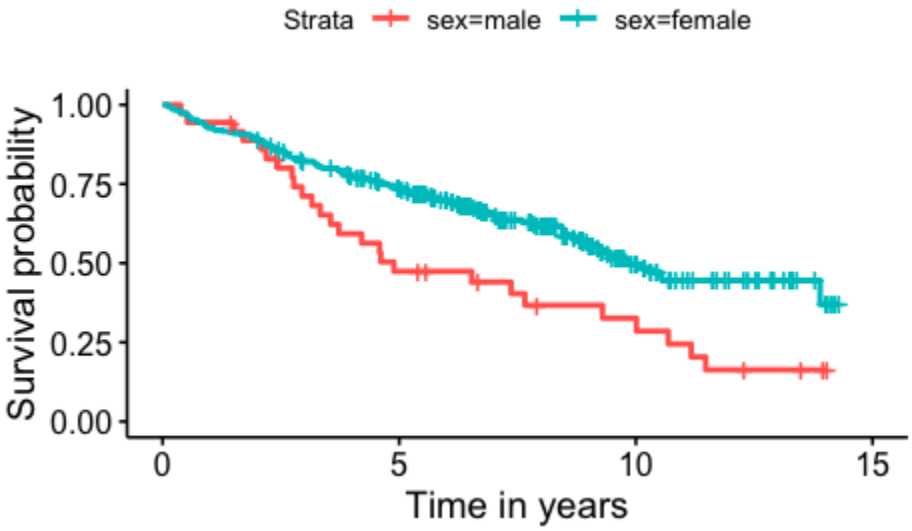

Figure 2: Kaplan-Meier survival curves for PBC data. In the plot, the tick marks indicate that the subject is censored. The survival probabilities for both male and female are 1 at time 0 and decrease over time. In general, females seemed to have a higher probability of survival than males. The median survival time was around 5 years for male patients while around 10 years for female patients.

Figure 2 shows the survival probabilities of PBC patients in different sex groups with no other covariates accounted. If the two curves cross, then the proportionality may be violated (Bouliotis & Billingham, 2011). However, small sample sizes may yield estimation error. In practice, the log-log survival curves are considered to be more robust.

2.2 Time-by-covariate interactions

A time-by-covariate interaction occurs when the effect of an explanatory variable on survival changes with time (Hess, 1995). We can test the statistical significance of the time-by-covariate interaction in the Cox model to detect the proportionality violation. If X_1 is being suspected of having time-varying effect on the hazard, the added interaction term is

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

In the equation, the term $f(t)$ is a function of time and can take the form of linear, logarithmic, and exponential functions (Hess, 1995). In this project, the function $f(t)$ is simply *t*, so the interaction term is $X_1 t$. The model is being assessed one-at-a-time by testing for significance of the product term to evaluate departure from non-proportionality. In other words, if bilirubin is being assessed, then the Cox model will have an interaction term (bilirubin x *t*). If the interaction term is significant, bilirubin has a time-varying effect and the proportionality is violated for bilirubin.

2.3 Goodness-of-fit test

The goodness-of-fit (GOF) technique tests the correlation between failure times and Schoenfeld residuals. Schoenfeld residuals are defined for every subject who has an event, one for each of predictors in the model (Kleinbaum & Klein, 2010). For example, if a Cox PH model has *n* covariates, there are *n* Schoenfeld residuals for each subject that has an event. They are the difference between the observed and weighted average values of suspiciously time-dependent covariates. In the PBC example, the Schoenfeld residual for bilirubin for ith subject that has the event is calculated

Schoenfeld residual = Observed bilirubin - Weighted Average bilirubin

where the weighted average of bilirubin is the values of the other subjects still at risk at time *t* and the weight is the hazard of each of the subjects. The correlation between Schoenfeld residuals and failure time is then tested. If Schoenfeld residuals are correlated with failure times, the proportionality assumption is violated; otherwise, the assumption is satisfied.

In the goodness-of-fit test, the null hypothesis states that there is not a correlation between them. Thus, an insignificant p-value from this test indicates that the proportional hazard is met. Each of the p-values checks the assumption for its corresponding covariate on the condition that the proportionality is satisfied for the other covariates in the model (Kleinbaum & Klein, 2010).

3. Assessing proportionality on simulated data

To assess the appropriateness of the Cox model and the effect of proportionality on the estimates, a simulation study is conducted with pre-specified parameters. There are two cases: one where the event rates are not too high so that the hazard ratio is close to being proportional and one

where the rates are high so that assumption is not fully met. I will do a thorough assessment of the proportionality with three different approaches and compare the estimates from the Cox models in both cases.

3.1 Simulation set-up

The simulation code is adapted from SAS and R: Data management, statistical analysis, and graphics (Kleinman & Horton, 2014) and the Cox analysis is run using the package *survival* (Therneau & Lumley, 2015). To simulate data from a Cox proportional hazards model, hazard functions for both time-to-event and time-to-censoring are modeled with random Weibull distributions. The probability density function of the Weibull distribution is

$$f(x) = \frac{\gamma}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{(\gamma-1)} = exp\,(-\,((x-\mu)/\alpha^{\gamma} \text{ where } x \geq \mu;\ \gamma, \alpha > 0$$

γ is the shape parameter and α is the scale parameter. In this simulation study, the baseline hazard is constant because in the Weibull distribution, shape $\gamma = 1$ The Weibull distribution can be increasing ($\gamma > 1$) and decreasing ($\gamma < 1$) depending on the value of γ.

