4 Evaluating the Proportional Hazards Assumption

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_4,

Introduction

We begin with a brief review of the characteristics of the Cox proportional hazards (PH) model. We then give an overview of three methods for checking the PH assumption: graphical, goodness-of-fit (GOF), and time-dependent variable approaches.

We then focus on each of the above approaches, starting with graphical methods. The most popular graphical approach involves the use of "log–log" survival curves. A second graphical approach involves the comparison of "observed" with "expected" survival curves.

The GOF approach uses a test statistic or equivalent p-value to assess the significance of the PH assumption. We illustrate this test and describe some of its advantages and drawbacks.

Finally, we discuss the use of time-dependent variables in an extended Cox model as a third method for checking the PH assumption. A more detailed description of the use of time-dependent variables is provided in Chapter 6.

Abbreviated Outline

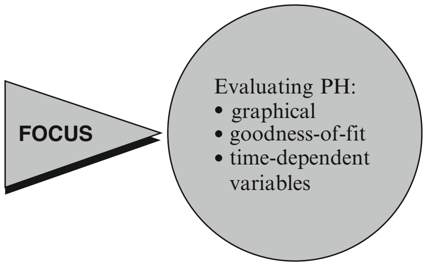
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize three general approaches for evaluating the PH assumption.
2. Summarize how log–log survival curves may be used to assess the PH assumption.
3. Summarize how observed versus expected plots may be used to assess the PH assumption.
4. Summarize how GOF tests may be used to assess the PH assumption.
5. Summarize how time-dependent variables may be used to assess the PH assumption.
6. Describe — given survival data or computer output from a survival analysis that uses a Cox PH model — how to assess the PH assumption for one or more variables in the model using:
   a. a graphical approach
   b. the GOF approach
   c. an extended Cox model with time-dependent covariates
7. State the formula for an extended Cox model that provides a method for checking the PH assumption for one or more of the time-independent variables in the model, given survival analysis data or computer output that uses a Cox PH model.

Presentation



This presentation describes three approaches for evaluating the proportional hazards (PH) assumption of the Cox model — a graphical procedure, a goodness-of-fit testing procedure, and a procedure that involves the use of time-dependent variables.

I. Background

Cox PH model:

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

Recall from the previous chapter that the general form of the Cox PH model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by the bold **X**.

$\mathbf{X} = (X_1, X_2, \ldots, X_p)$ explanatory/predictor variables

The Cox model formula says that the hazard at time t is the product of two quantities. The first of these, $h_0(t)$, is called the **baseline hazard** function. The second quantity is the exponential expression e to the linear sum of $\beta_i X_i$, where the sum is over the p explanatory X variables.

$h_0(t)$ $\times$	$e^{\sum_{i=1}^{p} \beta_i X_i}$
Baseline hazard Involves t but not X's	Exponential Involves X's but not t (X's are time-independent)

An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the X's, whereas the exponential expression involves the X's, but does not involve t. The X's here are called **time-independent** X's.

X's involving t: **time-dependent**
Requires **extended Cox model** (no PH)
↑
Chapter 6

It is possible, nevertheless, to consider X's that do involve t. Such X's are called **time-dependent** variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the **extended Cox model**. We will discuss this extended Cox model in Chapter 6 of this series.

Hazard ratio formula:

$$\widehat{HR} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

where $\mathbf{X}^* = \left(X_1^*, X_2^*, \cdots, X_p^*\right)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote the two sets of X's.

From the Cox PH model, we can obtain a general formula, shown here, for estimating a hazard ratio that compares two specifications of the X's, defined as $\mathbf{X}^*$ and $\mathbf{X}$.

Adjusted survival curves comparing E groups:

$$\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[\beta_1 E + \sum_{i \neq 1} \beta_i \overline{X}_i\right]}$$

Single curve:

$$S(t, \overline{\mathbf{X}}) = \left[\hat{S}_0(t)\right]^{\exp\left[\sum_{i=1}^{p} \hat{\beta}_i \overline{X}_i\right]}$$

We can also obtain from the Cox model an expression for an adjusted survival curve. Here we show a general formula for obtaining adjusted survival curves comparing two groups adjusted for other variables in the model. Below this, we give a formula for a single adjusted survival curve that adjusts for all X's in the model. Computer packages for these formulae use the mean value of each X being adjusted in the computation of the adjusted curve.

PH assumption:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}, \text{ constant over } t$$

i.e., $\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X})$

The Cox PH model assumes that the hazard ratio comparing any two specifications of predictors is constant over time. Equivalently, this means that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

Hazards cross: $\Rightarrow$ PH not met

Hazards don't cross $\not\Rightarrow$ PH met

The PH assumption is not met if the graph of the hazards cross for two or more categories of a predictor of interest. However, even if the hazard functions do not cross, it is possible that the PH assumption is not met. Thus, rather than checking for crossing hazards, we must use other approaches to evaluate the reasonableness of the PH assumption.

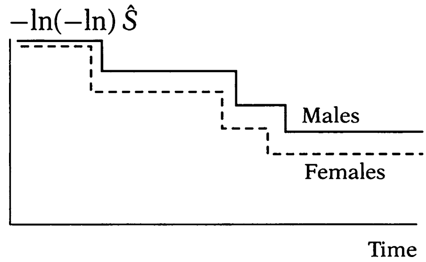
II. Checking the Proportional Hazards Assumption: Overview

There are three general approaches for assessing the PH assumption, again listed here. We now briefly overview each approach, starting with graphical techniques.

Three approaches:

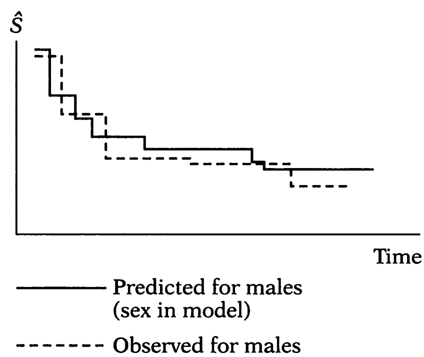
- graphical
- goodness-of-fit test
- time-dependent variables

Graphical techniques:
–ln(–ln) S curves parallel?



There are two types of graphical techniques available. The most popular of these involves comparing **estimated –ln(–ln) survivor curves** over different (combinations of) categories of variables being investigated. We will describe such curves in detail in the next section. Parallel curves, say comparing males with females, indicate that the PH assumption is satisfied, as shown in this illustration for the variable Sex.

Observed vs. predicted: Close?



An alternative graphical approach is to compare **observed with predicted** survivor curves. The observed curves are derived for categories of the variable being assessed, say, Sex, without putting this variable in a PH model. The predicted curves are derived with this variable included in a PH model. If observed and predicted curves are close, then the PH assumption is reasonable.

Goodness-of-fit (GOF) tests:

- Large sample Z or chi-square statistics
- Gives p-value for evaluating PH assumption for each variable in the model.

p-value large ⇒ PH satisfied
(e.g. $P > 0.10$)
p-value small ⇒ PH not satisfied
(e.g. $P < 0.05$)

A second approach for assessing the PH assumption involves goodness-of-fit (GOF) tests. This approach provides large sample Z or chi-square statistics which can be computed for each variable in the model, adjusted for the other variables in the model. A p-value derived from a standard normal statistic is also given for each variable. This p-value is used for evaluating the PH assumption for that variable. A nonsignificant (i.e., large) p-value, say greater than 0.10, suggests that the PH assumption is reasonable, whereas a small p-value, say less than 0.05, suggests that the variable being tested does not satisfy this assumption.

Time-dependent covariates:

Extended Cox model:
Add product term involving some function of time.

When time-dependent variables are used to assess the PH assumption for a time-independent variable, the Cox model is extended to contain **product** (i.e., interaction) **terms** involving the time-independent variable being assessed and some function of time.

EXAMPLE

$h(t, X) = h_0(t) \exp[\beta\text{Sex} + \delta(\text{Sex} \times t)]$
$\delta \neq 0$. PH assumption violated

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable "Sex $\times$ t" in addition to Sex. If the coefficient of the product term turns out to be significant, we can conclude that the PH assumption is violated for Sex.

GOF provides test statistic
Graphical: subjective
Time-dependent: computationally cumbersome
GOF: global, may not detect specific departures from PH

The GOF approach provides a single test statistic for each variable being assessed. This approach is not as subjective as the graphical approach nor as cumbersome computationally as the time-dependent variable approach. Nevertheless, a GOF test may be too "global" in that it may not detect specific departures from the PH assumption that may be observed from the other two approaches.

III. Graphical Approach 1: Log–Log Plots

- log–log survival curves
- observed versus expected survival curves

The two graphical approaches for checking the PH assumption are comparing log–log survival curves and comparing observed versus expected survival curves. We first explain what a −ln −ln survival curve is and how it is used.

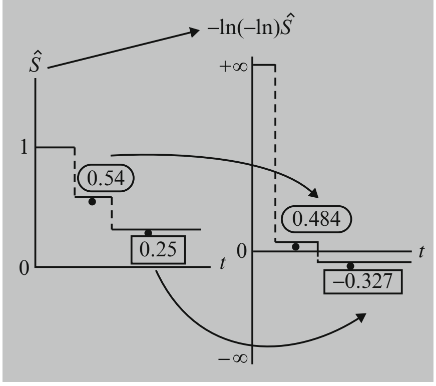
$\log-\log \hat{S}$ = transformation of $\hat{S}$
$= -\ln-\ln \hat{S})$

- $\ln \hat{S}$ is negative $\Rightarrow -$ ($\ln \hat{S}$) is positive.
- can't take log of $\ln \hat{S}$, but can take log of $(-\ln \hat{S})$.
- $-\ln(-\ln \hat{S})$ may be positive or negative.

