3 The Cox Proportional Hazards Model and Its Charact-eristics

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_3,

Introduction

We begin by discussing some computer results using the Cox PH model, without actually specifying the model; the purpose here is to show the similarity between the Cox model and standard linear regression or logistic regression.

We then introduce the Cox model and describe why it is so popular. In addition, we describe its basic properties, including the meaning of the proportional hazards assumption and the Cox likelihood. We also describe how and why we might consider using "age as the time scale" instead of "time-on follow-up" as the outcome variable.

Abbreviated Outline

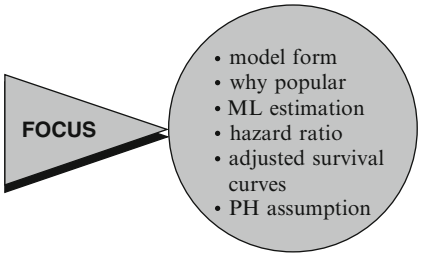
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

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

1. State or recognize the general form of the Cox PH model.
2. State the specific form of a Cox PH model appropriate for the analysis, given a survival analysis scenario involving one or more explanatory variables.
3. State or recognize the form and properties of the baseline hazard function in the Cox PH model.
4. Give three reasons for the popularity of the Cox PH model.
5. State the formula for a designated hazard ratio of interest given a scenario describing a survival analysis using a Cox PH model, when
   a. there are confounders but no interaction terms in the model;
   b. there are both confounders and interaction terms in the model.
6. State or recognize the meaning of the PH assumption.
7. Determine and explain whether the PH assumption is satisfied when the graphs of the hazard functions for two groups cross each other over time.
8. State or recognize what is an adjusted survival curve.
9. Compare and/or interpret two or more adjusted survival curves.
10. Given a computer printout involving one or more fitted Cox PH models,
    a. compute or identify any hazard ratio(s) of interest;
    b. carry out and interpret a designated test of hypothesis;
    c. carry out, identify or interpret a confidence interval for a designated hazard ratio;
    d. evaluate interaction and confounding involving one or more covariates.
11. Give an example of how the Cox PH likelihood is formed.
12. Given left truncated survival data, describe how and when you would consider using "age as the time scale" instead of "time-on follow-up" as the outcome variable.
13. Given left-truncated survival data, state the hazard function formula that uses "age as the time scale" as the outcome variable.
14. Illustrate the difference between an "open cohort" and a "closed cohort".

Presentation



This presentation describes the Cox proportional hazards (PH) model, a popular mathematical model used for analyzing survival data. Here, we focus on the model form, why the model is popular, maximum likelihood (ML) estimation of the model parameters, the formula for the hazard ratio, how to obtain adjusted survival curves, and the meaning of the PH assumption.

I. A Computer Example Using the Cox PH Model

We introduce the Cox PH model using computer output from the analysis of remission time data (Freireich et al., *Blood*, 1963), which we previously discussed in Chapters 1 and 2. The data set is listed here at the left.

EXAMPLE

Leukemia Remission Data

Group 1($n = 21$)		Group 2($n = 21$)	
t (weeks)	log WBC	t (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

+ denotes censored observation

These data involve two groups of leukemia patients, with 21 patients in each group. Group 1 is the treatment group, and group 2 is the placebo group. The data set also contains the variable log WBC, which is a well-known prognostic indicator of survival for leukemia patients.

For this example, the basic question of interest concerns comparing the survival experience of the two groups adjusting for the possible confounding and/or interaction effects of log WBC.

EXAMPLE: (continued)

T = weeks until going out of remission
X_1 = group status = E
X_2 = log WBC (confounding?)

Interaction?
$X_3 = X_1 \times X_2$ = group status × log WBC

Computer results for three Cox PH models using the Stata package

Other computer packages provide similar information.

Computer Appendix: uses Stata, SAS, and SPSS on the same dataset.

We are thus considering a problem involving two explanatory variables as predictors of survival time T, where T denotes "weeks until going out of remission." We label the explanatory variables X_1 (for group status) and X_2 (for log WBC). The variable X_1 is the primary study or exposure variable of interest. The variable X_2 is an extraneous variable that we are including as a possible confounder or effect modifier.

Note that if we want to evaluate the possible interaction effect of log WBC on group status, we would also need to consider a third variable, that is, the product of X_1 and X_2.

For this dataset, the computer results from fitting three different Cox proportional hazards models are presented below. The computer package used is Stata. This is one of several packages that have procedures for carrying out a survival analysis using the Cox model. The information printed out by different packages will not have exactly the same format, but they will provide similar information. A comparison of output using Stata, SAS, SPSS, and R procedures on the same dataset is provided in the computer appendix at the back of this text.

Edited Output From Stata:

Model 1:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.509	0.410	3.68	0.000	4.523	2.027	10.094
No. of subjects = 42		Log likelihood = −86.380			Prob > chi2 = 0.0001		

Model 2:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486
No. of subjects = 42		Log likelihood = −72.280			Prob > chi2 = 0.0000		

Model 3:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	2.355	1.681	1.40	0.161	10.537	0.391	284.201
log WBC	1.803	0.447	4.04	0.000	6.067	2.528	14.561
Rx x log WBC	−0.342	0.520	−0.66	0.510	0.710	0.256	1.967
No. of subjects = 42		Log likelihood = −72.066			Prob > chi2 = 0.0000		

EDITED OUTPUT FROM STATA

Model 1:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

No. of subjects = 42 Log likelihood = −86.380

Hazard ratios

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

No. of subjects = 42 Log likelihood = −72.066

We now describe how to use the computer printout to evaluate the possible effect of treatment status on remission time adjusted for the potential confounding and interaction effects of the covariate log WBC. For now, we focus only on five columns of information provided in the printout, as presented at the left for all three models.

For each model, the first column identifies the **variables** that have been included in the model. The second column gives estimates of **regression coefficients** corresponding to each variable in the model. The third column gives **standard errors** of the estimated regression coefficients. The fourth column gives **p-values** for testing the significance of each coefficient. The fifth column, labeled as **Haz. Ratio**, gives e^{Coef} for each variable in each model.

Models 1 and 2: $e^{coef} = HR$
Model 3: HR formula more complicated

As we discuss later in this chapter, e^{Coef} gives an estimated **hazard ratio (HR)** for the effect of each variable adjusted for the other variables in a model (e.g. Models 1 and 2) without product terms. With product terms such as Rx × log WBC in Model 3, the hazard ratio formula is more complicated, as we also discuss later.

Except for the Haz. Ratio column, these computer results are typical of output found in standard linear regression printouts. As the printout suggests, we can analyze the results from a Cox model in a manner similar to the way we would analyze a linear regression model.

EXAMPLE: (continued)

Same dataset for each model
$n = 42$ subjects
T = time (weeks) until out of remission
Model 1: *Rx* only
Model 2: *Rx* and log WBC
Model 3: *Rx*, log WBC, and *Rx* × log WBC

We now distinguish among the output for the three models shown here. All three models are using the same remission time data on 42 subjects. The outcome variable for each model is the same: time in weeks until a subject goes out of remission. However, the independent variables are different for each model. Model 1 contains only the treatment status variable, indicating whether a subject is in the treatment or placebo group. Model 2 contains two variables, treatment status and log WBC. And model 3 contains an interaction term defined as the product of treatment status and log WBC.

EDITED OUTPUT: ML ESTIMATION

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

No. of subjects = 42 Log likelihood = −72.066

We now focus on the output for model 3. The method of estimation used to obtain the coefficients for this model, as well as the other two models, is maximum likelihood (ML) estimation. Note that a p-value of 0.510 is obtained for the coefficient of the product term for the interaction of treatment with log WBC. This p-value indicates that there is no significant interaction effect, so that we can drop the product term from the model and consider the other two models instead.