The two cases of proportionality and non-proportionality share the following properties:

- The two covariates X_1 and X_2 follow normal distributions
- The baseline hazard: $\lambda_T = 0.002$
- The sample size: n = 10000
- The Weibull distribution of time-to-censoring: $f(x) = \frac{1}{\alpha}$
  Weibull distribution with shape $\gamma = 1$, scale $\alpha = \lambda_C$
- The Weibull distribution of time-to-event: $f(x) = \frac{1}{\alpha}$
  Weibull distribution with shape $\gamma = 1$, scale $\alpha = \lambda_T - exp\,(\beta_1 X_1 - \beta_2 X_2)$

The two cases have different hazards of censoring λ_C, and therefore different Weibull distributions of time-to-censoring. The intuition here is that if the event rate is not too high, the proportionality is presumably satisfied. If the event rate is high, the proportionality cannot hold in the long run. Thus:

- The hazard of censoring for proportional case: $\lambda_C = 0.004$
- The hazard of censoring for non-proportional case: $\lambda_C = 0.04$

The distributions of time-to-event are different because each study has different true coefficient values to satisfy the pre-specified proportionality assumptions.

- Proportional case: $\beta_1 = 2$; $\beta_2 = -1$
- Non-proportionality case: $\beta_1 = 4$; $\beta_2 = -1$

I run the simulation 100 times to get the distribution of the p-values for the time-by-covariate interactions and the Goodness-of-fit test. The summary table of averaged p-values is also represented. However, it is not feasible to plot all Kaplan-Meier and log-log survival curves from the simulation. Thus, in each simulation case, I will provide one Kaplan-Meier curve and log-log survival for each of covariates$X_1$and X_2. Since the covariates are continuous, they are stratified into 3 different groups: high, middle and low for visualizations.

3.2 Simulation results

3.2.1 Proportionality condition is presumably satisfied

- Event rate is about 60%
- The hazard of censoring: $\lambda_C = 0.004$
- True parameters: $\beta_1 = 2$; $\beta_2 = -1$

Visualization with Kaplan-Meier survival curves In Figure 3, it seems that the Kaplan-Meier survival curves do not cross, so the hazard ratios seem to be proportional for $X_1$and X_2.

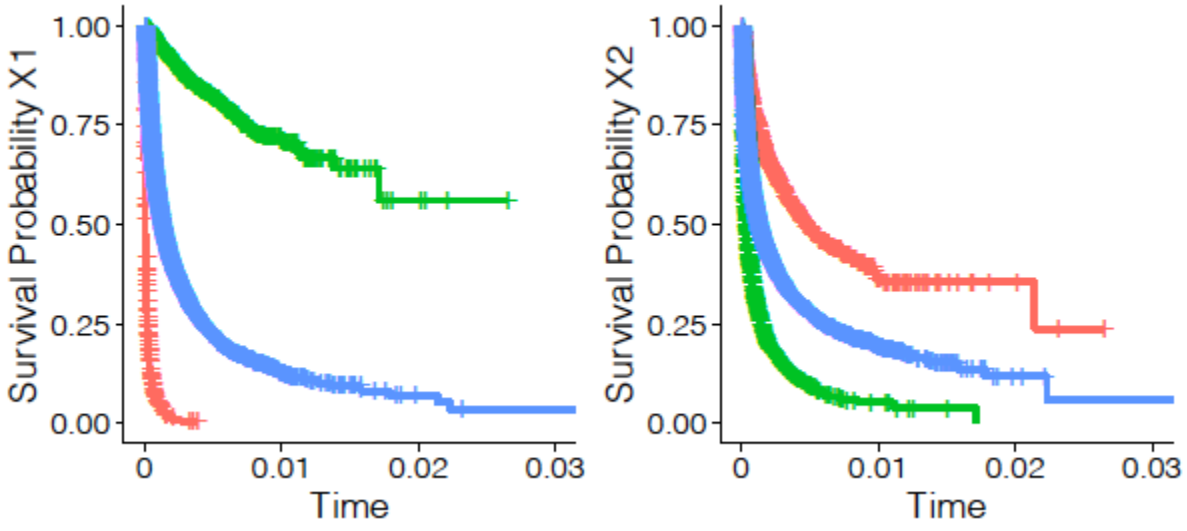

Figure 3: Kaplan-Meier survival curves for covariates X1 and X2 in the proportional case. In each plot, the curves represent the survival probability for groups high (red), medium (blue), and low (green) of the corresponding covariate.

Log-log survival curves In Figure 4, it seems that the log-log survival curves are approximately parallel for covariates $X_1$and X_2.