A log–log survival curve is simply a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability *twice*. Mathematically, we write a log–log curve as $-\ln(-\ln \hat{S})$. Note that the log of a probability such as $\hat{S}$ is always a negative number. Because we can only take logs of positive numbers, we need to negate the first log before taking the second log. The value for $-\ln(-\ln \hat{S})$ may be positive or negative, either of which is acceptable, because we are not taking a third log.[1]

[1]An equivalent way to write to $-\ln(-\ln\hat{S})$ is $-\ln\left(\int_0^t h(u)du\right)$, where $\int_0^t h(u)du$ is called the "cumulative hazard" function. This result follows from the formula $S(t) = \exp\left[-\int_0^t h(u)du\right]$, which relates the survivor function to the hazard function (see p. 15 in Chapter 1).

EXAMPLE



As an example, in the graph at left, the estimated survival probability of 0.54 is transformed to a log–log value of 0.484. Similarly, the point 0.25 on the survival curve is transformed to a $-\ln -\ln$ value of -0.327.

Note that because the survival curve is usually plotted as a step function, so will the log–log curve be plotted as a step function.

EXAMPLE

$\hat{S} = 0.54$: want $-\ln(-\ln 0.54)$
$-\ln(-\ln 0.54) = -\ln(0.616)$
since $\ln(0.54) = -0.616$
$-\ln(0.616) = 0.484$
since $\ln(0.616) = -0.484$
Thus, $\boxed{-\ln(-\ln 0.54) = 0.484}$

To illustrate the computation of a log–log value, suppose we start with an estimated survival probability of 0.54. Then the log–log transformation of this value is $-\ln(-\ln 0.54)$, which is $-\ln(0.616)$, because $\ln(0.54)$ equals -0.616. Now, continuing further, $-\ln(0.616)$ equals 0.484, because $\ln(0.616)$ equals -0.484. Thus, the transformation $-\ln(-\ln 0.54)$ equals 0.484.

ANOTHER EXAMPLE

$\hat{S} = 0.25$: want $-\ln(-\ln 0.25)$
$-\ln(-\ln 0.25) = -\ln(1.386) = -0.327$
Thus, $\boxed{-\ln(-\ln 0.25) = -0.327}$

As another example, if the estimated survival probability is 0.25, then $-\ln(-\ln 0.25)$ equals $-\ln(1.386)$, which equals -0.327.

y-axis scale:

$$\begin{array}{c|c} 1 & \\ & \hat{S} \\ 0 & \end{array} \qquad \begin{array}{c|c} +\infty & \\ & -\ln(1-\ln)\hat{S} \\ -\infty & \end{array}$$

Note that the scale of the y-axis of an estimated survival curve ranges between 0 and 1, whereas the corresponding scale for a $-\ln(-\ln)$ curve ranges between $-\infty$ and $+\infty$.

log–log $\hat{S}$ for the Cox PH model:

We now show why the PH assumption can be assessed by evaluating whether or not log–log curves are parallel. To do this, we must first describe the log–log formula for the Cox PH model.

Cox PH hazard function:

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

$\downarrow$ From math

Cox PH survival function:

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{\sum_{i=1}^{p} \beta_i X_i}}$$

$\nearrow$

Baseline survival function.

We start with the formula for the survival curve that corresponds to the hazard function for the Cox PH model. Recall that there is a mathematical relationship between any hazard function and its corresponding survival function. We therefore can obtain the formula shown here for the survival curve for the Cox PH model. In this formula, the expression $S_0(t)$ denotes the baseline survival function that corresponds to the baseline hazard function $h_0(t)$.

log–log$\Rightarrow$ takes logs twice

The log–log formula requires us to take logs of this survival function twice. The first time we take logs we get the expression shown here.

log #1:

$$\ln\ S(t, \mathbf{X}) = e^{\sum_{i=1}^{p} \beta_i X_i} \times\ \ln\ S_0(t)$$

$$0 \leq S(t, \mathbf{X}) \leq 1$$

ln(probability) = negative value, so $\ln S(t,\mathbf{X})$ and $\ln S_0(t)$ are negative.

But $-\ln\ S(t,\mathbf{X})$ is positive, which allows us to take logs again.

Now since $S(t, \mathbf{X})$ denotes a survival probability, its value for any t and any specification of the vector $\mathbf{X}$ will be some number between 0 and 1. It follows that the natural log of any number between 0 and 1 is a negative number, so that the log of $S(t, \mathbf{X})$ as well as the log of $S_0(t)$ are both negative numbers. This is why we have to put a minus sign in front of this expression before we can take logs a second time, because there is no such thing as the log of a negative number.

log #2:

$$\begin{aligned}
&\ln[-\ln S(t, \mathbf{X})] \\
&\quad = \ln\left[-e^{\sum_{i=1}^{p} \beta_i X_i} \times\ \ln\ S_0(t)\right] \\
&\quad = \ln\left[e^{\sum_{i=1}^{p} \beta_i X_i}\right] + \ln[-\ln\ S_0(t)] \\
&\quad = \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln\ S_0(t)] \\
&\qquad = \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln\ S_0(t)]
\end{aligned}$$

Thus, when taking the second log, we must obtain the log of $-\ln\ S(t, \mathbf{X})$, as shown here. After using some algebra, this expression can be rewritten as the sum of two terms, one of which is the **linear sum of the** $\beta_i X_i$ and the other is the **log of the negative log of the baseline survival function**.

or

$$\begin{aligned}
&-\ln[-\ln\ S(t, \mathbf{X})] \\
&\qquad = -\sum_{i=1}^{p} \beta_i X_i - \ln[-\ln\ S_0(t)]
\end{aligned}$$

This second log may be either positive or negative, and we aren't taking any more logs, so we actually don't have to take a second negative. However, for consistency's sake, a common practice is to put a minus sign in front of the second log to obtain the $-\ln\ -\ln$ expression shown here. Nevertheless, some software packages do not use a second minus sign.

Two individuals:
$\mathbf{X}_1 = (X_{11}, X_{12}, \ldots, X_{1p})$
$\mathbf{X}_2 = (X_{21}, X_{22}, \ldots, X_{2p})$

Now suppose we consider two different specifications of the **X** vector, corresponding to two different individuals, $\mathbf{X}_1$ and $\mathbf{X}_2$.

$$\begin{cases} \ln[-\ln\ S(t, \mathbf{X}_1)] \\ \quad = \sum_{i=1}^{p} \beta_i X_{1i} + \ln[-\ln\ S_0(t)] \\ \ln[-\ln\ S(t, \mathbf{X}_2)] \\ \quad = \sum_{i=1}^{p} \beta_i X_{2i} + \ln[-\ln\ S_0(t)] \end{cases}$$

Then the corresponding log–log curves for these individuals are given as shown here, where we have simply substituted $\mathbf{X}_1$ and $\mathbf{X}_2$ for **X** in the previous expression for the log–log curve for any individual **X**.

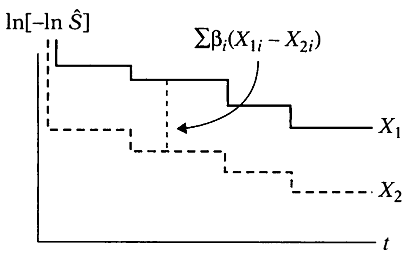
$$\ln[-\ln\ S(t, \mathbf{X}_1)] - (\ln[-\ln\ S(t, \mathbf{X}_2)]) = \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})$$

does not involve t

Subtracting the second log–log curve from the first yields the expression shown here. This expression is a linear sum of the differences in corresponding predictor values for the two individuals. Note that the baseline survival function has dropped out, so that the difference in log–log curves involves an expression that does not involve time t.

$$\boxed{\begin{aligned} -\ln[-\ln\ S(t, \mathbf{X}_1)] \\ = \ln[-\ln\ S(t, \mathbf{X}_2)] \\ + \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i}) \end{aligned}}$$

Alternatively, using algebra, we can write the above equation by expressing the log–log survival curve for individual $\mathbf{X}_1$ as the log–log curve for individual $\mathbf{X}_2$ plus a linear sum term that is independent of t.



The above formula says that if we use a Cox PH model and we plot the estimated log–log survival curves for individuals on the same graph, the two plots would be approximately parallel. The distance between the two curves is the linear expression involving the differences in predictor values, which does not involve time. Note, in general, if the vertical distance between two curves is constant, then the curves are parallel.

Graphical approach using log–log plots: PH model is appropriate if "empirical" plots of log–log survival curves are parallel.

The parallelism of log–log survival plots for the Cox PH model provides us with a graphical approach for assessing the PH assumption. That is, if a PH model is appropriate for a given set of predictors, one should expect that empirical plots of log–log survival curves for different individuals will be approximately parallel.

Empirical plots: use $-\ln[-\ln\ \hat{S}]$ where

1. $\hat{S}$ is a KM curve
2. $\hat{S}$ is an adjusted survival curve for predictors satisfying the PH assumption; predictor being assessed not included in model

By empirical plots, we mean plotting log–log survival curves based on Kaplan–Meier (KM) estimates that do not assume an underlying Cox model. Alternatively, one could plot log–log survival curves which have been adjusted for predictors already assumed to satisfy the PH assumption but have not included the predictor being assessed in a PH model.

EXAMPLE

Clinical trial of leukemia patients:
T = weeks until patient goes out of remission

Predictors (X's):
Rx (= 1 if placebo, 0 if treatment)
log WBC

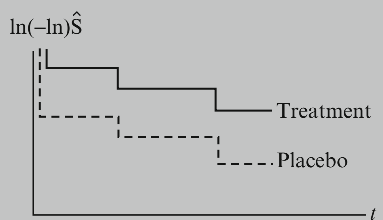
Cox PH model:

$h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$
Assessing PH assumption: compare log–log survival curves for categories of Rx and log WBC

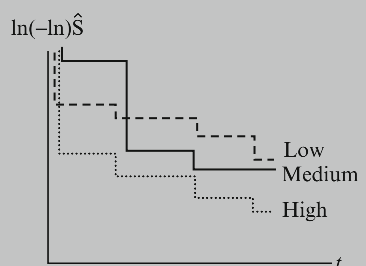
As an example, suppose we consider the comparison of treatment and placebo groups in a clinical trial of leukemia patients, where survival time is time, in weeks, until a patient goes out of remission. Two predictors of interest in this study are treatment group status (1 = placebo, 0 = treatment), denoted as Rx, and log white blood cell count (log WBC), where the latter variable is being considered as a confounder.