EXAMPLE: (continued)

$P = 0.510 : \dfrac{-0.342}{-0.520} = -0.66 = Z$ Wald statistic

LR statistic: uses Log likelihood = −72.066

−2 ln L (log likelihood statistic) = −2 × (−72.066) = 144.132

The p-value of 0.510 that we have just described is obtained by dividing the coefficient −0.342 of the product term by its standard error of 0.520, which gives −0.66, and then assuming that this quantity is approximately a standard normal or *Z* variable. This *Z* statistic is known as a **Wald statistic,** which is one of two test statistics typically used with ML estimates. The other test statistic, called the **likelihood ratio,** or *LR* statistic, makes use of the log likelihood statistic. The log likelihood statistic is obtained by multiplying the "Log likelihood" in the Stata output by −2 to get −2 ln L.

Edited Output

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280
−2 ln L = −2× (-72.280) = 144.550

We now look at the output for model 2, which contains two variables. The treatment status variable (*Rx*) represents the exposure variable of primary interest. The log WBC variable is being considered as a confounder. Our goal is to describe the effect of treatment status adjusted for log WBC. Note that for Model 2, −2 ln L equals 144.550.

EXAMPLE: (continued)

LR (interaction in model 3)
$= -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}})$

In general:
$LR = -2 \ln L_{\text{R}} - (-2 \ln L_{\text{F}})$

To use the likelihood ratio (LR) statistic to test the significance of the interaction term, we need to compute the difference between the log likelihood statistic of the reduced model which does not contain the interaction term (model 2) and the log likelihood statistic of the full model containing the interaction term (model 3). In general, the LR statistic can be written in the form $-2 \ln L_R$ minus $-2 \ln L_F$, where *R* denotes the reduced model and *F* denotes the full model.

EXAMPLE: (continued)

LR (interaction in model 3)
$= -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}})$
$= (-2 \times -72.280) - (-2 \times -72.066)$
$= 144.550 - 144.132 = 0.428$

(LR is χ^2 with 1 d.f. under H_0: no interaction.)
$0.40 < P < 0.50$, **not significant**
Wald test $P = 0.510$

To obtain the LR statistic in this example, we compute 144.550 minus 144.132 to obtain 0.428. Under the null hypothesis of no interaction effect, the test statistic has a chi-square distribution with p degrees of freedom, where p denotes the number of predictors being assessed. The p-value for this test is between 0.40 and 0.50, which indicates no significant interaction. Although the p-values for the Wald test (0.510) and the LR test are not exactly the same, both p-values lead to the same conclusion.

LR $\neq$ Wald

When in doubt, use the LR test.

In general, the LR and Wald statistics may not give exactly the same answer. Statisticians have shown that of the two test procedures, the LR statistic has better statistical properties, so when in doubt, you should use the LR test.

OUTPUT

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

We now focus on how to assess the effect of treatment status adjusting for log WBC using the model 2 output, again shown here.

Three statistical objectives.

1. **test for significance of effect**
2. **point estimate of effect**
3. **confidence interval for effect**

There are three statistical objectives typically considered. One is to **test for the significance** of the treatment status variable, adjusted for log WBC. Another is to obtain a **point estimate of the effect** of treatment status, adjusted for log WBC. And a third is to obtain a **confidence interval for this effect**. We can accomplish these three objectives using the output provided, without having to explicitly describe the formula for the Cox model being used.

EXAMPLE: (continued)

Test for treatment effect:
Wald statistic: $P = 0.002$ (highly significant)
LR statistic: compare
−2 log L from model 2 with
−2 log L from model without Rx variable
Printout not provided here

Conclusion: treatment effect is significant, after adjusting for log WBC

To test for the significance of the treatment effect, the p-value provided in the table for the Wald statistic is 0.002, which is highly significant. Alternatively, a likelihood ratio (LR) test could be performed by comparing the log likelihood statistic (144.559) for model 2 with the log likelihood statistic for a model which does not contain the treatment variable. This latter model, which should contain only the log WBC variable, is not provided here, so we will not report on it other than to note that the LR test is also very significant. Thus, these test results show that using model 2, the treatment effect is significant, after adjusting for log WBC.

EXAMPLE: (continued)

Point estimate:

$$\widehat{HR} = 3.648 = e^{1.294}$$

Coefficient of treatment variable

A point estimate of the effect of the treatment is provided in the *HR* column by the value 3.648. This value gives the estimated hazard ratio (HR) for the effect of the treatment; in particular, we see that the hazard for the placebo group is 3.6 times the hazard for the treatment group. Note that the value 3.648 is calculated as e to the coefficient of the treatment variable; that is, *e* to the 1.294 equals 3.648.

To describe the confidence interval for the effect of treatment status, we consider the output for the extended table for model 2 given earlier.

Output

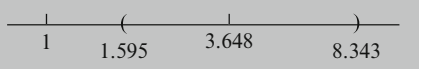
Model 2:

	Coef.	Std. Err.	z	P>|z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
Log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486

No. of subjects = 42 Log likelihood = −72.280 Prob > chi2 = 0.0000

EXAMPLE: (continued)

95% confidence interval for the *HR*: (1.595, 8.343)



95% CI for β_1: $1.294 \pm (1.96)\,(0.422)$

95% CI for $HR = e^{\beta_1}$

$$\exp\left[\hat{\beta}_1 \pm 1.96 s_{\hat{\beta}_1}\right] = e^{1.294 \pm 1.96(0.422)}$$

From the table, we see that a 95% confidence interval for the treatment effect is given by the range of values 1.595–8.343. This is a confidence interval for the hazard ratio (HR), which surrounds the point estimate of 3.648 previously described. Notice that this confidence interval is fairly wide, indicating that the point estimate is somewhat unreliable. As expected from the low p-value of 0.002, the confidence interval for HR does not contain the null value of 1.

The calculation of the confidence interval for HR is carried out as follows:

1. Compute a 95% confidence interval for the regression coefficient of the *Rx* variable (β_1). The large sample formula is 1.294 plus or minus 1.96 times the standard error 0.422, where 1.96 is the 97.5 percentile of the standard normal or *Z* distribution.
2. Exponentiate the two limits obtained for the confidence interval for the regression coefficient of *Rx*.

Stata: provides CI directly

Other packages: provide $\hat{\beta}$ and $s_{\hat{\beta}}$

The Stata output provides the required confidence interval directly, so that the user does not have to carry out the computations required by the large sample formula. Other computer packages may not provide the confidence interval directly, but, rather, may provide only the estimated regression coefficients and their standard errors.

EDITED OUTPUT

Model 1:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

No. of subjects = 42 Log likelihood = −86.380

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

To this point, we have made use of information from outputs for models 2 and 3, but have not yet considered the model 1 output, which is shown again here. Note that model 1 contains only the treatment status variable, whereas model 2, shown below, contains log WBC in addition to treatment status. Model 1 is sometimes called the "crude" model because it ignores the effect of potential covariates of interest, like log WBC.

EXAMPLE: (continued)

HR for model 1 (4.523) is higher than *HR* for model 2 (3.648)

Confounding: crude versus adjusted $\widehat{HR}$ are meaningfully different.

Model 1 can be used in comparison with model 2 to evaluate the potential confounding effect of the variable log WBC. In particular, notice that the value in the *HR* column for the treatment status variable is 4.523 for model 1, but only 3.648 for model 2. Thus, the crude model yields an estimated hazard ratio that is somewhat higher than the corresponding estimate obtained when we adjust for log WBC. If we decide that the crude and adjusted estimates are meaningfully different, we then say that there is confounding due to log WBC.

Confounding due to log WBC ⇒ must control for log WBC, i.e., prefer model 2 to model 1.