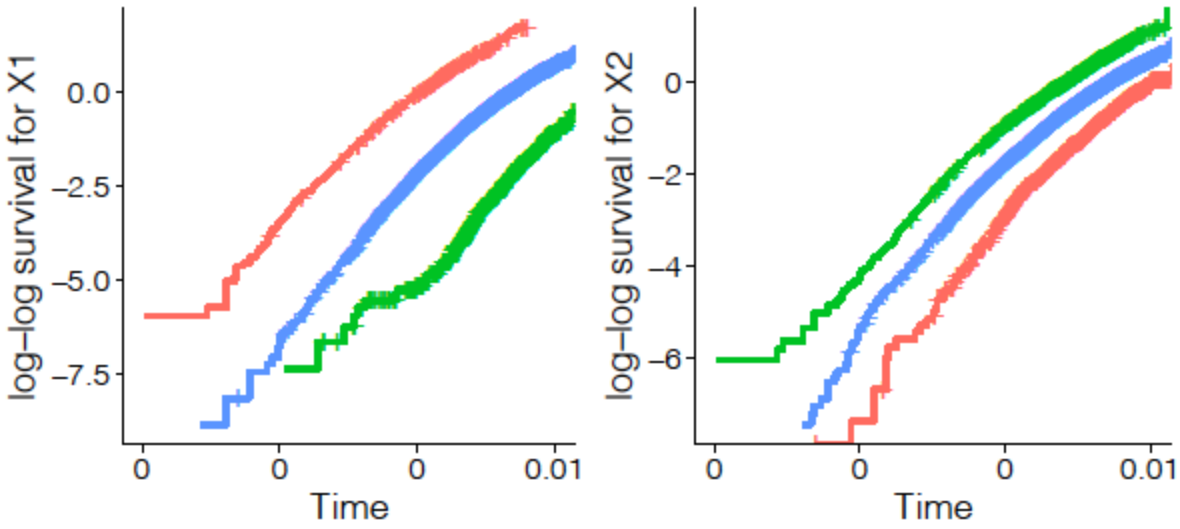

Figure 4: Log-log survival curves for covariates X1 and X2 in the proportional case. In each plot, the curves represent the log-log transformation of the estimated survival curves for groups high (red), medium (blue), and low (green) of the corresponding covariate.

Interaction terms

In Figure 5, the majority of the p-values for x1:time interaction term are close to 0, with a few exceptions; the largest p-value for x1:time interaction term is 0.3. The distribution for p-values for x2:time interaction term vary from 0 to 1. Though many p-values for x2:time interaction terms are larger than the significance threshold of 0.05, there are quite a few p-values close to 0.

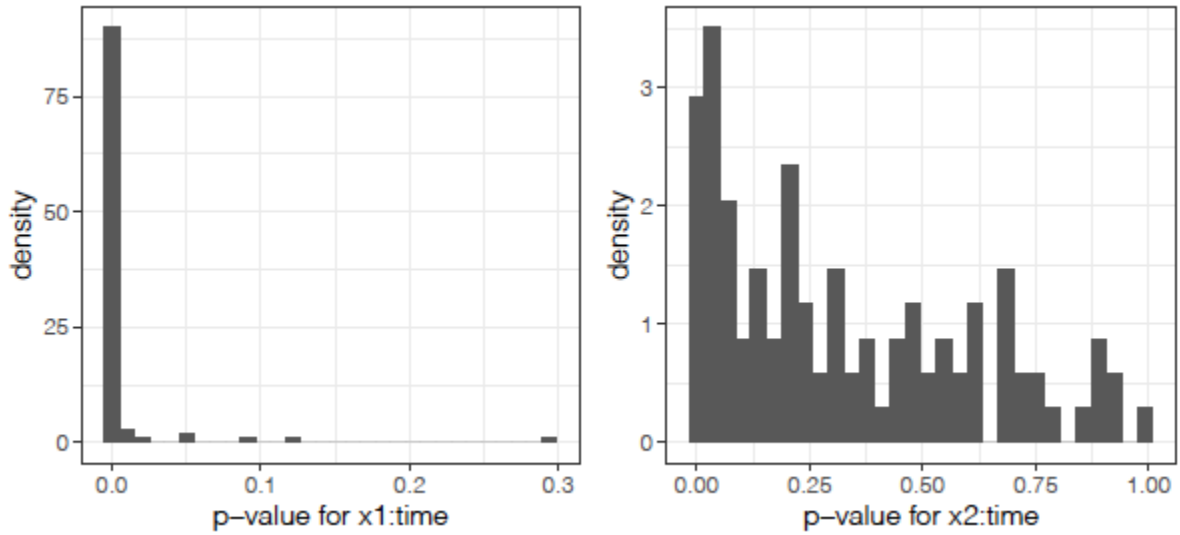

Figure 5: Distribution of p-values for time by covariate interaction terms in the proportionality case

In Table 2, the averaged p-value for x1:time interaction is 0.007 while the averaged p-value for x2:time interaction is 0.324. Thus, I will conclude that generally X_1 has a time-varying effect while X_2 is not time-varying.

Interaction	Averaged p-value
x1:time	0.007
x2:time	0.324

Table 2: Averaged p-value for covariate by time interaction term across 100 simulations in the proportionality case

Goodness-of-fit test with Schoenfeld residuals

In Figure 6, the majority of the p-values for X_1 and X_2 from the Goodness-of-fit test are larger than the significance threshold of 0.05.