A Cox PH model involving both these predictors would have the form shown at the left. To assess whether the PH assumption is satisfied for either or both of these variables, we would need to compare log–log survival curves involving categories of these variables.

One-at-a-time strategy: Rx variable



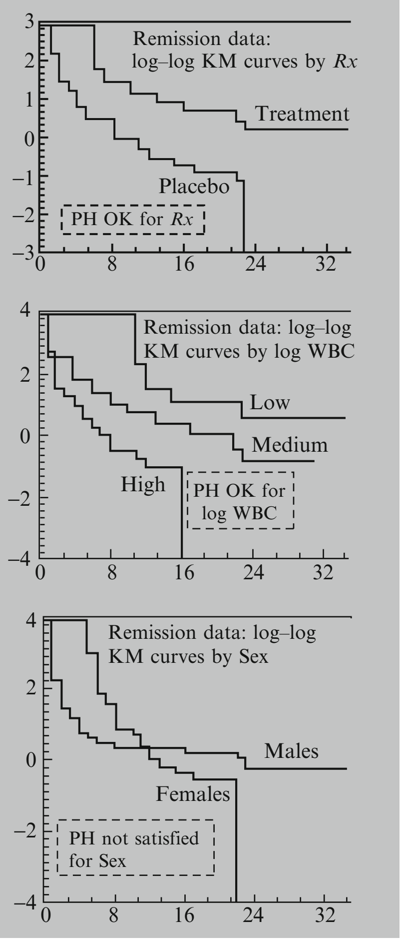
One-at-a-time strategy: log WBC



One strategy to take here is to consider the variables one at a time. For the Rx variable, this amounts to plotting log–log KM curves for treatment and placebo groups and assessing parallelism. If the two curves are approximately parallel, as shown here, we would conclude that the PH assumption is satisfied for the variable Rx. If the two curves intersect or are not parallel in some other way, we would conclude that the PH assumption is not satisfied for this variable.

For the log WBC variable, we need to categorize this variable into categories – say, low, medium, and high – and then compare plots of log–log KM curves for each of the three categories. In this illustration, the three log–log Kaplan–Meier curves are clearly nonparallel, indicating that the PH assumption is not met for log WBC.

EXAMPLE: Computer Results



The above examples are sketches of some of the possibilities that could occur from comparisons of log–log curves. For the actual data set containing 42 leukemia patients, computer results are shown here for each variable separately. Similar output using Stata, SAS, SPSS, and R packages is provided in the Computer Appendix.

We first show the log–log KM curves by treatment, *Rx*. Notice that the two log–log curves are roughly parallel, indicating that the *Rx* variable satisfies the PH assumption when being considered by itself.

Here we show the log–log KM curves by log WBC, where we have divided this variable into low (below 2.3), medium (between 2.3 and 3), and high (above 3) values. Notice that there is some indication of nonparallelism below 8 days, but that overall the three curves are roughly parallel. Thus, these plots suggest that the PH assumption is more or less satisfied for the variable log WBC, when considered alone.

As a third example, we consider the log–log KM plots categorized by Sex from the remission data. Notice that the two curves clearly intersect, and are therefore noticeably nonparallel. Thus, the variable, Sex, when considered by itself, does not appear to satisfy the PH assumption and therefore should not be incorporated directly into a Cox PH model containing the other two variables, *Rx* and log WBC.

Problems with log–log survival curve approach:

How parallel is parallel?
Recommend:

- subjective decision
- conservative strategy: assume PH is OK unless strong evidence of nonparallelism

The above examples suggest that there are some problems associated with this graphical approach for assessing the PH assumption. The main problem concerns how to decide "how parallel is parallel?" This decision can be quite subjective for a given data set, particularly if the study size is relatively small. We recommend that one should use a conservative strategy for this decision by assuming the PH assumption is satisfied unless there is **strong** evidence of nonparallelism of the log–log curves.

How parallel is parallel?
Recommend:

- many categories ⇒ data "thins out"
- different categorizations may give different graphical pictures

Another problem concerns how to categorize a continuous variable like log WBC. If many categories are chosen, the data "thins out" in each category, making it difficult to compare different curves. [Also, one categorization into, say, three groups may give a different graphical picture from a different categorization into three groups.]

Recommend:

- small # of categories (2 or 3)
- meaningful choice
- reasonable balance (e.g., terciles)

In categorizing continuous variables, we recommend that the number of categories be kept reasonably small (e.g., two or three) if possible, and that the choice of categories be as meaningful as possible and also provide reasonable balance of numbers (e.g., as when using terciles).

How to evaluate several variables simultaneously?

In addition to the two problems just described, another problem with using log–log survival plots concerns how to evaluate the PH assumption for several variables simultaneously.

Strategy:

- categorize variables separately
- form combinations of categories
- compare log–log curves on same graph

One strategy for simultaneous comparisons is to categorize all variables separately, form combinations of categories, and then compare log–log curves for all combinations on the same graph.

Drawback:

- data "thins out"
- difficult to identify variables responsible for nonparallelism

A drawback of this strategy is that the data will again tend to "thin out" as the number of combinations gets even moderately large. Also, even if there are sufficient numbers for each combined category, it is often difficult to determine which variables are responsible for any nonparallelism that might be found.

EXAMPLE

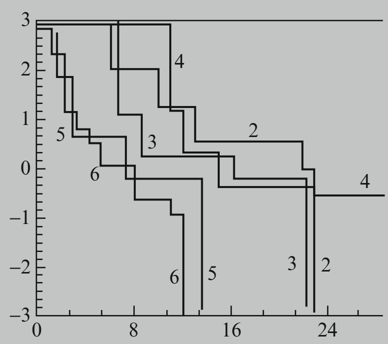
Remission Data:

Rx \ log WBC	Low	Medium	High
Treatment	✓	✓	✓
Placebo	✓	✓	✓

As an example of this strategy, suppose we use the remission data again and consider both *Rx* and log WBC together. Because we previously had two categories of *Rx* and three categories of log WBC, we get a total of six combined categories, consisting of treated subjects with low log WBC, placebo subjects with low log WBC, treated subjects with medium log WBC, and so on.

EXAMPLE: (continued)

Log–log $\widehat{\text{KM}}$ curves by six combinations of *Rx* by log WBC



Plots suggest PH not satisfied. However, the study is small, i.e., plots are unreliable.

The computer results are shown here for the log–log curves corresponding to each of the six combinations of *Rx* with log WBC. Notice that there are several points of intersection among the six curves. Therefore, these results suggest that the PH assumption is not satisfied when considering *Rx* and log WBC together.

However, the sample sizes used to estimate these curves are quite small, ranging between four subjects for group 4 (*Rx* = 1, log WBC = low) to 12 subjects for group 6 (*Rx* = 1, log WBC = high), with the total study size being 42. Thus, for this small study, the use of six log–log curves provides unreliable information for assessing the PH assumption.

Alternative strategy:

Adjust for predictors already satisfying PH assumption, i.e., use adjusted log–log $\hat{S}$ curves

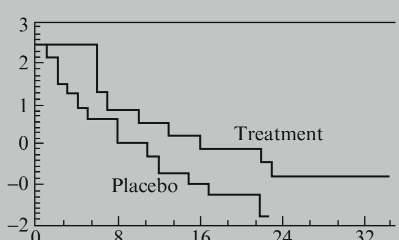
An alternative graphical strategy for considering several predictors together is to assess the PH assumption for one predictor adjusted for other predictors that are assumed to satisfy the PH assumption. Rather than using Kaplan–Meier curves, this involves a comparison of adjusted log–log survival curves.

EXAMPLE

Remission data:

- compare *Rx* categories adjusted for log WBC
- fit PH model for each *Rx* stratum
- obtain adjusted survival curves using overall mean of log WBC

Log–log $\hat{S}$ curves for *Rx* groups using PH model adjusted for log WBC

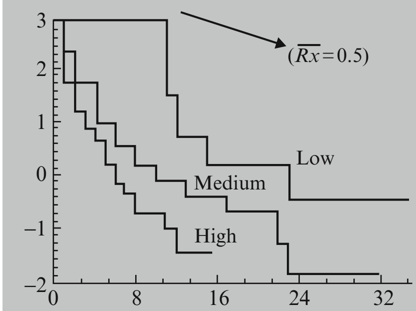


As an example, again we consider the remission data and the predictors *Rx* and log WBC. To assess the PH assumption for *Rx* adjusted for log WBC, we would compare adjusted log–log survival curves for the two treatment categories, where each adjusted curve is derived from a PH model containing log WBC as a predictor. In computing the adjusted survival curve, we need to stratify the data by treatment, fit a PH model in each stratum, and then obtain adjusted survival probabilities using the overall mean log WBC in the estimated survival curve formula for each stratum.

For the remission data example, the estimated log–log survival curves for the two treatment groups adjusted for log WBC are shown here. Notice that these two curves are roughly parallel, indicating that the PH assumption is satisfied for treatment.

EXAMPLE: (continued)

Log–log $\hat{S}$ curves for log WBC groups using PH model adjusted for *Rx*



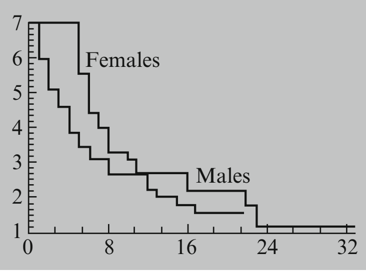
As another example, we consider adjusted log–log survival curves for three categories of log WBC, adjusted for the treatment status (*Rx*) variable. The adjusted survival probabilities in this case use the overall mean *Rx* score, i.e., 0.5, the proportion of the 42 total subjects that are in the placebo group (i.e., half the subjects have a score of *Rx* = 1).