Once we decide that confounding is present, we then **must** control for the confounder, in this case, log WBC, in order to obtain a valid estimate of the effect. Thus, we prefer model 2, which controls for log WBC, to model 1, which does not.

If no confounding, then consider precision: e.g., if 95% CI is narrower for model 2 than model 1, we prefer model 2.

Note that if we had decided that there is no "meaningful" confounding, then we would not need to control for log WBC to get a valid answer. Nevertheless, we might wish to control for log WBC anyhow, to obtain a more precise estimate of the hazard ratio. That is, if the confidence interval for the HR is narrower when using model 2 than when using model 1, we would prefer model 2 to model 1 for **precision** gain.

EDITED OUTPUT: Confidence Intervals

	[95% Conf. Interval]	
Rx model 1	2.027	10.094
	width = **8.067**	
	width = **6.748**	
Rx model 2	1.595	8.343
log WBC	2.609	9.486

The confidence intervals for *Rx* in each model are shown here at the left. The interval for *Rx* in model 1 has width equal to 10.094 minus 2.027, or **8.067**; for model 2, the width is 8.343 minus 1.595, or **6.748**. Therefore, model 2 gives a more precise estimate of the hazard ratio than does model 1.

EXAMPLE: (continued)

Model 2 is best model.

$\widehat{HR} = 3.648$ statistically significant

95% CI for *HR*: (1.6, 8.3)

Our analysis of the output for the three models has led us to conclude that model 2 is the best of the three models and that, using model 2, we get a statistically significant hazard ratio of 3.648 for the effect of the treatment, with a 95% confidence interval ranging between 1.6 and 8.3.

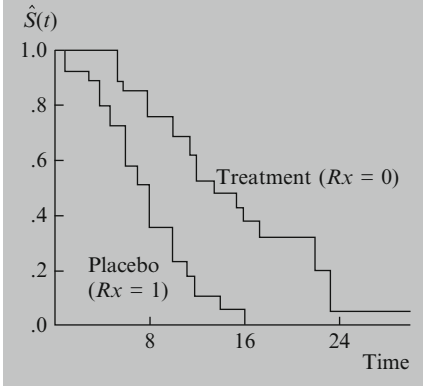
Cox model formulae not specified

Analysis strategy and methods for Cox model analogous to those for logistic and classical linear models.

Note that we were able to carry out this analysis without actually specifying the formulae for the Cox PH models being fit. Also, the strategy and methods used with the output provided have been completely analogous to the strategy and methods one uses when fitting logistic regression models (see Kleinbaum and Klein, *Logistic Regression*, Chapters 6 and 7, 2010), and very similar to carrying out a classical linear regression analysis (see Kleinbaum et al., *Applied Regression Analysis*, 4th ed., Chapter 16, 2008).

EXAMPLE: (continued)

Survival Curves Adjusted for log WBC (Model 2)



In addition to the above analysis of this data, we can also obtain survival curves for each treatment group, **adjusted** for the effects of log WBC and based on the model 2 output. Such curves, sketched here at the left, give additional information to that provided by estimates and tests about the hazard ratio. In particular, these curves describe how the treatment groups compare over the time period of the study.

For these data, the survival curves show that the treatment group consistently has higher survival probabilities than the placebo group after adjusting for log WBC. Moreover, the difference between the two groups appears to widen over time.

Adjusted survival curves	KM curves
Adjusted for covariates	No covariates
Use fitted Cox model	No Cox model fitted

Note that adjusted survival curves are mathematically different from Kaplan–Meier (KM) curves. KM curves do not adjust for covariates and, therefore, are not computed using results from a fitted Cox PH model.

Nevertheless, for these data, the plotted KM curves (which were described in Chapter 2) are similar in appearance to the adjusted survival curves.

Remainder:

- Cox model formula
- basic characteristics of Cox model
- meaning of PH assumption

In the remainder of this presentation, we describe the Cox PH formula and its basic characteristics, including the meaning of the PH assumption and the Cox likelihood.

II. The Formula for the Cox PH Model

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

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

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

The Cox PH model is usually written in terms of the hazard model formula shown here at the left. This 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 $\mathbf{X}$. That is, the bold $\mathbf{X}$ represents a collection (sometimes called a "vector") of predictor variables that is being modeled to predict an individual's hazard.

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.

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. In contrast, the exponential expression shown here, 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)

It is possible, nevertheless, to consider *X*'s which 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**.

Time-dependent variables: Chapter 6

The use of time-dependent variables is discussed in Chapter 6. For the remainder of this presentation, we will consider time-independent *X*'s only.

Time-independent variable: Values for a given individual do not change over time; e.g., SEX and SMK

Assumed not to change once measured

A time-independent variable is defined to be any variable whose value for a given individual does not change over time. Examples are SEX and smoking status (SMK). Note, however, that a person's smoking status may actually change over time, but for purposes of the analysis, the SMK variable is assumed not to change once it is measured, so that only one value per individual is used.

AGE and WGT values do not change much, or effect on survival depends on one measurement.

Also note that although variables like AGE and weight (WGT) change over time, it may be appropriate to treat such variables as time-independent in the analysis if their values do not change much over time or if the effect of such variables on survival risk depends essentially on the value at only one measurement.

$$X_1 = X_2 = \cdots = X_p = 0$$

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

$$= h_0(t)\, e^0$$

$$= h_0(t)$$

Baseline hazard

The Cox model formula has the property that if all the *X*'s are equal to zero, the formula reduces to the baseline hazard function. That is, the exponential part of the formula becomes *e* to the zero, which is 1. This property of the Cox model is the reason why $h_0(t)$ is called the baseline function.

No *X*'s in model: $h(t,\mathbf{X}) = h_0(t)$.

Or, from a slightly different perspective, the Cox model reduces to the baseline hazard when no *X*'s are in the model. Thus, $h_0(t)$ may be considered as a starting or "baseline" version of the hazard function, prior to considering any of the *X*'s.

$h_0(t)$ is unspecified.

Cox model: **semiparametric**

Another important property of the Cox model is that the baseline hazard, $h_0(t)$, is an unspecified function. It is this property that makes the Cox model a **semiparametric** model.

Example: Parametric Model

Weibull:
$h(t, \mathbf{X}) = \lambda p t^{p-1}$
where $\lambda = \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$
and $h_0(t) = p t^{p-1}$

In contrast, a **parametric** model is one whose functional form is completely specified, except for the values of the unknown parameters. For example, the Weibull hazard model is a parametric model and has the form shown here, where the unknown parameters are λ, p, and the β_i's. Note that for the Weibull model, $h_0(t)$ is given by $\lambda p t^{p-1}$ (see Chapter 7).

Semiparametric property
⇓
Popularity of the Cox model

One of the reasons why the Cox model is so popular is that it is semiparametric. We discuss this and other reasons in the next section (III) concerning why the Cox model is so widely used.

III. Why the Cox PH Model Is Popular

Cox PH model is "robust": Will closely approximate correct parametric model

A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox PH model is a "robust" model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

If correct model is:

Weibull	⇒	Cox model will approximate Weibull
Exponential	⇒	Cox model will approximate exponential

For example, if the correct parametric model is Weibull, then use of the Cox model typically will give results comparable to those obtained using a Weibull model. Or, if the correct model is exponential, then the Cox model results will closely approximate the results from fitting an exponential model.

Prefer parametric model if sure of correct model, e.g., use goodness-of-fit test (Lee, 1982).

We would prefer to use a parametric model if we were sure of the correct model. Although there are various methods for assessing goodness of fit of a parametric model (for example, see Lee, *Statistical Methods for Survival Data Analysis*, 1982), we may not be completely certain that a given parametric model is appropriate.

When in doubt, the Cox model is a "safe" choice.