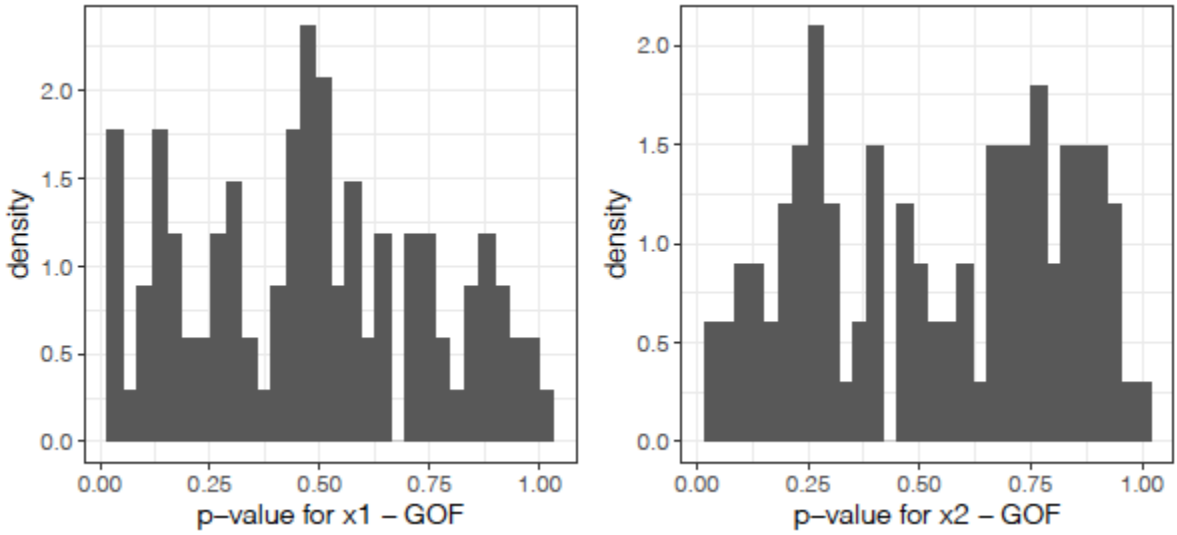

Figure 6: Distribution of p-values from the Goodness-of-fit test for covariates in the proportionality case

Table 3 shows the averaged results of goodness-of-fit test for the Cox model in the proportional hazards case. Since the averaged p.values for both covariates X_1 and X_2 are not statistically insignificant, I conclude that in general, covariates X_1 and X_2 satisfy the proportionality. The averaged estimates across 100 simulations are close to the true parameters.

Covariate	True parameter	Estimate	Averaged p-value

X_1	2	1.999	0.479
X_2	-1	-0.999	0.536

Table 3: Averaged p-value from the Goodness-of-fit test for covariates across 100 simulations in the proportionality case

3.2.2 Proportionality condition is presumably not satisfied

- Event rate is around 75%
- The hazard of censoring: $\lambda_C = 0.04$
- True parameters: $\beta_1 = 4$; $\beta_2 = -1$

Visualization with Kaplan-Meier survival curves

In Figure 7, it seems that the proportionality is not satisfied for X_1 because the red (high X_1) and the blue (middle X_1) curves seem to cross each other, a little bit. The red (high X_1) curve in group X_1 has very low survival probabilities. The proportionality for X_2 appears to be met. Overall, the Kaplan-Meier curves are not clear enough to help assess the proportionality assumption.

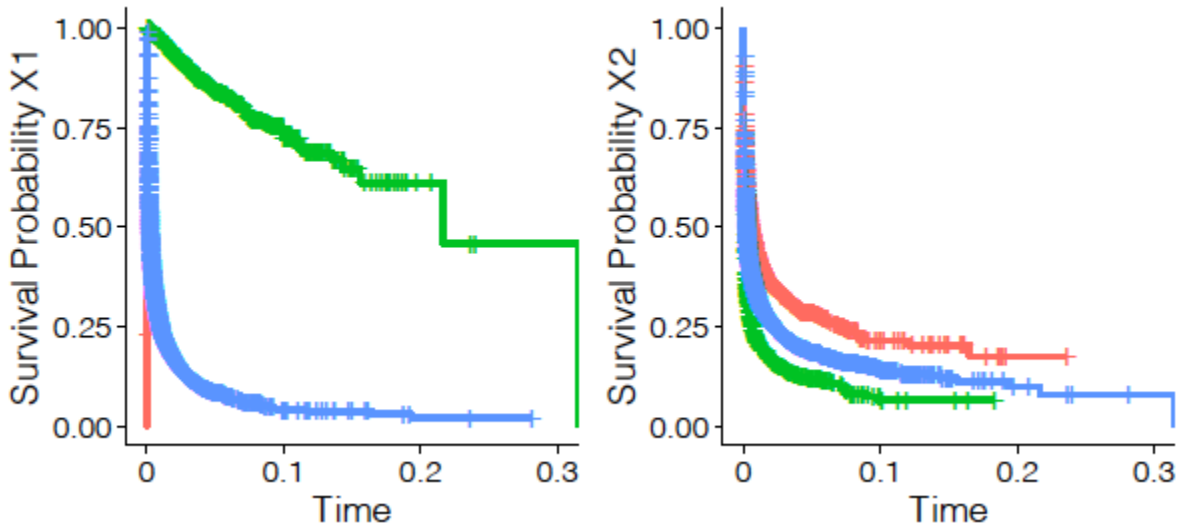

Figure 7: Kaplan-Meier survival curves for covariates X1 and X2 in the non-proportional case. In each plot, the curves represent the survival probability for groups high (red), medium (blue), and low (green) of the corresponding covariate.

Log-log survival curves In Figure 8, it seems that the survival curves are parallel for covariate X_2. It does not appear to be parallel for covariateX_1.