The three log–log curves adjusted for treatment status are shown here. Although two of these curves intersect early in follow-up, they do not suggest a strong departure from parallelism overall, suggesting that the PH assumption is reasonable for log WBC, after adjusting for treatment status.

Remission data:
Assess PH assumption for Sex:

- use PH model containing *Rx* and log WBC
- use *Rx* and log WBC in survival probability formula

Log–log $\hat{S}$ curves for Sex adjusted for *Rx* and log WBC



As a third example, again using the remission data, we assess the PH assumption for Sex, adjusting for both treatment status and log WBC in the model. This involves obtaining log–log survival curves for males and females separately, using a PH model that contains both treatment status and log WBC. The adjustment uses the overall mean treatment score and the overall mean log WBC score in the formula for the estimated survival probability.

The estimated log–log survival curves for Sex, adjusted for treatment and log WBC are shown here. These curves clearly cross, indicating that the PH assumption is not satisfied for Sex, after adjusting for treatment and log WBC.

✓ 1. log–log survival curves
2. observed versus expected survival curves

We have thus described and illustrated one of the two graphical approaches for checking the PH assumption, that is, using log–log survival plots. In the next section, we describe an alternative approach that compares “observed” with “expected” survival curves.

IV. Graphical Approach 2: Observed Versus Expected Plots

Graphical analog of GOF test

The use of observed versus expected plots to assess the PH assumption is the graphical analog of the goodness-of-fit (GOF) testing approach to be described later, and is therefore a reasonable alternative to the log–log survival curve approach.

Two strategies:

1. One-at-a-time: uses KM curves to obtain observed plots
2. Adjusting for other variables: uses stratified Cox PH model to obtain observed plots (see Chapter 5)

As with the log–log approach, the observed versus expected approach may be carried out using either or both of two strategies—(1) assessing the PH assumption for variables one-at-a-time, or (2) assessing the PH assumption after adjusting for other variables. The strategy which adjusts for other variables uses a stratified Cox PH model to form observed plots, where the PH model contains the variables to be adjusted and the stratified variable is the predictor being assessed. The stratified Cox procedure is described in Chapter 5.

Here, we describe only the one-at-a-time strategy, which involves using KM curves to obtain observed plots.

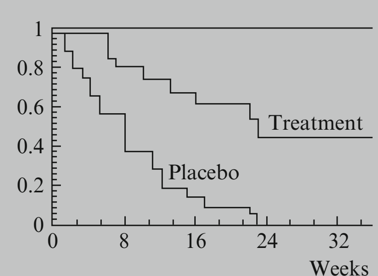
One-at-a-time:

- stratify data by categories of predictor
- obtain KM curves for each category

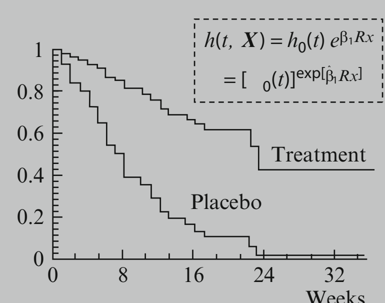
Using the one-at-a-time strategy, we first must stratify our data by categories of the predictor to be assessed. We then obtain observed plots by deriving the KM curves separately for each category.

EXAMPLE: Remission Data

KM (Observed) Plots by *Rx* Group



Expected Survival Plots by *Rx* Group Using PH Model



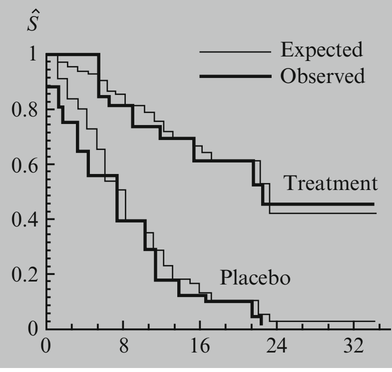
As an example, for the remission data on 42 leukemia patients we have illustrated earlier, the KM plots for the treatment and placebo groups, with 21 subjects in each group, are shown here. These are the "observed" plots.

To obtain "expected" plots, we fit a Cox PH model containing the predictor being assessed. We obtain expected plots by separately substituting the value for each category of the predictor into the formula for the estimated survival curve, thereby obtaining a separate estimated survival curve for each category.

As an example, again using the remission data, we fit the Cox PH model with *Rx* as its only variable. Using the corresponding survival curve formula for this Cox model, as given in the box at the left, we then obtain separate expected plots by substituting the values of 0 (for treatment group) and 1 (for placebo group). The expected plots are shown here.

EXAMPLE: (continued)

Observed Versus Expected Plots by *Rx*



To compare observed with expected plots we then put both sets of plots on the same graph as shown here.

If observed and expected plots are:

- **close**, complies with PH assumption
- **discrepant**, PH assumption violated

If for each category of the predictor being assessed, the observed and expected plots are "close" to one another, we then can conclude that the PH assumption is satisfied. If, however, one or more categories show quite discrepant observed and expected plots, we conclude that the PH assumption is violated.

Example: Remission Data (continued)

Observed and expected plots are close for each treatment group.
Conclude PH assumption not violated.

For the example shown above, observed and expected curves appear to be quite close for each treatment group. Thus, we would conclude using this graphical approach that the treatment variable satisfies the PH assumption.

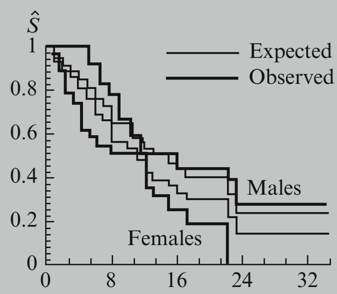
Drawback: How close is close?

Recommend: PH not satisfied *only* when plots are strongly discrepant.

An obvious drawback to this graphical approach is deciding "how close is close" when comparing observed versus expected curves for a given category. This is analogous to deciding "how parallel is parallel" when comparing log–log survival curves. Here, we recommend that the PH assumption be considered as not satisfied only when observed and expected plots are strongly discrepant.

Example: Remission Data

Observed Versus Expected Plots by Sex



PH assumption not satisfied for Sex. Same conclusion as with log–log curves.

As another example, again using the remission data, we consider observed versus expected plots by Sex, as shown here. Note that the observed plots for males and females, which are described by the thicker lines, cross at about 12 weeks, whereas the expected plots don't actually intersect, with the female plot lying below the male plot throughout follow-up. Moreover, for males and females separately, the observed and expected plots are quite different from one another.

Thus, the above plots suggest that the PH assumption is not satisfied for the variable Sex. We came to the same conclusion when using log–log survival curves, which crossed one another and were therefore clearly nonparallel.

Continuous variable:

- form strata from categories
- observed plots are KM curves for each category

When using observed versus expected plots to assess the PH assumption for a continuous variable, observed plots are derived, as for categorical variables, by forming strata from categories of the continuous variable and then obtaining KM curves for each category.

- two options for expected plots
  1. Use PH model with $k-1$ dummy variables X_c for k categories, i.e.,

     $$h(t,\mathbf{X}) = h_0(t)\exp\left(\sum_{i=1}^{k-1}\beta_i X_{ci}\right)$$

     Obtain adjusted survival curve:

     $$\hat{S}(t,\mathbf{X}_c) = \left[\hat{S}_0(t)\right]^{\exp\left(\sum \hat{\beta}_i X_{ci}\right)}$$

     where $\mathbf{X}_c = (X_{c1}, X_{c2}, \ldots, X_{c,k-1})$ gives values of dummy variables for category c.

However, for continuous predictors, there are two options available for computing expected plots. One option is to use a Cox PH model which contains $k-1$ dummy variables to indicate k categories. The expected plot for a given category is then obtained as an adjusted survival curve by substituting the values for the dummy variables that define the given category into the formula for the estimated survival curve, as shown here for category c.

Options for a continuous variable:

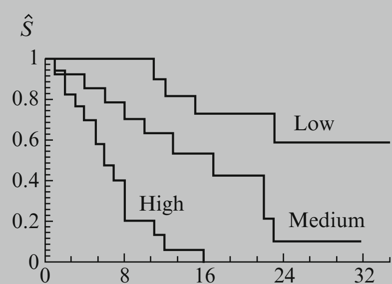
2. Use PH model:
   $h(t, \mathbf{X}) = h_0(t)\exp(\beta X)$
   so $\mathbf{X} = X$ ← Continuous
   Obtain adjusted survival curve:
   $$\hat{S}(t, \overline{X}_c) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}\overline{X}_c)}$$

where $\overline{X}_c$ denotes the mean value for the variable X within category c.

The second option is to use a Cox PH model containing the continuous predictor (say, X) being assessed. Expected plots are then obtained as adjusted survival curves by specifying predictor values that distinguish categories, as, for example, when using mean predictor values for each category.

Example: Remission Data

Observed (KM) Plots by log WBC Categories



Option 1:
$h(t, \mathbf{X}) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2)$

where $X_1 = \begin{cases} 1 & \text{if high} \\ 0 & \text{if other} \end{cases}$ $X_2 = \begin{cases} 1 & \text{if medium} \\ 0 & \text{if other} \end{cases}$

so that
high = (1, 0); medium = (0, 1); low = (0, 0)
Expected survival plots:

$X_1 = 1, X_2 = 0: \hat{S}(t, X_{\text{high}}) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}_1)}$

$X_1 = 0, X_2 = 1: \hat{S}(t, X_{\text{medium}}) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}_2)}$

$X_1 = 0, X_2 = 0: \hat{S}(t, X_{\text{low}}) = \left[\hat{S}_0(t)\right]$

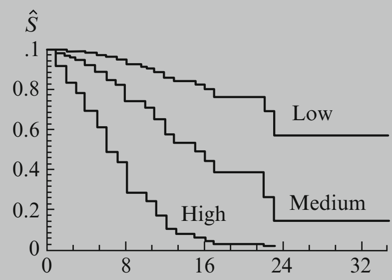
As an example to illustrate both options, we consider the continuous variable log WBC from the remission data example. To assess the PH assumption for this variable, we would first stratify log WBC into, say, three categories – low, medium, and high. The observed plots would then be obtained as KM curves for each of the three strata, as shown here.