Thus, when in doubt, as is typically the case, the Cox model will give reliable enough results so that it is a "safe" choice of model, and the user does not need to worry about whether the wrong parametric model is chosen.

In addition to the general "robustness" of the Cox model, the specific form of the model is attractive for several reasons.

$$h(t,\mathbf{X}) = \underbrace{h_0(t)}_{\text{Baseline hazard}} \times \underbrace{e^{\sum_{i=1}^{p}\beta_i X_i}}_{\begin{array}{c}\text{Exponential}\\ \Downarrow \\ 0 \le h(t,\mathbf{X}) < \infty \text{ always}\end{array}}$$

As described previously, the specific form of the Cox model gives the hazard function as a product of a baseline hazard involving t and an exponential expression involving the X's without t. The exponential part of this product is appealing because it ensures that the fitted model will always give estimated hazards that are non-negative.

$$h_0(t) \times \underbrace{\sum_{i=1}^{p}\beta_i X_i}_{\text{Linear}} \Downarrow \text{Might be} < 0$$

We want such nonnegative estimates because, by definition, the values of any hazard function must range between zero and plus infinity, that is, a hazard is always nonnegative. If, instead of an exponential expression, the X part of the model were, for example, linear in the X's, we might obtain negative hazard estimates, which are not allowed.

Even though $h_0(t)$ is unspecified, we can estimate the β's.

Measure of effect: hazard ratio (*HR*) involves only β's, without estimating $h_0(t)$.

Another appealing property of the Cox model is that, even though the baseline hazard part of the model is unspecified, it is still possible to estimate the β's in the exponential part of the model. As we will show later, all we need are estimates of the β's to assess the effect of explanatory variables of interest. The measure of effect, which is called a hazard ratio, is calculated without having to estimate the baseline hazard function.

Can estimate $h(t,\mathbf{X})$ and $S(t, \mathbf{X})$ for Cox model using a minimum of assumptions.

Note that the hazard function $h(t,\mathbf{X})$ and its corresponding survival curves $\mathbf{S}(t,\mathbf{X})$ can be estimated for the Cox model even though the baseline hazard function is not specified. Thus, with the Cox model, using a minimum of assumptions, we can obtain the primary information desired from a survival analysis, namely, a hazard ratio and a survival curve.

Cox model preferred to **logistic** model.

⇓	⇓
Uses survival times and censoring	Uses (0,1) outcome; ignores survival times and censoring

One last point about the popularity of the Cox model is that it is preferred over the logistic model when survival time information is available and there is censoring. That is, the Cox model uses more information, the survival times, than the logistic model, which considers a (0, 1) outcome and ignores survival times and censoring.

IV. ML Estimation of the Cox PH Model

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

We now describe how estimates are obtained for the parameters of the Cox model. The parameters are the β's in the general Cox model formula shown here. The corresponding estimates of these parameters are called maximum likelihood (ML) estimates and are denoted as $\hat{\beta}_i$.

ML estimates: $\hat{\beta}_i$

	Coef.	Std.Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

As an example of ML estimates, we consider once again the computer output for one of the models (model 2) fitted previously from remission data on 42 leukemia patients.

Estimated model:

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t)\, e^{1.294\, Rx\, +\, 1.604\, \log\, \mathrm{WBC}}$$

The Cox model for this example involves two parameters, one being the coefficient of the treatment variable (denoted here as *Rx*) and the other being the coefficient of the log WBC variable. The expression for this model is shown at the left, which contains the estimated coefficients 1.294 for *Rx* and 1.604 for log white blood cell count.

ML estimates: maximize likelihood function L

L = joint probability of observed data = $L(\beta)$

As with logistic regression, the ML estimates of the Cox model parameters are derived by maximizing a likelihood function, usually denoted as L. The likelihood function is a mathematical expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters (the β's) in the model being considered. L is sometimes written notationally as $L(\beta)$ where β denotes the collection of unknown parameters.

The expression for the likelihood is developed at the end of the chapter. However, we give a brief overview below.

L is a partial likelihood:

- considers probabilities only for subjects who fail
- does not consider probabilities for subjects who are censored

The formula for the Cox model likelihood function is actually called a "partial" likelihood function rather than a (complete) likelihood function. The term "partial" likelihood is used because the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. Thus the likelihood for the Cox model does not consider probabilities for all subjects, and so it is called a "partial" likelihood.

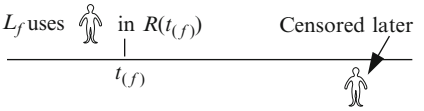
Number of failure times

$$L = L_1 \times L_2 \times L_3 \times \cdots \times L_k = \prod_{j=1}^{k} L_j$$

where
L_f = portion of L for the jth failure time given the risk set $R\,(t_{(f)})$

In particular, the partial likelihood can be written as the product of several likelihoods, one for each of, say, k failure times. Thus, at the f-th failure time, L_f denotes the likelihood of failing at this time, given survival up to this time. Note that the set of individuals at risk at the jth failure time is called the "risk set," $R(t_{(f)})$, and this set will change – actually get smaller in size – as the failure time increases.

Information on censored subjects used prior to censorship.



Thus, although the partial likelihood focuses on subjects who fail, survival time information prior to censorship is used for those subjects who are censored. That is, a person who is censored *after* the f-th failure time is part of the risk set used to compute L_f even though this person is censored later.

Steps for obtaining ML estimates:

- form L from model
- maximize ln L by solving

$$\frac{\partial \ln L}{\partial \beta_i} = 0$$
$$i = 1, \ldots, p(\#\text{ of parameters})$$

Once the likelihood function is formed for a given model, the next step for the computer is to maximize this function. This is generally done by maximizing the natural log of L, which is computationally easier.

Solution by iteration:

- guess at solution
- modify guess in successive steps
- stop when solution is obtained

The maximization process is carried out by taking partial derivatives of log of L with respect to each parameter in the model, and then solving a system of equations as shown here. This solution is carried out using **iteration**. That is, the solution is obtained in a stepwise manner, which starts with a guessed value for the solution, and then successively modifies the guessed value until a solution is finally obtained.

Statistical inferences for hazard ratios: (See Section I, pages 100–107)

Test hypotheses	Confidence intervals
Wald test LR test	Large sample 95% CI

$\widehat{HR} = e^{\hat{\beta}}$ for a (0, 1) exposure variable (no interaction)

Once the ML estimates are obtained, we are usually interested in carrying out statistical inferences about hazard ratios defined in terms of these estimates. We illustrated previously how to test hypotheses and form confidence intervals for the hazard ratio in Section I above. There, we described how to compute a Wald test and a likelihood ratio (LR) test. We also illustrated how to calculate a large sample 95% confidence interval for a hazard ratio. The estimated hazard ratio (HR) was computed by exponentiating the coefficient of a (0,1) exposure variable of interest. Note that the model contained no interaction terms involving exposure.

V. Computing the Hazard Ratio

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

where

$$\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)$$

and

$$\mathbf{X} = (X_1, X_2, \cdots, X_p)$$

denote the set of X's for two individuals

In general, a hazard ratio (HR) is defined as the hazard for one individual divided by the hazard for a different individual. The two individuals being compared can be distinguished by their values for the set of predictors, that is, the X's.

We can write the hazard ratio as the estimate of $h(t,\mathbf{X}^*)$ divided by the estimate of $h(t,\mathbf{X})$, where $\mathbf{X}^*$ denotes the set of predictors for one individual, and $\mathbf{X}$ denotes the set of predictors for the other individual.

To interpret $\widehat{HR}$, want $\widehat{HR} > 1$, i.e., $\hat{h}(t, \mathbf{X}^*) > \hat{h}(t, \mathbf{X})$.