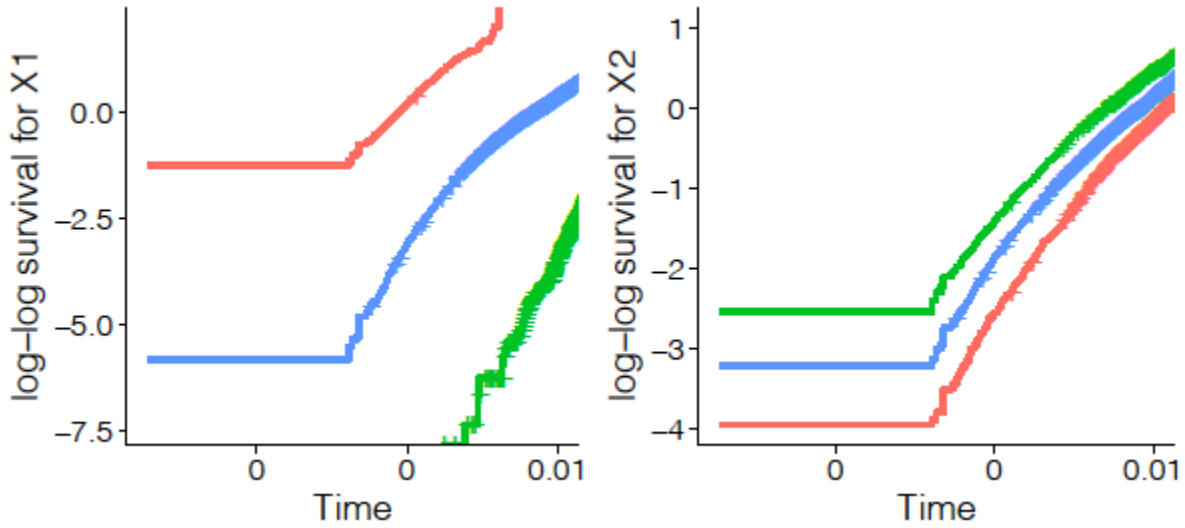

Figure 8: Log-log survival curves for covariates X1 and X2 in the non-proportional case. In each plot, the curves represent the log-log transformation of the estimated survival curves for groups high (red), medium (blue), and low (green) of the corresponding covariate.

Interaction terms

In Figure 9, the majority of the p-values for x1:time interaction term are close to 0, with a few exceptions; the largest p-values for x1:time interaction is around 0.33. The distribution of p-values for x2:time interaction terms vary from 0 to 1; most of them are larger than the significance threshold of 0.05.

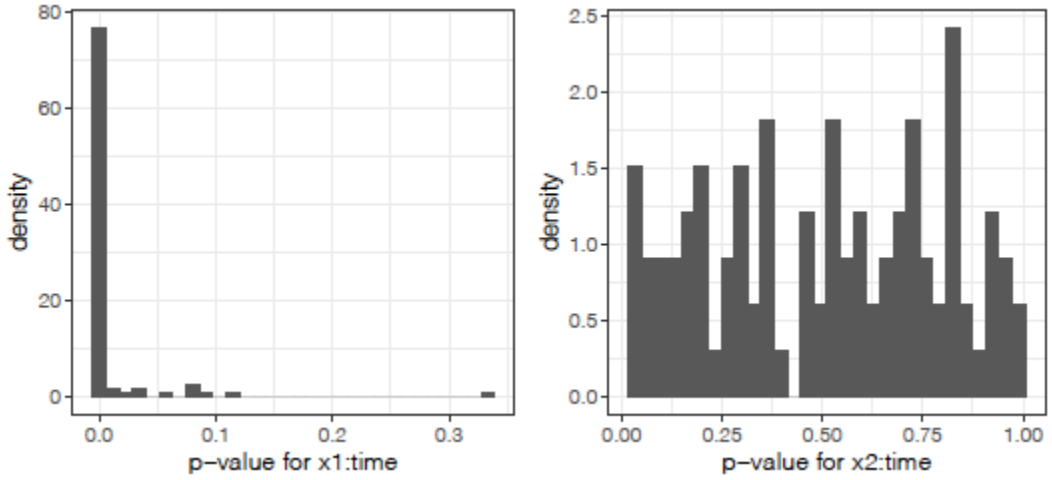

Figure 9: Distribution of p-values for covariate by time interaction terms in the non-proportionality case

In Table 4, the averaged p-value for x1:time interaction is 0.01 while the averaged p-value for x2:time interaction is 0.505. Thus, I will conclude that generally X_1 has a time-varying effect while X_2 is not time-varying.

Interaction	Averaged p-value
x1:time	0.010
x2:time	0.505

Table 4: Averaged p-value for covariate by time interaction term across 100 simulations in the non-proportionality case

Goodness-of-fit test with Schoenfeld residuals

In Figure 10, the majority of the p-values for both covariates X_1 and X_2 from the Goodness-of-fit test are close to 0, especially for X_1. The p-values for X_2 vary from 0 to 1, with many close to 0.

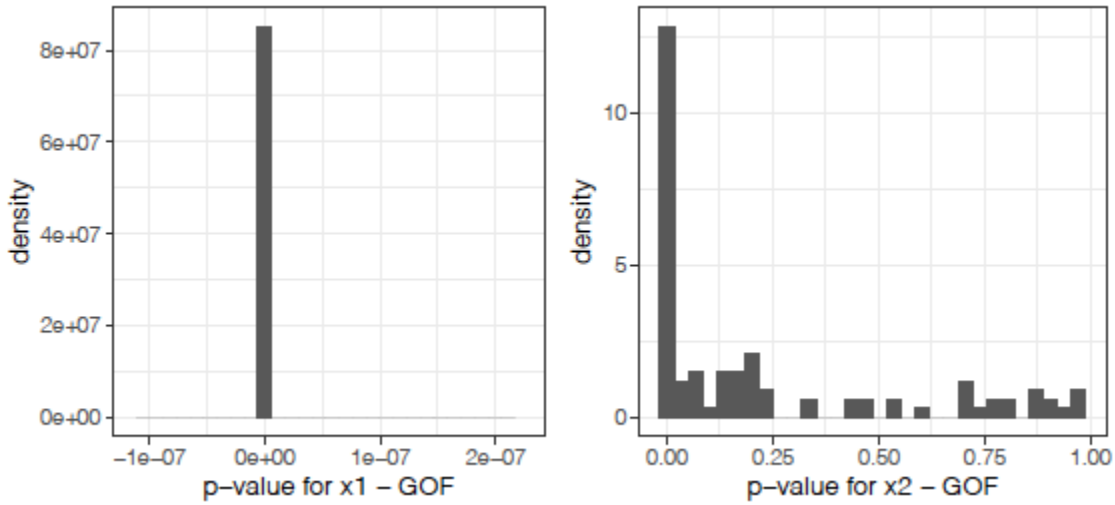

Figure 10: Distribution of p-values from the Goodness-of-fit test for the covariates in the non-proportionality case

Table 5 shows the averaged results of goodness-of-fit test for the Cox model in the non-proportional hazards case. Since the averaged p.value for covariate X_1 is statistically significant, I conclude that in general, covariate X_1 does not satisfy the proportionality assumption. Because the averaged p-value for covariate X_2 is larger than the significance level, I conclude that X_2 satisfies the proportionality assumption. The averaged estimates across 100 simulations are far different from the true parameters.