Using option 1, expected plots would be obtained by fitting a Cox PH model containing two dummy variables X_1 and X_2, as shown here, where X_1 takes the values 1 if high or 0 if other and X_2 takes the values 1 if medium or 0 if other. Thus, when log WBC is high, the values of X_1 and X_2 are 1 and 0, respectively; whereas when log WBC is medium, the values are 0 and 1, respectively; and when log WBC is low, the values are both 0.

The expected survival plots for high, medium, and low categories are then obtained by substituting each of the three specifications of X_1 and X_2 into the formula for the estimated survival curve, and then plotting the three curves.

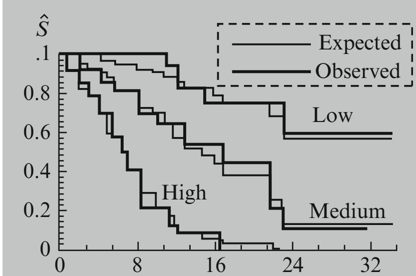
EXAMPLE: (continued)

Expected Plots for log WBC Using Option 1 (Dummy Variables)



The expected plots using option 1 (the dummy variable approach) are shown here for the three categories of log WBC.

Observed Versus Expected Plots Using Option 1



Here we put the observed and expected plots on the same graph. Although there are some discrepancies, particularly early in follow-up for the low log WBC category, these plots suggest overall that the PH assumption is satisfied for log WBC.

Option 2: Treat log WBC as continuous
$h(t, \mathbf{X}) = h_0(t)\exp[\beta(\log \text{WBC})]$

$\overline{\log \text{WBC}}_{\text{high}} = 3.83$:

$$\hat{S}(t, X_{\text{high}}) = [\hat{S}_0(t)]^{\exp[3.83\hat{\beta}]}$$

$\overline{\log \text{WBC}}_{\text{med}} = 2.64$:

$$\hat{S}(t, X_{\text{med}}) = [\hat{S}_0(t)]^{\exp[2.64\hat{\beta}]}$$

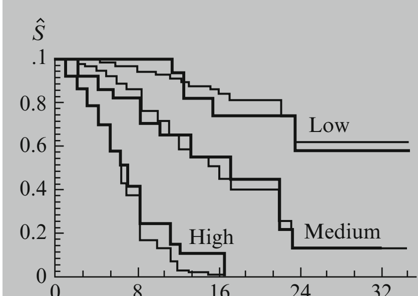
$\overline{\log \text{WBC}}_{\text{low}} = 1.71$:

$$\hat{S}(t, X_{\text{low}}) = [\hat{S}_0(t)]^{\exp[1.71\hat{\beta}]}$$

Using option 2, expected plots would be obtained by first fitting a Cox PH model containing the continuous variable log WBC, as shown here.

Adjusted survival curves are then obtained for specified values of log WBC that summarize the three categories used to form observed curves. Here, we find that the mean log WBC scores for low, medium, and high categories are, respectively, 1.71, 2.64, and 3.83. These values are substituted into the estimated survival curve formula as shown here.

Observed Versus Expected Plots for log WBC Using Option 2.



Here are the observed and expected plots using option 2. As with option 1, although there are some discrepancies within categories, overall, these plots suggest that the PH assumption is satisfied for the log WBC variable.

V. The Goodness of Fit (GOF) Testing Approach

Statistical test appealing

- Provides p-value
- More objective decision than when using graphical approach

The GOF testing approach is appealing because it provides a test statistic and p-value for assessing the PH assumption for a given predictor of interest. Thus, the researcher can make a more objective decision using a statistical test than is typically possible when using either of the two graphical approaches described above.

Test of Harrel and Lee (1986)

- Variation of test of Schoenfeld
- Uses Schoenfeld residuals

A number of different tests for assessing the PH assumption have been proposed in the literature. We present the test of Harrel and Lee (1986), a variation of a test originally proposed by Schoenfeld (1982) and based on the residuals defined by Schoenfeld, now called the Schoenfeld residuals.

Schoenfeld residuals defined for

- Each predictor in model
- Every subject who has event

Consider Cox PH model

$$h(t,\mathbf{X}) = h_0(t)\exp(\beta_1 RX + \beta_2 \log WBC + \beta_3 SEX)$$

3 predictors → 3 Schoenfeld residuals for each subject who has event

For each predictor in the model, Schoenfeld residuals are defined for every subject who has an event. For example, consider a Cox PH model with three predictors: RX, LOGWBC, and SEX. Then there are three Schoenfeld residuals defined for each subject who has an event, one for each of the three predictors.

Schoenfeld residual for ith subject for LOGWBC:

Observed LOGWBC
– LOGWBC weighted average

Weights are other subjects' hazard (from subjects still at risk)

Suppose subject i has an event at time t. Then her Schoenfeld residual for LOGWBC is her observed value of log white blood cell count minus a weighted average of the log white blood cell counts for the other subjects still at risk at time t. The weights are each subject's hazard.

Underlying idea of test
If PH holds then Schoenfeld residuals uncorrelated with time

The idea behind the statistical test is that **if the PH assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time**.

Steps for test implementation

1. Obtain Schoenfeld residuals
2. Rank failure times
3. Test correlation of residuals to ranked failure time H_0: $\rho = 0$

The implementation of the test can be thought of as a three-step process.

Step 1. Run a Cox PH model and obtain Schoenfeld residuals for each predictor.

Step 2. Create a variable that ranks the order of failures. The subject who has the first (earliest) event gets a value of 1, the next gets a value of 2, and so on.

Step 3. Test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is zero.

H_0 rejected
Conclude PH assumption violated

Rejection of the null hypothesis leads to a conclusion that the PH assumption is violated.

PH test in Stata, SAS, SPSS, R shown in Computer Appendix

The implementation of the test for the PH assumption in Stata, SAS, SPSS, and R is shown in the Computer Appendix. Stata uses a slight variation of the test we just described in that it uses the scaled Schoenfeld residual rather than the Schoenfeld residual (Grambsch and Therneau, 1994). The tests typically (but not always) yield similar results.

Stata uses scaled Schoenfeld residuals rather than Schoenfeld residuals (typically similar results)

EXAMPLE: Remission Data

Column name.	Coeff.	StErr.	*P(PH)*
Rx	1.294	0.422	0.917
log WBC	1.604	0.329	0.944

Both variables satisfy PH assumption.

Note: P(PH) = 0.917 assesses PH for *Rx*, assuming PH OK for log WBC.

To illustrate the statistical test approach, we return to the remission data example. The printout on the left gives p-values P(PH) for treatment group and log WBC variables based on fitting a Cox PH model containing these two variables.

The P(PH) values are quite high for both variables, suggesting that both variables satisfy the PH assumption. Note that each of these p-values tests the assumption for one variable given that the other predictors are included in the model. For example, the P(PH) of 0.917 assesses the PH assumption for Rx, assuming the PH assumption is satisfied for log WBC.

EXAMPLE

Column name	Coeff.	StErr.	$P(PH)$
Rx	1.391	0.457	0.935
log WBC	1.594	0.330	0.828
Sex	0.263	0.449	0.038

log WBC and Rx satisfy PH.

Sex does not satisfy PH.

(Same conclusions using graphical approaches).

As another example, consider the computer results shown here for a Cox PH model containing the variable SEX in addition to log WBC and treatment group. The P(PH) values for log WBC and treatment group are still nonsignificant. However, the P(PH) value for SEX is significant below the 0.05 level. This result suggests that log WBC and treatment group satisfy the PH assumption, whereas SEX does not. We came to the same conclusion about these variables using the graphical procedures described earlier.

Statistical Tests

Null is never proven

- May say not enough evidence to reject

p-value can be driven by sample size

- Small sample – gross violation of null may not be significant
- Large sample – slight violation of null may be highly significant

An important point concerning a testing approach is that the null hypothesis is never proven with a statistical test. The most that may be said is that there is not enough evidence to reject the null. A p-value can be driven by sample size. A gross violation of the null assumption may not be statistically significant if the sample is very small. Conversely, a slight violation of the null assumption may be highly significant if the sample is very large.

Test – more objective
Graph – more subjective, but can detect specific violations
Recommend – Use both graphs and tests

A statistical test offers a more objective approach for assessing the PH assumption compared to the subjectivity of the graphical approach. However, the graphical approach enables the researcher to detect specific kinds of departures from the PH assumption; the researcher can see what is going on from the graph. Consequently, we recommend that when assessing the PH assumption, the investigator use both graphical procedures and statistical testing before making a final decision.

VI. Assessing the PH Assumption Using Time-Dependent Covariates

Extended Cox model:
contains product terms of the form $X \times g(t)$, where $g(t)$ is a function of time.

When time-dependent variables are used to assess the PH assumption for a time-independent variable, the Cox model is extended to contain product (i.e., interaction) terms involving the time-independent variable being assessed and some function of time.

One-at-a-time model:

$h(t,\mathbf{X}) = h_0(t)\exp[\beta X + \delta(X \times g(t))]$

When assessing predictors one-at-a-time, the extended Cox model takes the general form shown here for the predictor X.

Some choices for $g(t)$:

$g(t) = t$

$g(t) = \log t$

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases} \quad \text{(heaviside function)}$$

One choice for the function $g(t)$ is simply $g(t)$ equal to t, so that the product term takes the form $X \times t$. Other choices for $g(t)$ are also possible, for example, $\log t$, or the "heaviside function" shown at the left.

H_0: $\delta = 0$

Under H_0, the model reduces to:

$h(t,\mathbf{X}) = h_0(t)\exp[\beta X]$

Using the above one-at-a-time model, we assess the PH assumption by testing for the significance of the product term. The null hypothesis is therefore "δ equal to zero." Note that if the null hypothesis is true, the model reduces to a Cox PH model containing the single variable X.