Typical coding: $\mathbf{X}^*$: group with larger h
$\mathbf{X}$: group with smaller h

Note that, as with an odds ratio, it is easier to interpret an HR that exceeds the null value of 1 than an HR that is less than 1. Thus, the X's are typically coded so that group with the larger hazard corresponds to $\mathbf{X}^*$, and the group with the smaller hazard corresponds to $\mathbf{X}$. As an example, for the remission data described previously, the placebo group is coded as $X_1^* = 1$, and the treatment group is coded as $X_1 = 0$.

EXAMPLE: Remission Data

$\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)$, where $X_1^* = 1$ denotes **placebo** group.

$\mathbf{X} = (X_1, X_2, \ldots, X_p)$, where $X_1 = 0$ denotes **treatment** group.

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\cancel{\hat{h}_0(t)} e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}}$$

We now obtain an expression for the *HR* formula in terms of the regression coefficients by substituting the Cox model formula into the numerator and denominator of the hazard ratio expression. This substitution is shown here. Notice that the only difference in the numerator and denominator are the *X**'s versus the *X*'s. Notice also that the baseline hazards will cancel out.

$$\widehat{HR} = \frac{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)}$$

Using algebra involving exponentials, the hazard ratio formula simplifies to the exponential expression shown here. Thus, the hazard ratio is computed by exponentiating the sum of each β_i "hat" times the difference between X_i^* and X_i.

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

An alternative way to write this formula, using exponential notation, is shown here. We will now illustrate the use of this general formula through a few examples.

EXAMPLE

$\mathbf{X} = (X_1, X_2, \ldots, X_p) = (X_1)$, where X_1 denotes (0, 1) exposure status ($p = 1$)
$X_1^* = 1, X_1 = 0$

$$\widehat{HR} = \exp\left[\hat{\beta}_1 \left(X_1^* - X_1\right)\right]$$
$$= \exp\left[\hat{\beta}_1 (1 - 0)\right] = e^{\hat{\beta}_1}$$

Suppose, for example, there is only one X variable of interest, X_1, which denotes (0,1) exposure status, so that $p = 1$. Then, the hazard ratio comparing exposed to unexposed persons is obtained by letting $X_1^* = 1$ and $X_1 = 0$ in the hazard ratio formula. The estimated hazard ratio then becomes e to the quantity β_1 "hat" times 1 minus 0, which simplifies to e to the β_1 "hat."

Model 1:

	Coef.	Std. Err.	P > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

Recall the remission data printout for Model 1, which contains only the *Rx* variable, again shown here. Then the estimated hazard ratio is obtained by exponentiating the coefficient 1.509, which gives the value 4.523 shown in the *HR* column of the output.

EXAMPLE 2

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

$\mathbf{X}^* = (1, \log \text{WBC})$, $\mathbf{X} = (0, \log \text{WBC})$
HR for effect of *Rx* adjusted for log WBC:

As a second example, consider the output for Model 2, which contains two variables, the *Rx* variable and log WBC. Then to obtain the hazard ratio for the effect of the *Rx* variable adjusted for the log WBC variable, we let the vectors $\mathbf{X}^*$ and $\mathbf{X}$ be defined as $\mathbf{X}^* = (1, \log \text{WBC})$ and $\mathbf{X} = (0, \log \text{WBC})$. Here we assume that log WBC is the same for $\mathbf{X}^*$ and $\mathbf{X}$ though unspecified.

EXAMPLE 2: (continued)

$$\widehat{HR} = \exp\left[\hat{\beta}_1(X_1^* - X_1) + \hat{\beta}_1(X_2^* - X_2)\right]$$
$$= \exp[1.294(1-0) + 1.604(\log\ \text{WBC} - \log\ \text{WBC})]$$
$$= \exp[1.294(1) + 1.604(0)] = \boxed{e^{1.294}}$$

The estimated hazard ratio is then obtained by exponentiating the sum of two quantities, one involving the coefficient 1.294 of the *Rx* variable, and the other involving the coefficient 1.604 of the log WBC variable. Since the log WBC value is fixed, however, this portion of the exponential is zero, so that the resulting estimate is simply e to the 1.294.

General rule: If X_1 is a (0,1) exposure variable, then $\widehat{HR} = e^{\hat{\beta}_1}$ (= effect of exposure adjusted for other X's) provided no other X's are product terms involving exposure.

This second example illustrates the general rule that the hazard ratio for the effect of a (0,1) exposure variable which adjusts for other variables is obtained by exponentiating the estimated coefficient of the exposure variable. This rule has the proviso that the model does not contain any product terms involving exposure.

EXAMPLE 3

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

Want *HR* for effect of *Rx* adjusted for log WBC.

Placebo subject:

$$\mathbf{X}^* = (X_1^* = 1,\ X_2^* = \log\ \text{WBC},\ X_3^* = 1 \times \log\ \text{WBC})$$

Treated subject:

$$\boldsymbol{X} = (X_1 = 0, X_2 = \log \text{WBC}, X_3 = 0 \times \log \text{WBC})$$

$$\widehat{HR} = \exp\left[\sum_{i=1}^{3} \hat{\beta}_i(X_i^* - X_i)\right]$$
$$\widehat{HR} = \exp[2.355(1-0) + 1.803(\log\ \text{WBC} - \log\ \text{WBC}) + (-0.342)(1 \times \log\ \text{WBC} - 0 \times \log\ \text{WBC})]$$
$$\boxed{= \exp[2.355 - 0.342\ \log\ \text{WBC}]}$$

We now give a third example which illustrates how to compute a hazard ratio when the model does contain product terms. We consider the printout for Model 3 of the remission data shown here.

To obtain the hazard ratio for the effect of *Rx* adjusted for log WBC using Model 3, we consider $\mathbf{X}^*$ and $\mathbf{X}$ vectors which have three components, one for each variable in the model. The $\mathbf{X}^*$ vector, which denotes a placebo subject, has components $X_1^* = 1$, $X_2^* = \log \text{WBC}$ and $1 \times \log \text{WBC}$. The $\mathbf{X}$ vector, which denotes a treated subject, has components $X_1 = 0$, $X_2 = \log \text{WBC}$ and $X_3 = 0 \times \log \text{WBC}$. Note again that, as with the previous example, the value for log WBC is treated as fixed, though unspecified.

Using the general formula for the hazard ratio, we must now compute the exponential of the sum of three quantities, corresponding to the three variables in the model. Substituting the values from the printout and the values of the vectors $\mathbf{X}^*$ and $\mathbf{X}$ into this formula, we obtain the exponential expression shown here. Using algebra, this expression simplifies to the exponential of 2.355 minus 0.342 times log WBC.

EXAMPLE: (continued)

log WBC = 2:

$$\widehat{HR} = \exp[2.355 - 0.342(2)]$$
$$= e^{1.671} = 5.32$$

log WBC = 4:

$$\widehat{HR} = \exp[2.355 - 0.342(4)]$$
$$= e^{0.987} = 2.68$$

In order to get a numerical value for the hazard ratio, we must specify a value for log WBC. For instance, if log WBC = 2, the estimated hazard ratio becomes 5.32, whereas if log WBC = 4, the estimated hazard ratio becomes 2.68. Thus, we get different hazard ratio values for different values of log WBC, which should make sense since log WBC is an effect modifier in Model 3.

General rule for (0, 1) exposure variables when there are product terms:

$$\boxed{\widehat{HR} = \exp\left[\hat{\beta} + \sum \hat{\delta}_j W_j\right]}$$

where

$\hat{\beta}$ = coefficient of E

$\hat{\delta}_j$ = coefficient of $E \times W_j$

($\widehat{HR}$ does not contain coefficients of non-product terms)

The example we have just described using Model 3 illustrates a general rule which states that the hazard ratio for the effect of a (0,1) exposure variable in a model which contains product terms involving this exposure with other X's can be written as shown here. Note that β "hat" denotes the coefficient of the exposure variable and the δ "hats" are coefficients of product terms in the model of the form $E \times W_j$. Also note that this formula does not contain coefficients of nonproduct terms other than those involving E.