Covariate	True parameter	Estimate	Averaged p-value
X_1	4	3.406	0.000
X_2	-1	-0.852	0.234

Table 3: Averaged p-value from the Goodness-of-fit test for covariates across 100 simulations in the non-proportionality case

4. Discussion

In Table 6, *proportional* indicates that the covariate satisfies the PH assumption, while *time-varying* means it does not. From the simulation studies, there is an inconsistency in the conclusion from different assessing approaches. Specifically, for covariate X_1 in the first case, which assumes proportionality, the averaged p-value for the time-by-covariate interaction is statistically significant, which indicates that X_1 has a time-varying effect. However, the visualizations and the insignificant averaged p-value from the Goodness-of-fit states otherwise. Other than this, the conclusions drawn from three approaches are consistent.

Assessing Techniques	Proportional Hazards		Non-proportional hazards	
	X_1	X_2	X_1	X_2
Approach 1: Kaplan-Meier	Proportional	Proportional	Time-varying	Proportional
Approach 1: Log-log survival	Parallel	Parallel	Time-varying	Parallel
Approach 2: Time interaction	Time-varying	Proportional	Time-varying	Proportional
Approach 3: GOF test	Proportional	Proportional	Time-varying	Proportional

Table 6: Summary Table of the Proportionality Assessment from the Simulation Study. The input is from the results in the Simulation Study for both cases: proportional hazards and non-proportional hazards

When I was running the simulation, I tried different values for the censoring hazard and the true parameters so that the covariates are proportional in the proportionality case and are time-dependent in the non-proportionality case. However, I wasn't able to get the insignificant averaged p-values for both the interaction terms despite many different values for censoring hazard and true parameter when the proportionality is assumed to hold, even though the GOF test states otherwise and the event rate is not too high.

Among the three approaches to assess the appropriateness of the Cox model, the first approach using Kaplan-Meier curves and log-log survival curves tends to be subjective. In particular, log-log survival curves indicate proportionality if they are parallel, but the standard of "parallel" varies among people. From the distributions of p-values for time interaction terms in approach 2 and from the GOF test in approach 3, there is a variation for the values of p-values, ranging from 0 to 1. It can partly explain the inconsistency in the summary table. In addition, in this simulation, I used the linear function for time *t*. However, there are many other options such as logarithmic, exponential, polynomial, spline or step function choices for the function of *t*. In this approach with interaction terms, the correct functional form for $f(t)$ is critical and different choices may lead to different conclusions about PH assumption. Since there may be insufficient information to judge the precise nature of $f(t)$, in order to gauge the general form of $f(t)$, one can plot both the step function estimates and one or more simple parametric estimates (Hess, 1995). However, this step is beyond the scope of my project.

For approach 3, each of the p-values tests the PH assumption for one covariate given the other ones included in the model satisfy the assumption (Kleinbaum & Klein, 2010). However, there is no guarantee that the other covariates, not being tested, satisfy the assumption. One can use the GOF test with the strategy one-at-a-time for each of the covariates in the model but no specific strategy is clearly preferable. Thus, it is recommended to use at least 2 of the available techniques to assess the appropriateness of the Cox model, preferably a graphical and a test-based approach so that the assessment can be as comprehensive as possible.

From the Summary Table, two among three approaches show that X_1 is not time-varying, so I will conclude that the proportional hazards for X_1 is constant over time. With that information, it seems that when the proportional hazards assumption is not met, the inferences from the Cox model might not be valid. The estimates are close to the true parameter values if the PH assumption is satisfied (see case 1: Proportionality). The estimates are different from the true parameters if proportionality is violated (see case 2: Non-proportionality). Thus, it is very crucial to examine the proportionality and assess the appropriateness of the model.

5. Motivating data

5.1 Description of the PBC clinical data

The dataset is from the Primary Biliary Cirrhosis (PBC) clinical trial in 312 patients conducted by the Mayo Clinic for a 10-year period from 1975 to 1984 (Murtaugh, 1994). Among these patients, 154 of them were randomly put in the placebo group and the rest were in the treatment group. By the end of the follow-up, which was extended to 1988, 140 of the patients died, 29 had received orthotopic liver transplantation and 143 were still alive. After accounting for death and censoring, there were 1945 patient visits, with repeated measurements capturing the progression of PBC longitudinal biomarkers (indicators of progression and stages diseases) such as serum bilirubin, albumin, prothrombin time, the presence of spiders - blood vessel malformation in the skin, etc. These biomarkers were measured at specific visits at six months, one year, and annually thereafter. In addition to these biomarkers, there were baseline covariates such as age, drug and gender.

The primary objective of this randomized placebo controlled trial was to investigate the treatment effect of the drug D-penicillamine on overall survival of the PBC patients. In this project, the main goal is to study how serum bilirubin is in association with overall survival of PBC patients. When the liver fails to excrete bilirubin, high levels of this serum can cause jaundice of the skin, which is one of known clinical features of cirrhosis (Schuppan & Afdhal 2008). Insights on how serum bilirubin is associated with the hazard of death of PBC patients are valuable. Researchers can better adjust personal care and improve the treatment regimen of PBC patients.