Use either Wald statistic or likelihood ratio statistic:

χ^2 with 1 df under H_0

The test can be carried out using either a Wald statistic or a likelihood ratio statistic. In either case, the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis.

Example

$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1\text{Sex} + \beta_2 (\text{Sex} \times t)]$

$\beta_2 \neq 0$ = PH assumption violated

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable Sex $\times$ t in addition to Sex. If the coefficient of the product term turns out to be significant, we can conclude that the PH assumption is violated for Sex.[2]

Strategies for assessing PH:

- one-at-a-time
- several predictors simultaneously
- for a given predictor adjusted for other predictors

In addition to a one-at-a-time strategy, the extended Cox model can also be used to assess the PH assumption for several predictors simultaneously as well as for a given predictor adjusted for other predictors in the model.

[2] In contrast, if the test for H_0: $\beta_2 = 0$ is nonsignificant, we can conclude only that the particular version of the extended Cox model being considered is not supported by the data.

Several predictors simultaneously:

$$h(t,\mathbf{X}) = h_0(t)\ \exp\left(\sum_{i=1}^{p}[\beta_i X_i + \delta_i(X_i \times g_i(t))]\right)$$

$g_i(t)$ = function of time for ith predictor

To assess the PH assumption for several predictors simultaneously, the form of the extended model is shown here. This model contains the predictors being assessed as main effect terms and also as product terms with some function of time. Note that different predictors may require different functions of time; hence, the notation $g_i\ (t)$ is used to define the time function for the ith predictor.

H_0: $\delta_1 = \delta_2 = \ldots = \delta_p = 0$

$$LR = -2 \ln L_{\text{PH model}} - (-2 \ln L_{\text{ext. Cox model}})$$
$$\dot{\sim}\chi^2_{\text{p}} \text{ under } H_0$$

Cox PH (reduced) model:

$$h(t,\mathbf{X}) = h_0(t)\ \exp\left(\sum_{i=1}^{p} \beta_i X_i\right)$$

With the above model, we test for the PH assumption simultaneously by assessing the null hypothesis that all the δ_i coefficients are equal to zero. This requires a likelihood ratio chi-square statistic with p degrees of freedom, where p denotes the number of predictors being assessed. The LR statistic computes the difference between the log likelihood statistic (i.e., $-2\ \ln\ L$) for the PH model and the log likelihood statistic for the extended Cox model. Note that under the null hypothesis, the model reduces to the Cox PH model shown here.

EXAMPLE: Remission Data

$$\begin{aligned} h(t,X) = h_0(t)\exp[&\ \beta_1(Rx) \\ &+ \beta_2(\text{log WBC}) + \beta_3(\text{Sex}) \\ &+ \delta_1(Rx \times g(t)) \\ &+ \delta_2(\text{log WBC} \times \text{g}(t)) \\ &+ \delta_3(\text{Sex} \times g(t))] \end{aligned}$$

$$\text{where } g(t) = \begin{cases} 1 & \text{if t} \geq 7 \\ 0 & \text{if t} < 7 \end{cases}$$

H_0: $\delta_1 = \delta_2 = \delta_3 = 0$

$LR \dot{\sim} \chi^2$ with 3 df under H_0

If test is significant, use backward elimination to find predictors not satisfying PH assumption.

As an example, we assess the PH assumption for the predictors Rx, log WBC, and Sex from the remission data considered previously. The extended Cox model is given as shown here, where the functions $g_i\ (t)$ have been chosen to be the same "heaviside" function defined by $g(t)$ equals 1 if t is 7 weeks or more and $g(t)$ equals 0 if t is less than 7 weeks. The null hypothesis is that all three δ coefficients are equals to zero. The test statistic is a likelihood-ratio chi-square with 3 degrees of freedom.

If the above test is found to be significant, then we can conclude that the PH assumption is not satisfied for at least one of the predictors in the model. To determine which predictor(s) do not satisfy the PH assumption, we could proceed by backward elimination of nonsignificant product terms until a final model is attained.

Heaviside function:

$$g(t) = \begin{cases} 1 & \text{if } t \geq 7 \\ 0 & \text{if } t < 7 \end{cases}$$

$h(t,\mathbf{X})$ differs for $t \geq 7$ and $t < 7$.

Properties of heaviside functions and numerical results are described in Chapter 6.

Note that the use of a heaviside function for $g(t)$ in the above example yields different expressions for the hazard function depending on whether t is greater than or equal to 7 weeks or t is less than 7 weeks. Chapter 6 provides further details on the properties of heaviside functions, and also provides numerical results from fitting extended Cox models.

Assessing PH for a given predictor adjusted for other predictors:

$$h(t,\mathbf{X}) = h_0(t)\exp\left[\sum_{i=1}^{p-1} \beta_i X_i + \beta^* X^* + \delta^*\left(X^* \times g(t)\right)\right]$$

X^* = Predictor of interest
H_0: $\delta^* = 0$
Wald or LR statistic $\dot{\sim}\chi^2$ with 1 df

We show here an extended Cox model that can be used to evaluate the PH assumption for a given predictor **adjusted for predictors already satisfying the PH assumption**. The predictor of interest is denoted as X^*, and the predictors considered to satisfy the PH assumption are denoted as X_i. The null hypothesis is that the coefficient δ^* of the product term $X^*g(t)$ is equal to zero. The test statistic can either be a Wald statistic or a likelihood ratio statistic, with either statistic having a chi-square distribution with 1 degree of freedom under the null hypothesis.

Example: Remission Data

for Sex, adjusted for Rx and log WBC:

$$h(t,\mathbf{X}) = \exp[\beta_1 Rx + \beta_2 \log \text{ WBC} + \beta^* \text{Sex} + \delta^*(\text{Sex} \times g(t))]$$

As an example, suppose, again considering the remission data, we assess the PH assumption for the variable, Sex, adjusted for the variables Rx and log WBC, which we assume already satisfy the PH assumption. Then, the extended Cox model for this situation is shown here.

Two models for LR test of PH:

1. Cox PH model
2. extended Cox model

See Computer Appendix for SAS, Stata, SPSS, and R

To carry out the computations for any of the likelihood ratio tests described above, two different types of models, a PH model and an extended Cox model, need to be fit. See the Computer Appendix for details on how the extended Cox model is fit using SAS, Stata, SPSS, and R.

Drawback: choice of $g_i(t)$

Different choices may lead to different conclusions about PH assumption.

The primary drawback of the use of an extended Cox model for assessing the PH assumption concerns the choice of the functions $g_i(t)$ for the time-dependent product terms in the model. This choice is typically not clear-cut, and it is possible that different choices, such as $g(t)$ equal to t versus log t versus a heaviside function, may result in different conclusions about whether the PH assumption is satisfied.

Chapter 6: Time-dependent covariates

Further discussion of the use of time-dependent covariates in an extended Cox model is provided in Chapter 6.

This presentation:
Three methods for assessing PH.

i. graphical
ii. GOF
iii. time-dependent covariates

Recommend using at least two methods.

This presentation is now complete. We have described and illustrated three methods for assessing the PH assumption: graphical, goodness-of-fit (GOF), and time-dependent covariate methods. Each of these methods has both advantages and drawbacks. We recommend that the researcher use at least two of these approaches when assessing the PH assumption.

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Survival Curves and the Log–Rank Test
3. The Cox Proportional Hazards Model and Its Characteristics

✓4. $\boxed{\text{Evaluating the Proportional Hazards Assumption}}$

Next:

5. The Stratified Cox Procedure
6. Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

We suggest that the reader review this presentation using the detailed outline that follows. Then answer the practice exercises and the test that follow.

The next Chapter (5) is entitled "The Stratified Cox Procedure." There, we describe how to use a stratification procedure to fit a PH model when one or more of the predictors do not satisfy the PH assumption.

Detailed Outline

I. Background (pages 164–165)

A. The formula for the Cox PH model:

$$\boxed{h(t,\mathbf{X}) = h_0(t)\ \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]}$$

B. Formula for hazard ratio comparing two individuals,

$\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_p^*\right)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

$$\boxed{\frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp\left[\sum_{i=1}^{p} \beta_i\left(X_i^* - X_i\right)\right]}$$

C. Adjusted survival curves using the Cox PH model:

$$S(t,\mathbf{X}) = [S_0(t)]^{\exp[\sum \beta_1 X_1]}$$

i. To graph $S\ (t,\ \mathbf{X})$, must specify values for $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

ii. To obtain "adjusted" survival curves, usually use overall mean values for the X's being adjusted.

D. The meaning of the PH assumption

i. Hazard ratio formula shows that hazard ratio is independent of time:

$$\frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \hat{\theta}$$

ii. Hazard ratio for two X's are proportional:

$$\hat{h}(t,\mathbf{X}^*) = \hat{\theta}\hat{h}(t,\mathbf{X})$$

II. Checking the proportional hazards assumption: overview (pages 165–167)

A. Three methods for checking the PH assumption:

i. Graphical: compare –ln –ln survival curves or observed versus predicted curves.

ii. Goodness-of-fit test: use a large sample Z statistic.

iii. Time-dependent covariates: use product (i.e., interaction) terms of the form $X \times g(t)$.

B. Abbreviated illustrations of each method are provided.

III. Graphical approach 1: log–log plots (pages 167–175)

A. A log–log curve is a transformation of an estimated survival curve, where the scale for a log–log curve is $-\infty$ to $+\infty$.

B. The log–log expression for the Cox model survival curve is given by

$$\ln[-\ln \mathcal{S}(t, \mathbf{X})] = \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln S_0(t)]$$

C. For the Cox model, the log–log survival curve for individual $\mathbf{X}_1$ can be written as the log–log curve for individual $\mathbf{X}_2$ plus a linear sum term that is independent of time t. This formula is given by

$$\ln[-\ln \mathcal{S}(t, \mathbf{X}_1)] = \ln[-\ln \mathcal{S}(t, \mathbf{X}_2)] + \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})$$

D. The above log–log formula can be used to check the PH assumption as follows: the PH model is appropriate if "empirical" plots of log–log survival curves are parallel.