EXAMPLE

Model 3:

$\hat{\beta}$ = coefficient of Rx (E)

$\hat{\delta}_1$ = coefficient of $Rx \times$ log WBC (W_1 = log WBC)

$$\widehat{HR}\,(\text{Model 3}) = \exp\left[\hat{\beta} + \hat{\delta}_1 \log \text{WBC}\right]$$
$$= \exp[2.355 - 0.342 \log \text{WBC}]$$

For Model 3, β "hat" is the coefficient of the Rx variable, and there is only one δ "hat" in the sum, which is the coefficient of the product term $Rx \times$ log WBC. Thus, there is only one W, namely W_1 = log WBC. The hazard ratio formula for the effect of exposure is then given by exponentiating β "hat" plus δ "hat" times log WBC. Substituting the estimates from the printout into this formula yields the expression obtained previously, namely the exponential of 2.355 minus 0.342 times log WBC.

VI. Interval Estimation: Interaction

We have previously illustrated in Model 2 of the Remission Time Data how to obtain a 95% interval estimate of the HR when there is only one regression coefficient of interest, e.g., the HR is of the form $\exp[\beta_1]$.

Model 2:

$h(t, X) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$

$HR = \exp[\beta_1]$

Large sample 95% confidence interval:

$$\exp\left[\hat{\beta}_1 \pm 1.96\sqrt{\hat{Var}\hat{\beta}_1}\right]$$

where

$$s_{\hat{\beta}_1} = \sqrt{\hat{Var}\hat{\beta}_1}$$

The procedure typically used to obtain a large sample 95% confidence interval (CI) for the parameter is to compute *the exponential of the estimate of the parameter plus or minus a percentage point of the normal distribution times the estimated standard error of the estimate*. Note that the square root of the estimated variance is the standard error.

No interaction: simple formula

Interaction: complex formula

This computation is relatively simple when there are no interaction effects in the model. However, when there is interaction, the computational formula for the estimated standard error is more complex.

Model 3:

$$h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log WBC + \beta_3 (Rx \times \log WBC)]$$

Suppose we focus on Model 3, shown here on the left, and, again, we assume that Rx is a (0, 1) exposure variable of interest. Then the formula for the HR for the effect of Rx controlling for the variable log WBC is given on the left underneath the model formula.

$$HR = \exp[\beta_1 + \beta_3 \log WBC]$$

Interaction: variance calculation difficult

No interaction: variance directly from printout

The difficult part in computing the CI for a HR involving interaction effects is the calculation for the estimated variance. When there is no interaction, so that the parameter of interest is a single regression coefficient, this variance is obtained directly from the listing of estimated coefficients and corresponding standard errors.

$$\hat{HR} = exp[\hat{\ell}],$$
where $\ell = \beta_1 + \beta_3 \log WBC$

For Model 3, we can alternatively write the estimated HR formula as $\exp[\hat{\ell}]$, where ℓ is the linear function $\beta_1 + \beta_3 \log WBC$ and $\hat{\ell}$ is the estimate of this linear function using the ML estimates.

$$\boxed{\begin{array}{l}95\%\text{ CI for HR} = \exp[\hat{\ell}] \\ \exp[\hat{\ell} \pm 1.96\sqrt{\text{Var}\hat{\ell})}]\end{array}}$$

To obtain a 95% CI for $\exp[\ell]$ we must exponentiate the CI for ℓ. The formula is shown on the left.

General Formula:
can consider any ℓ, e.g.,

$$\ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots \delta_k W_k,$$

where $X_1 = (0, 1)$ exposure variable and β_1 = coeff of X_1,
δ_j = coeff of $X_1 \times W_j$, j=1,..., k

This CI formula, though motivated by our example using Model 3, is actually the general formula for the 95% CI for any HR of interest from a Cox PH model. In general, for a model with a (0, 1) exposure variable X_1 and interaction terms $X_1 \times W_1$, ..., $X_1 \times W_k$, the linear function may take any form of interest, as shown in the left.

$Var(\hat{\ell}) = Var(\hat{\beta}_1 + \hat{\delta}_1 W_1 + \cdots \hat{\delta}_k W_k)$
where the estimates $\hat{\beta}_1, \hat{\delta}_i, \ldots, \hat{\delta}_k$ are correlated, so one must use
$Var(\hat{\beta}_1), Cov(\hat{\beta}_1, \hat{\delta}_i)$ *and* $Cov(\hat{\delta}_i, \hat{\delta}_j)$

When the HR involves interaction effects, the estimated variance considers a linear sum of estimated regression coefficients. Because the coefficients in the linear sum are estimated from the same data set, these coefficients are correlated with one another. Consequently, the calculation of the estimated variance must consider both the variances and the covariances of the estimated coefficients, which makes computations somewhat cumbersome.

Computer packages SAS and STATA compute $\hat{\text{Var}}\,\hat{\ell}$ as part of the program options (see Computer Appendix).

Nevertheless, most computer packages that have procedures for fitting survival analysis models like the Cox model provide for computing the estimated variance of linear functions like $\hat{\ell}$ as part of the program options. See the Computer Appendix for details on the use of the "contrast" option in SAS and the "lincom" option in STATA.

General formula for $\hat{\text{Var}}\,\hat{\ell}$:

$$\hat{\text{Var}}(\hat{\ell}) = \hat{\text{Var}}(\hat{\beta}_1) + \sum_j W_j^2 \hat{\text{Var}}(\hat{\delta}_1) + 2\sum_j W_j \hat{Cov}(\hat{\beta}_1, \hat{\delta}_j) + 2\sum_j \sum_k W_j W_k \hat{Cov}(\hat{\delta}_j, \hat{\delta}_k)$$

For the interested reader, we provide here the general formula for the estimated variance of the linear function $\hat{\ell}$.

- Variances and covariances provided in the computer output
- User specifies W's values of interest.

In applying this formula, the user obtains the estimated variances and covariances from the variancecovariance output. The user must specify values of interest for the effect modifiers defined by the Ws in the model.

Model 3:
$\ell = \beta_1 + \beta_3 \log \text{WBC}$

$$\text{Var}(\hat{\ell}) = \text{Var}(\hat{\beta}_1) + (\log WBC)^2 \text{Var}(\hat{\beta}_3) + 2(\log WBC)^2 Cov(\hat{\beta}_1, \hat{\beta}_3)$$

Applying this variance formula to Model 3, we obtain the variance expression shown on the left. Since log WBC is the only effect modifier here, the user would need to specify log WBC values of interest, e.g., log WBC =2 and log WBC=4.

95% CI for Rx in Model 3
(SAS edited output):

log WBC	$\widehat{HR}$	S.E.	Conf	Limits
2	5.3151	3.8410	1.2894	21.9101
4	2.6809	1.6520	0.8013	8.9700

Using the "contrast" option in SAS's PHREG procedure, we show on the left computed 95% CI's for the Rx variable for two choices of log WBC. When log WBC is 2, the estimated HR is 5.32 with a 95% CI given by the limits (1.29, 21.91), whereas when log WBC is 4, the estimated HR is 2.68 with a 95% CI given by the limits (0.80, 8.97).

CI results suggest log WBC by Rx interaction but conflict with nonsignificant interaction test result.
Note: small study size (n=42)

These results suggest the interaction of log WBC with Rx, and they conflict with the previously reported nonsignificance of the test for interaction in Model 3, which might primarily be attributed to the small sample size (n=42) of this study.