In the PBC clinical data, the time variable is the number of years between enrollment and the earlier of death, transplantation, or study analysis time. The event of interest is death. Thus the event is 1 if death occurs and 0 for alive/transplanted patients. The baseline measurements of biomarkers are used in the Cox model.

5.2 Results from the Cox model

The PBC analysis follows this procedure:

- Step 1: Univariate analysis with the threshold of 0.1
- Step 2: Backward elimination with the threshold of 0.05
- Step 3: Multivariate analysis with chosen covariates from Step (2)

This is a commonly used variable selection procedure in survival modeling. Examples of this procedure can be seen in the studies of survival in colorectal carcinoma (SchmitzMoormann et al. 1987), gastric cancer (Maruyama 1987), and advanced non-small-cell lung cancer (Paesmans et al. 1995). The summary of the univariate analysis is run with functions in the package *gtsummary* (Sjoberg et al. 2020)

5.2.1 Univariate Analysis

In Table 6, from the Univariate Analysis, *drug* is the only variable that is not statistically significant at the level of 0.10. Thus, *drug* is not included in the variable selection. In this summary, the hazard ratio for bilirubin is 1.16. That is, each unit increase in the level of bilirubin is associated with an 16% increase in the hazard for death of PBC patients, not accounting for any effects of other covariates in the data.

Univariate Analysis			
Covariate	**Hazard Ratio**	**95% CI**	**p-value**
Drug: placebo	*ref*		
D-penicillin	1.00	(0.72 - 1.39)	>0.9
Age	1.05	(1.03 - 1.06)	<0.001
Sex: male	*ref*		
female	0.52	(0.34 - 0.80)	0.005

Bilirubin	1.16	(1.13 – 1.19)	<0.001
Alkaline	1.00	(1.00 – 1.00)	0.094
SGOT	1.01	(1.00 – 1.01)	<0.001
Platelets	1.00	(0.99 – 1.00)	<0.001
Prothrombin	2.12	(1.81 – 2.48)	<0.001
Ascites: No	*ref*		<0.001
Yes	7.58	(4.78 – 12.0)	
Hepatomegaly: No	*ref*		<0.001
Yes	3.06	(2.14 – 4.38)	
Spiders: No	*ref*		<0.001
Yes	2.42	(1.72 – 3.42)	
Edema: No edema	*ref*		<0.001
Edema no diuretics	1.63	(1.04 – 2.55)	
Edema diuretics	10.9	(6.61 – 18.0)	
Histologic: 1	*ref*		<0.001
2	6.39	(0.86 – 47.5)	
3	9.66	(1.33 – 70.1)	
4	24.0	(3.33 – 174)	

Table 6: Univariate Cox PH Model

5.2.2 Multivariate Analysis

In table 7, from the Multivariate Analysis, after accounting for the other covariates in the model, the hazard ratio for bilirubin is 1.11. That is, each unit increase in the level of bilirubin is associated with an 11% increase in the hazard for death of PBC patients.

Multivariate Analysis			
Characteristics	**Hazard Ratio**	**95% CI**	**p–value**
Bilirubin	1.11	(1.06 – 1.15)	<0.001
Albumin	0.52	(0.32 – 0.85)	0.008
Age	1.04	(1.03 – 1.06)	0.094
Edema: No edema	*ref*		0.019
Edema no diuretics	1.04	(0.65 – 1.67)	

Edema diuretics	2.37	(1.33 – 4.22)	
Histologic: 1	*ref*		0.014
2	4.49	(0.60 – 33.8)	
3	5.79	(0.79 – 42.5)	
4	8.04	(1.09 – 59.5)	
SGOT	1.00	(1.00 – 1.01)	0.012
Prothrombin	1.46	(1.20 – 1.78)	<0.001

Table 7: Multivariate Cox PH Model

5.3 Assessing Proportionality

Since bilirubin is a continuous covariate, it is stratified into two groups: high and normal bilirubin with a clinical cut-off of 1.2 mg/dl. Since Kaplan-Meier curves may yield estimation error, especially with a small sample size, I only use log-log survival curves in Approach 1 using visualization because log-log survival curves are more robust.

5.3.1 Approach 1: Graphics Log-log survival curves

In Figure 13, the difference in the log-log curves does not seem to be constant. The two survival curves are not parallel. It seems that the proportionality is not met for Bilirubin.

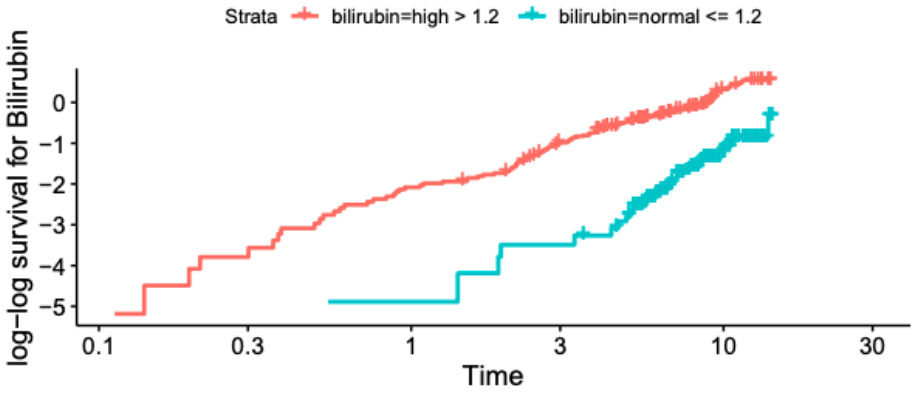

Figure 13: Log-log survival curves for Bilirubin

5.3.2 Approach 2: Interaction terms

In Table 8, Bilirubin does not have time-varying effect because the p-value for the time:bilirubin interaction term at 0.607 is statistically insignificant.