E. Two kinds of empirical plots for $-\ln -\ln \hat{\mathcal{S}}$:
   i. $\hat{\mathcal{S}}$ is a KM curve
   ii. $\hat{\mathcal{S}}$ is an adjusted survival curve where predictor being assessed is not included in the Cox regression model.

F. Several examples of log–log plots are provided using remission data from a clinical trial of leukemia patients.

G. Problems with log–log curves:
   i. How parallel is parallel?
   ii. How to categorize a continuous variable?
   iii. How to evaluate several variables simultaneously?

H. Recommendation about problems:
   i. Use small number of categories, meaningful choice, reasonable balance.
   ii. With several variables, two options:
      a. Compare log–log curves from combinations of categories.
      b. Adjust for predictors already satisfying PH assumption.

IV. **Graphical approach 2: observed versus expected plots** (pages 175–180)

A. Graphical analog of the GOF test.

B. Two strategies
   i. One-at-a-time: uses KM curves to obtain observed plots.
   ii. Adjusting for other variables: uses stratified Cox PH model to obtain observed plots.

C. Expected plots obtained by fitting a Cox model containing the predictor being assessed; substitute into the fitted model the value for each category of the predictor to obtain the expected value for each category.
D. If observed and expected plots are close, conclude PH assumption is reasonable.
E. Drawback: how close is close?
F. Recommend: conclude PH not satisfied *only* if plots are strongly discrepant.
G. Another drawback: what to do if assessing continuous variable.
H. Recommend for continuous variable:
   i. Form strata from categories.
   ii. Observed plots are KM curves for each category.
   iii. Two options for expected plots:
      a. Use PH model with $k - 1$ dummy variables for k categories.
      b. Use PH model with continuous predictor and specify predictor values that distinguish categories.

V. The goodness-of-fit (GOF) testing approach (pages 181–183)

A. Appealing approach because
   i. provides a test statistic (p-value).
   ii. researcher can make clear-cut decision.
B. References
   i. methodological: Schoenfeld (1982), Harrel and Lee (1986).
   ii. SAS and Stata use different GOF formulae.
C. The method:
   i. Schoenfeld residuals for each predictor uses a chi-square statistic with 1 df.
   ii. Correlations between Schoenfeld's residuals and ranked failure times.
   iii. If p-value small, then departure from PH.
D. Examples using remission data.
E. Drawbacks:
   i. global test: may fail to detect a specific kind of departure from PH; recommend using both graphical and GOF methods.
   ii. several strategies to choose from, with no one strategy clearly preferable (one-at-a-time, all variables, each variable adjusted for others).

VI. Assessing the PH assumption using time-dependent covariates (pages 183–187)

A. Use extended Cox model: contains product terms of form $X \times g(t)$, where $g(t)$ is function of time, e.g., $g(t) = t$, or $\log t$, or heaviside function.

B. One-at-a-time model:
$h(t, \mathbf{X}) = h_0(t) \exp[\beta X + \delta X g(t)]$
Test H_0: $\delta = 0$ using Wald or LR test (chi- square with 1 df).

C. Evaluating several predictors simultaneously:

$$h(t, \mathbf{X}) = h_0(t) \exp\left(\sum_{i=1}^{p} [\beta_i X_i + \delta_i X_i g_i(t)]\right)$$

where $g_i(t)$ is function of time for ith predictor.
Test H_0: $\delta_1 = \delta_2 = \cdots = \delta_p = 0$ using LR (chi-square) test with p df.

D. Examples using remission data.

E. Two computer programs, required for test:
   i. Cox PH model program.
   ii. Extended Cox model program.

F. Drawback: choice of $g(t)$ not always clear; different choices may lead to different conclusions about PH assumption.

Practice Exercises

The dataset "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (*The Statistical Analysis of Survival Time Data*, Wiley, 2002). The exposure variable of interest is treatment status (standard = 1, test = 2). Other variables of interest as control variables are cell type (four types, defined by dummy variables), performance status, disease duration, age, and prior therapy status. Failure status is defined by the status variable (0 if censored, 1 if died). A complete list of the variables is given below.

Column 1: Treatment (standard = 1, test = 2)
Column 2: Cell type 1 (large = 1, other = 0)
Column 3: Cell type 2 (adeno = 1, other = 0)
Column 4: Cell type 3 (small = 1, other = 0)
Column 5: Cell type 4 (squamous = 1, other = 0)
Column 6: Survival time (days)
Column 7: Performance status (0 = worst,. . ., 100 = best)

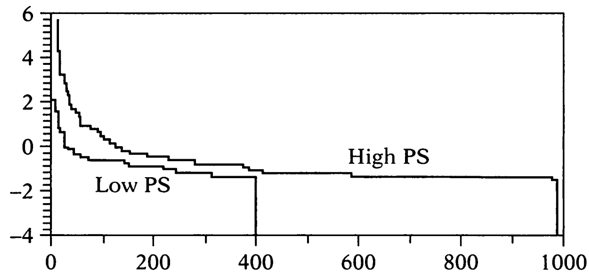
Column 8: Disease duration (months)
Column 9: Age
Column 10: Prior therapy (none = 0, some = 10)
Column 11: Status (0 = censored, 1 = died)

1. State the hazard function form of the Cox PH model that describes the effect of the treatment variable and controls for the variables, cell type, performance status, disease duration, age, and prior therapy. In stating this model, make sure to incorporate the cell type variable using dummy variables, but do not consider possible interaction variables in your model.
2. State three general approaches that can be used to evaluate whether the PH assumption is satisfied for the variables included in the model you have given in question 1.
3. The following printout is obtained from fitting a Cox PH model to these data. Using the information provided, what can you conclude about whether the PH assumption is satisfied for the variables used in the model? Explain briefly.

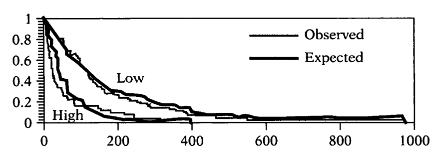
Cox regression	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]	*P*(*PH*)
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Performance status	−0.033	0.006	0.000	0.968	0.958	0.978	0.000
Disease duration	0.000	0.009	0.992	1.000	0.982	1.018	0.919
Age	−0.009	0.009	0.358	0.991	0.974	1.010	0.198
Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.145

4. For the variables used in the PH model in question 1, describe a strategy for evaluating the PH assumption using log–log survival curves for variables considered one-at-a-time.
5. Again considering the variables used in question 1, describe a strategy for evaluating the PH assumption using log–log survival curves that are adjusted for other variables in the model.

6. For the variable "performance status," describe how you would evaluate the PH assumption using observed versus expected survival plots?
7. For the variable "performance status," log–log plots which compare high ($\geq$50) with low ($<$50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



8. What are some of the drawbacks of using the log–log approach for assessing the PH assumption and what do you recommend to deal with these drawbacks?
9. For the variable "performance status," observed versus expected plots that compare high ($\geq$50) with low ($<$50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



10. State the form of an extended Cox model that allows for the one-at-a-time assessment of the PH assumption for the variable "performance status," and describe how you would carry out a statistical test of the assumption for this variable.
11. State the form of an extended Cox model that allows for the simultaneous assessment of the PH assumption for the variables treatment, cell type, performance status, disease duration, age, and prior therapy. For this model, describe how you would carry out a statistical test of the PH assumption for these variables. Also, provide a strategy for assessing which of these variables satisfy the PH assumption and which do not using the extended Cox model approach.

12. Using any of the information provided above and any additional analyses that you perform with this dataset, what do you conclude about which variables satisfy the PH assumption and which variables do not? In answering this question, summarize any additional analyses performed.

Test

The following questions consider a dataset from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," *Med. J. Aust.*, 1991). These data comprise the times in days spent by heroin addicts from entry to departure from one of two methadone clinics. There are two additional covariates, namely, prison record and maximum methadone dose, believed to affect the survival times. The dataset name is **addicts.dat**. A listing of the variables is given below:

Column 1: Subject ID
Column 2: Clinic (1 or 2)
Column 3: Survival status (0 = censored, 1 = departed from clinic)
Column 4: Survival time in days
Column 5: Prison record (0 = none, 1 = any)
Column 6: Maximum methadone dose (mg/day)

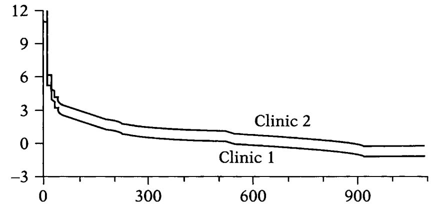
1. The following edited printout was obtained from fitting a Cox PH model to these data:

Cox regression Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]		*P*(*PH*)
Clinic	−1.009	0.215	0.000	0.365	0.239	0.556	0.001
Prison	0.327	0.167	0.051	1.386	0.999	1.924	0.332
Dose	−0.035	0.006	0.000	0.965	0.953	0.977	0.347

No. of subjects: 238 Log likelihood = −673.403

Based on the information provided in this printout, what do you conclude about which variables satisfy the PH assumption and which do not? Explain briefly.

2. Suppose that for the model fit in question 1, log–log survival curves for each clinic adjusted for prison and dose are plotted on the same graph. Assume that these curves are obtained by substituting into the formula for the estimated survival curve the values for each clinic and the overall mean values for the prison and dose variables. Below, we show these two curves. Are they parallel? Explain your answer.