VII. Adjusted Survival Curves Using the Cox PH model

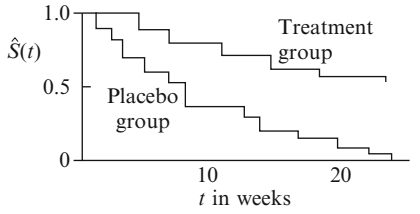
The two primary quantities desired from a survival analvsis point of viem are estimated hazard ratios and estimated survival curves. Having just described how to compute hazard ratios, we now turn to estimation of survival curves using the Cox model.

Two primary quantities:

1. estimated hazard ratios
2. estimated survival curves

No model: use KM cunzes

Recall that if no model is used to fit survival data, a survival curve can be estimated using a Kaplan–Meier method. Such KM curves are plotted as step functions as shown here for the remission data example.



Cox model: adjusted survival curves (also step fuuctions).

When a Cox model is used to fit survival data, survival curves can be obtained that adjust for the explanatory variables used as predictors. These are called **adjusted survival curves**, and, like KM curves, these are also plotted as step functions.

Cox model hazard function:

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

Cox model survival function:

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

The hazard function formula for the Cox PH model, shown here again, can be converted to a corresponding survival function formula as shown below. This survival function formula is the basis for determining adjusted survival curves. Note that this formula says that the survival function at time t for a subject with vector $\mathbf{X}$ as predictors is given by a baseline survival function $S_0(t)$ raised to a power equal to the exponential of the sum of β_i times X_i.

Estimated survival function:

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

The expression for the estimated survival function can then be written with the usual "hat" notation as shown here.

$\hat{S}_0(t)$ and $\hat{\beta}_i$ are provided by the computer program. The X_i must be specified by the investigator.

The estimates of $\hat{S}_0(t)$ and $\hat{\beta}_i$ are provided by the computer program that fits the Cox model. The X's, however, must first be specified by the investigator before the computer program can compute the estimated survival curve.

EXAMPLE: Model 2 Remission Data

$\hat{h}(t,\mathbf{X}) = \hat{h}_0(t)e^{1.294\,Rx\,+\,1.604\,\log\,\mathrm{WBC}}$
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp(1.294\,Rx\,+\,1.604\,\log\,\mathrm{WBC})}$

For example, if we consider model 2 for the remission data, the fitted model written in terms of both the hazard function and corresponding survival function is given here.

Specify values for $\mathbf{X} = (Rx, \log \mathrm{WBC})$

We can obtain a specific survival curve by specifying values for the vector $\mathbf{X}$, whose component variables are Rx and log WBC.

$Rx = 1$, log WBC = 2.93:

$$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_1\overline{Rx}+\hat{\beta}_2\overline{\log\ \mathrm{WBC}}\right)}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(1.294(0.5)+1.604(2.93))}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(5.35)} = \boxed{\left[\hat{S}_0(t)\right]^{210.6}}$$

For instance, if $Rx = 1$ and log WBC = 2.93, the estimated survival curve is obtained by substituting these values in the formula as shown here, and carrying out the algebra to obtain the expression circled. Note that the value 2.93 is the overall mean log WBC for the entire dataset of 42 subjects.

$Rx = 0$, log WBC = 2.93:

$$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp(1.294(0)+1.604(2.93))}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(4.70)} = \boxed{\left[\hat{S}_0(t)\right]^{109.9}}$$

Also, if $Rx = 0$ and log WBC = 2.93, the estimated survival curve is obtained as shown here.

Each of the circled expressions gives **adjusted** survival curves, where the adjustment is for the values specified for the X's. Note that for each expression, a survival probability can be obtained for any value of t.

Adjusted Survival Curves

$Rx = 1$, log WBC = 2.93:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{400.9}$
$Rx = 0$, log WBC = 2.93:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{109.9}$

The two formulae just obtained, again shown here, allow us to compare survival curves for different treatment groups adjusted for the covariate log WBC. Both curves describe estimated survival probabilities over time assuming the same value of log WBC, in this case, the value 2.93.

Typically, use $X = \overline{X}$ or X_{median}

Computer uses $\overline{X}$

Typically, when computing adjusted survival curves, the value chosen for a covariate being adjusted is an average value like an arithmetic mean or a median. In fact, most computer programs for the Cox model automatically use the mean value over all subjects for each covariate being adjusted.

EXAMPLE: (continued)

Remission data ($n = 42$):
$\overline{\log\ \mathrm{WBC}} = 2.93$

In our example, the mean log WBC for all 42 subjects in the remission data set is 2.93. That is why we chose this value for log WBC in the formulae for the adjusted survival curve.

General formulae for adjusted survival curves comparing two groups:

Exposed subjects:

$$\hat{S}(t, \mathbf{X}_1) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1(1) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i\right]}$$

Unexposed subjects:

$$\hat{S}(t, \mathbf{X}_0) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1(0) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i\right]}$$

More generally, if we want to compare survival curves for two levels of an exposure variable, and we want to adjust for several covariates, we can write the formula for each curve as shown here. Note that we are assuming that the exposure variable is variable X_1, whose estimated coefficient is β_1 "hat," and the value of X_1 is 1 for exposed and 0 for unexposed subjects.

General formula for adjusted survival curve for all covariates in the model:

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

Also, if we want to obtain an adjusted survival curve which adjusts for all covariates in the model, the general formula which uses the mean value for each covariate is given as shown here. This formula will give a single adjusted survival curve rather than different curves for each exposure group.

EXAMPLE

Single survival curve for Cox model containing *Rx* and log WBC:

$\overline{Rx} = 0.50$

$\overline{\log\ \text{WBC}} = 2.93$

$$\begin{aligned}\hat{S}(t, \mathbf{X}) &= \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_1 \overline{Rx} + \hat{\beta}_2 \overline{\log\ \text{WBC}}\right)} \\ &= \left[\hat{S}_0(t)\right]^{\exp(1.294(0.5) + 1.604(2.93))} \\ &= \left[\hat{S}_0(t)\right]^{\exp(5.35)} = \boxed{\left[\hat{S}_0(t)\right]^{210.6}}\end{aligned}$$

To illustrate this formula, suppose we again consider the remission data, and we wish to obtain a single survival curve that adjusts for both *Rx* and log WBC in the fitted Cox model containing these two variables. Using the mean value of each covariate, we find that the mean value for *Rx* is 0.5 and the mean value for log WBC is 2.93, as before.

To obtain the single survival curve that adjusts for *Rx* and log WBC, we then substitute the mean values in the formula for the adjusted survival curve for the model fitted. The formula and the resulting expression for the adjusted survival curve are shown here. (Note that for the remission data, where it is of interest to compare two exposure groups, the use of a single survival curve is not appropriate.)

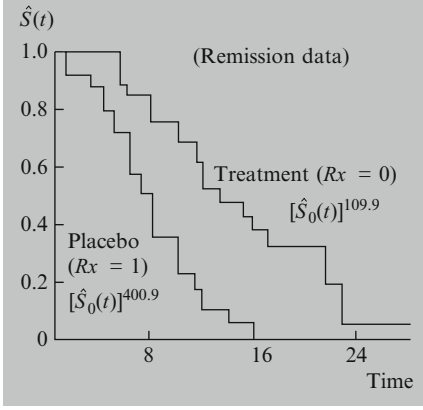
Compute survival probability by specifying value for t in $\hat{S}(t, \overline{\mathbf{X}}) = [\hat{S}_0(t)]^{210.6}$

Computer uses t's which are failure times.

From this expression for the survival curve, a survival probability can be computed for any value of t that is specified. When graphing this survival curve using a computer package, the values of t that are chosen are the failure times of all persons in the study who got the event. This process is automatically carried out by the computer without having the user specify each failure time.