Covariates	**p-value**

bilirubin	0.960
time	2e-16
time:bilirubin	0.607

Table 8: Time by Bilirubin interaction

5.3.3 Approach 3: GOF test

From the GOF test in Table 9, bilirubin is statistically significant with a statistically significant p-value of 0.012.

Proportionality Assumption in Cox PH Model	
Characteristics	**p – value**
Bilirubin	0.012
Albumin	0.081
Age	0.659
Edema	0.122
Histologic	0.086
SGOT	0.525
Prothrombin	0.223
GLOBAL	0.022

Table 9: GOF test for PBC data

5.4 Future direction

Two among the three approaches indicate that bilirubin is time-varying, so I conclude that there is a time-varying effect of bilirubin on the hazard of death. Since the hazards are not proportional for bilirubin over time in the multivariate analysis, the estimated hazards ratio from the result is not valid. There are several ways to address non-proportionality depending on the study objectives. If non-proportionality is unlikely problematic on short time intervals, one can shorten follow-up time (Bellera et al. 2010). If the covariate with time-varying effect detected from the GOF test is not of interest, it is possible to stratify this covariate using the stratified Cox model (Kleinbaum & Klein 2010).

In the PBC clinical data, the motivation aims to study the effect of bilirubin on the overall survival of PBC patients, one can use the Time-Dependent Cox model (extended Cox model), which takes into account the progression of bilirubin over time. The Time-Dependent Cox model uses the Counting Process data format. In the Counting Process data format, there are multiple lines holding information for the same subject to divide the total follow-up time into small time

intervals with START and END points. In addition, the format holds information for repeated measurements of longitudinal biomarkers for subjects.

id	years	drug	age	sex	ascites	serBilir	start	end	event
1	1.0951703	D-penicil	58.76684	female	Yes	14.5	0.0000000	0.5256817	0
1	1.0951703	D-penicil	58.76684	female	Yes	21.3	0.5256817	1.0951703	1
2	14.1523382	D-penicil	56.44782	female	No	1.1	0.0000000	0.4983025	0
2	14.1523382	D-penicil	56.44782	female	No	0.8	0.4983025	0.9993429	0
2	14.1523382	D-penicil	56.44782	female	No	1.0	0.9993429	2.1027270	0
2	14.1523382	D-penicil	56.44782	female	No	1.9	2.1027270	4.9008871	0
2	14.1523382	D-penicil	56.44782	female	Yes	2.6	4.9008871	5.8892783	0
2	14.1523382	D-penicil	56.44782	female	Yes	3.6	5.8892783	6.8858833	0
2	14.1523382	D-penicil	56.44782	female	Yes	4.2	6.8858833	7.8907020	0
2	14.1523382	D-penicil	56.44782	female	Yes	3.6	7.8907020	8.8325485	0
2	14.1523382	D-penicil	56.44782	female	Yes	4.6	8.8325485	14.1523382	0

Table 10: Counting Process data format for repeated measurements

References

Bellera, C. A., MacGrogan, G., Debled, M., de Lara, C. T., Brouste, V. & Mathoulin Pelissier, S. (2010), `Variables with time-varying effects and the cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer', BMC medical research methodology 10(1), 20.

Bouliotis, G. & Billingham, L. (2011), `Crossing survival curves: alternatives to the log-rank test', Trials 12(S1), A137.

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

Hess, K. R. (1995), `Graphical methods for assessing violations of the proportional hazards assumption in cox regression', Statistics in medicine 14(15), 1707-1723.

Kleinbaum, D. G. & Klein, M. (2010), Survival analysis, Springer.

Kleinman, K. & Horton, N. J. (2014), SAS and R: Data management, statistical analysis, and graphics, CRC Press.

Leung, K.-M., Elasho, R. M. & A, A. A. (1997), 'Censoring issues in survival analysis', Annual review of public health 18(1), 83-104.

Maruyama, K. (1987), `The most important prognostic factors for gastric cancer patients: a study using univariate and multivariate analyses', Scandinavian Journal of Gastroenterology 22(sup133), 63-68.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L., & Gips, C. H. (1994). Primary biliary cirrhosis: prediction of short‐term survival based on repeated patient visits. *Hepatology*, *20*(1), 126-134.

Paesmans, M., Sculier, J.-P., Libert, P., Bureau, G., Dabouis, G., Thiriaux, J., Michel, J., Van Cutsem, O., Sergysels, R. & Mommen, P. (1995), `Prognostic factors for survival in advanced non-small-cell lung cancer: univariate and multivariate analyses including recursive partitioning and amalgamation algorithms in 1,052 patients. The european lung cancer working party.', Journal of Clinical Oncology 13(5), 1221-1230.

Schmitz-Moormann, P., Himmelmann, G., Baum, U. & Nilles, M. (1987), `Morphological predictors of survival in colorectal carcinoma: univariate and multivariate analysis', Journal of cancer research and clinical oncology 113(6), 586-592.

Schuppan, D. & Afdhal, N. H. (2008), `Liver cirrhosis', The Lancet 371(9615), 838-851.

Sjoberg, D., Hannum, M., Whiting, K. & Zabor, E. (2020), `Gtsummary: Presentation ready data summary and analytic result tables'.

Therneau, T. M. & Lumley, T. (2015), `Package `survival'', R Top Doc 128, 112.