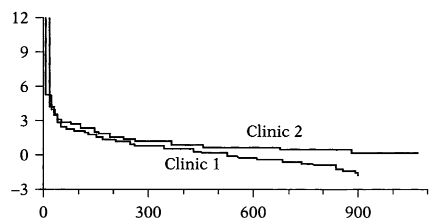


3. The following printout was obtained from fitting a stratified Cox PH model to these data, where the variable being stratified is clinic:

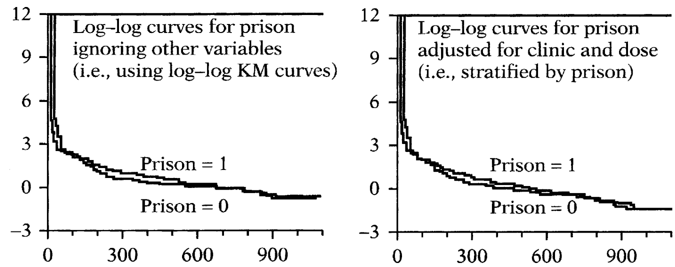
Stratified Cox regression Analysis time.t: survt (in days)	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Prison	0.389	0.169	0.021	1.475	1.059	2.054
Dose	−0.035	0.006	0.000	0.965	0.953	0.978

No. of subjects = 238 Log likelihood = −597.714 Stratified by clinic

Using the above fitted model, we can obtain the log–log curves below that compare the log–log survival for each clinic (i.e., stratified by clinic) adjusted for the variables prison and dose. Using these curves, what do you conclude about whether or not the clinic variable satisfies the PH assumption? Explain briefly.



4. Consider the two plots of log–log curves below that compare the log–log survival for the prison variable ignoring other variables and adjusted for the clinic and dose variables. Using these curves, what do you conclude about whether or not the prison variable satisfies the PH assumption? Explain briefly.



5. How do your conclusions from question 1 compare with your conclusions from question 4? If the conclusions differ, which conclusion do you prefer? Explain.
6. Describe briefly how you would evaluate the PH assumption for the variable maximum methadone dose using observed versus expected plots.
7. State an extended Cox model that would allow you to assess the PH assumption for the variables clinic, prison, and dose simultaneously. For this model, state the null hypothesis for the test of the PH assumption and describe how the likelihood ratio statistic would be obtained and what its degrees of freedom would be under the null hypothesis.
8. State at least one drawback to the use of the extended Cox model approach described in question 7.
9. State an extended Cox model that would allow you to assess the PH assumption for the variable clinic alone, assuming that the prison and dose variables already satisfy the PH assumption. For this model, state the null hypothesis for the test of the PH assumption, and describe how the likelihood ratio (LR) statistic would be obtained. What is the degrees of freedom of the LR test under the null hypothesis?

10. Consider the situation described in question 9, where you wish to use an extended Cox model that would allow you to assess the PH assumption for the variable clinic alone, assuming that the assumption is satisfied for the prison and dose variables. Suppose you use the following extended Cox model:

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1(\text{prison}) + \beta_2(\text{dose}) + \beta_3(\text{clinic}) + \delta_1(\text{clinic})g(t)]$$

where $g(t)$ is defined as follows:

$$g(t) = \begin{cases} 1 & \text{if } t > 365 \text{ days} \\ 0 & \text{if } t \leq 365 \text{ days} \end{cases}$$

For the above model, what is the formula for the hazard ratio that compares clinic 1 to clinic 2 when t is greater than 365 days? when t is less than or equal to 365 days? In terms of the hazard ratio formulae just described, what specific departure from the PH assumption is being tested when the null hypothesis is H_0: $\delta_1 = 0$?

Answers to Practice Exercises

1. $h(t,\mathbf{X}) = h_0(t) \exp[\beta_1(\text{treatment}) + \beta_2(\text{CT1}) + \beta_3(\text{CT2}) + \beta_4(\text{CT3}) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT})]$

   where *CTi* denotes the cell type *i* dummy variable, PS denotes the performance status variable DD denotes the disease duration variable, and PT denotes the prior therapy variable.

2. The three general approaches for assessing the PH model for the above model are:
   (a) graphical, using either log–log plots or observed versus expected plots;
   (b) statistical test;
   (c) an extended Cox model containing product terms involving the variables being assessed with some function(s) of time.

3. The $P(PH)$ values given in the printout provide goodness-of-fit tests for each variable in the fitted model adjusted for the other variables in the model. The $P(PH)$ values shown indicate that the large cell type variables and the performance status variable do not satisfy the PH assumption, whereas the treatment, age, disease duration, and prior therapy variables satisfy the PH assumption, and the adeno and small cell type variable are of borderline significance.

4. A strategy for evaluating the PH assumption using log–log survival curves for variables considered one-at-a-time is given as follows:

   For each variable separately, obtain a plot of log–log Kaplan–Meier curves for the different categories of that variable. For the cell type variable, this requires obtaining a plot of four log–log KM curves, one for each cell type. (Note that this is not the same as obtaining four separate plots of two log–log curves, where each plot corresponds to one of the dummy variables used in the model.) For the variables PS, DD, and Age, which are interval variables, each variable must be separately categorized into two or more groups – say, low versus high values – and KM curves are obtained for each group. For the variable PT, which is a dichotomous variable, two log–log curves are obtained which compare the "none" versus "some" groups.

   For each plot (i.e., one for each variable), those plots that are noticeably nonparallel indicate variables which do not satisfy the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

5. One strategy for evaluating the PH assumption for each variable adjusted for the others is to use adjusted log–log survival curves instead of KM curves separately for each of the variables in the model. That is, for each variable separately, a stratified Cox model is fit stratifying on the given variable while adjusting for the other variables. Those variables that yield adjusted log–log plots that are noticeably nonparallel are then to be considered as not satisfying the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

   A variation of the above strategy uses adjusted log–log curves for only those variables *not satisfying* the PH assumption from a one-at-a-time approach, adjusting for those variables *satisfying* the PH assumption from the one-at-a-time approach. This second iteration would flag a subset of the one-at-a-time flagged variables for further iteration. At each new iteration, those variables found to satisfy the assumption get added to the list of variables previously determined to satisfy the assumption.

6. For the performance status (PS) variable, **observed plots** are obtained by categorizing the variable into strata (say, two strata: low versus high) and then obtaining KM survival plots for each stratum. **Expected plots** can be obtained by fitting a Cox

model containing the (continuous) PS variable and then obtaining estimated survival curves for values of the performance status (PS) variable that represent summary descriptive statistics for the strata previously identified. For example, if there are two strata, say, high (PS $>$ 50) and low (PS $\leq$ 50), then the values of PS to be used could be the mean or median PS score for persons in the high stratum and the mean or median PS score for persons in the low stratum.

An alternative method for obtaining expected plots involves first dichotomizing the PS variable – say, into high and low groups – and then fitting a Cox model containing the dichotomized PS variable instead of the original continuous variable. The expected survival plots for each group are estimated survival curves obtained for each value of the dichotomized PS variable.

Once observed and expected plots are obtained for each stratum of the PS variable, they are then compared on the same graph to determine whether or not corresponding observed and expected plots are "close." If it is determined that, overall, comparisons for each stratum are close, then it is concluded that the PH assumption is satisfied for the PH variable. In determining how close is close, the researcher should look for noticeably discrepant observed versus expected plots.

7. The log–log plots that compare high versus low PS groups (ignoring other variables) are *arguably* parallel early in follow-up, and are not comparable later because survival times for the two groups do not overlap after 400 days. These plots do not strongly indicate that the PH assumption is violated for the variable PS. This contradicts the conclusion previously obtained for the PS variable using the $P(PH)$ results.

8. Drawbacks of the log–log approach are:

   - How parallel is parallel?
   - How to categorize a continuous variable?
   - How to evaluate several variables simultaneously?

   Recommendations about problems:

   - Look for noticeable nonparallelism; otherwise PH assumption is OK.
   - For continuous variables, use a small number of categories, a meaningful choice of categories, and a reasonable balance in sample size for categories.

- With several variables, there are two options:

  i. Compare log–log curves from combinations of categories.
  ii. Adjust for predictors already satisfying PH assumption.

9. The observed and expected plots are relatively close for low and high groups separately, although there is somewhat more discrepancy for the high group than for the low group. Deciding how close is close is quite subjective for these plots. Nevertheless, because there are no major discrepancies for either low or high groups, we consider the PH assumption satisfied for this variable.

10. $h(t,\mathbf{X}) = h_0(t)\ \exp[\beta_1(\text{PS}) + \delta(\text{PS})g(t)]$

    where $g(t)$ is a function of t, such as $g(t) = t$, or $g(t) = \log t$, or a heaviside function. The PH assumption is tested using a 1 df Wald or LR statistic for H_0: $\delta = 0$.

11. $h(t,\mathbf{X}) = h_0(t)\ \exp[\beta_1(\text{treatment}) + \beta_2(\text{CT1}) + \beta_3(\text{CT2}) + \beta_4(\text{CT3}) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT}) + \delta_1(\text{treatment} \times g(t)) + \delta_2(\text{CT1} \times g(t)) + \delta_3(\text{CT2} \times g(t)) + \delta_4(\text{CT3} \times g(t)) + \delta_5(\text{PS} \times g(t)) + \delta_6(\text{DD} \times g(t)) + \delta_7(\text{Age} \times g(t)) + \delta_8(\text{PT} \times g(t))]$

    where $g(t)$ is some function of time, such as $g(t) = t$, or $g(t) = \log t$, or a heaviside function. To test the PH assumption simultaneously for all variables, the null hypothesis is stated as H_0: $\delta_1 = \delta_2 = \ldots = \delta_8 = 0$. The test statistic is a likelihood-ratio statistic of the form

    $LR = -2 \ln LR - (-2 \ln L_F)$

    where R denotes the reduced (PH) model obtained when all δ's are 0, and F denotes the full model given above. Under H_0, the LR statistic is approximately chi-square with 8 df.

12. The question here is somewhat open-ended, leaving the reader the option to explore additional graphical, GOF, or extended Cox model approaches for evaluating the PH assumption for the variables in the model. The conclusions from the GOF statistics provided in question 3 are likely to hold up under further scrutiny, so that a reasonable conclusion is that cell type and performance status variables do not satisfy the PH assumption, with the remaining variables satisfying the assumption.