EXAMPLE

Adjusted Survival Curves for Treatment and Placebo Groups



The graph of adjusted survival curves obtained from fitting a Cox model is usually plotted as a step function. For example, we show here the step functions for the two adjusted survival curves obtained by specifying either 1 or 0 for treatment status and letting log WBC be the mean value 2.93.

Next section: PH assumption

- explain meaning
- when PH **not** satisfied

We now turn to the concept of the proportional hazard (PH) assumption. In the next section, we explain the meaning of this assumption and we give an example of when this assumption is not satisfied.

Later presentations:

- how to evaluate PH
- analysis when PH not met

In later presentations, we expand on this subject, describing how to evaluate statistically whether the assumption is met and how to carry out the analysis when the assumption is not met.

VIII. The Meaning of the PH Assumption

PH: *HR* is constant over time, i.e., $\hat{h}(t,\mathbf{X}^*) = \text{constant} \times \hat{h}(t,\mathbf{X})$

$$\widehat{HR} = \frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \frac{\hat{h}_0(t)\exp\left[\sum \hat{\beta}_i X_i^*\right]}{\hat{h}_0(t)\exp\left[\sum \hat{\beta}_i X_i\right]} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

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

The PH assumption requires that the *HR* is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

To understand the PH assumption, we need to reconsider the formula for the *HR* that compares two different specifications $\mathbf{X}^*$ and $\mathbf{X}$ for the explanatory variables used in the Cox model. We derived this formula previously in Section V, and we show this derivation again here. Notice that the baseline hazard function $\hat{h}_0(t)$ appears in both the numerator and denominator of the hazard ratio and cancels out of the formula.

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

does not involve t.

The final expression for the hazard ratio therefore involves the estimated coefficients β_i "hat" and the values of $\mathbf{X}^*$ and $\mathbf{X}$ for each variable. However, because the baseline hazard has canceled out, the final expression does not involve time t.

Let (Constant)

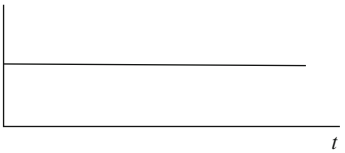
$$\hat{\theta} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right]$$

then

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

Thus, once the model is fitted and the values for $\mathbf{X}^*$ and $\mathbf{X}$ are specified, the value of the exponential expression for the estimated hazard ratio is a constant, which does not depend on time. If we denote this constant by θ "hat," then we can write the hazard ratio as shown here. This is a mathematical expression which states the proportional hazards assumption.

$\widehat{HR}$ ($\mathbf{X}^*$ versus $\mathbf{X}$)



Graphically, this expression says that the estimated hazard ratio comparing any two individuals plots as a constant over time.

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

Proportionality constant (not dependent on time)

Another way to write the proportional hazards assumption mathematically expresses the hazard function for individual $\mathbf{X}^*$ as θ "hat" times the hazard function for individual $\mathbf{X}$, as shown here. This expression says that the hazard function for one individual is proportional to the hazard function for another individual, where the proportionality constant is θ "hat," which does not depend on time.

EXAMPLE: Remission Data

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t) e^{1.294\,Rx + 1.604\ \log\ \mathrm{WBC}}$$

$$\widehat{HR} = \frac{\hat{h}(t, Rx = 1, \log \mathrm{WBC} = 2.93)}{\hat{h}(t, Rx = 0, \log \mathrm{WBC} = 2.93)} = \exp[1.294] = 3.65 \text{ Constant}$$

Placebo

$$\hat{h}(t, Rx = 1, \log \mathrm{WBC} = 2.93) = 3.65\, \hat{h}(t, Rx = 0, \log \mathrm{WBC} = 2.93)$$

Treatment

3.65 = proportionality constant

To illustrate the proportional hazard assumption, we again consider the Cox model for the remission data involving the two variables Rx and log WBC. For this model, the estimated hazard ratio that compares placebo ($Rx = 1$) with treated ($Rx = 0$) subjects controlling for log WBC is given by e to the 1.294, which is 3.65, a constant.

Thus, the hazard for placebo group ($Rx = 1$) is 3.65 times the hazard for the treatment group ($Rx = 0$), and the value, 3.65, is the same regardless of time. In other words, using the above model, the hazard for the placebo group is proportional to the hazard for the treatment group, and the proportionality constant is 3.65.

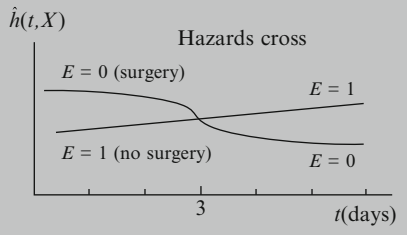
EXAMPLE: PH Not Satisfied

Cancer patients → Surgery
Cancer patients → Radiation with no surgery

$$E = \begin{cases} 0 & \text{if surgery} \\ 1 & \text{if no surgery} \end{cases}$$

$$h(t,\mathbf{X}) = h_0(t)e^{\beta E}$$

Is the above Cox PH model appropriate?
Note:
Serious surgery $\Rightarrow$ High risk for death early



2 days: $\dfrac{\hat{h}(t=2,\ E=1)}{\hat{h}(t=2,\ E=0)} < 1$

but

5 days: $\dfrac{\hat{h}(t=5,\ E=1)}{\hat{h}(t=5,\ E=0)} > 1$

To further illustrate the concept of proportional hazards, we now provide an example of a situation for which the proportional hazards assumption is *not* satisfied.

For our example, we consider a study in which cancer patients are randomized to either surgery or radiation therapy without surgery. Thus, we have a (0,1) exposure variable denoting surgery status, with 0 if a patient receives surgery and 1 if not. Suppose further that this exposure variable is the only variable of interest, so that a Cox PH model for the analysis of this data, as shown here, will contain only the one variable E, denoting exposure.

Now the question we consider here is whether the above Cox model containing the variable E is an appropriate model to use for this situation. To answer this question we note that when a patient undergoes serious surgery, as when removing a cancerous tumor, there is usually a high risk for complications from surgery or perhaps even death early in the recovery process, and once the patient gets past this early critical period, the benefits of surgery, if any, can then be observed.

Thus, in a study that compares surgery to no surgery, we might expect to see hazard functions for each group that appear as shown here. Notice that these two functions cross at about 3 days, and that prior to 3 days, the hazard for the surgery group is higher than the hazard for the no surgery group, whereas after 3 days, the hazard for the surgery group is lower than the hazard for the no surgery group.

Looking at the above graph more closely, we can see that at 2 days, when $t = 2$, the hazard ratio of non-surgery ($E = 1$) to surgery ($E = 0$) patients yields a value less than 1. In contrast, at $t = 5$ days, the hazard ratio of nonsurgery to surgery yields a value greater than 1.

EXAMPLE: (continued)

Given the above description, ***HR* is not constant over time**.

Cox PH model inappropriate because PH model assumes constant *HR*:

$h(t,\mathbf{X}) = h_0(t)e^{\beta E}$

$\widehat{HR} = \dfrac{\hat{h}(t,\ E=1)}{\hat{h}(t,\ E=0)} = e^{\hat{\beta}}$

Thus, if the above description of the hazard functions for each group is accurate, the hazard ratios are not constant over time. That is, the hazard ratio is some number less than 1 before 3 days and greater than 1 after 3 days.

It is therefore inappropriate to use a Cox PH model for this situation, because the PH model assumes a constant hazard ratio across time, whereas our situation yields a hazard ratio that varies with time.

In fact, if we use a Cox PH model, shown here again, the estimated hazard ratio comparing exposed to unexposed patients at any time is given by the constant value e to the β "hat," which does not vary over time.

General rule:
If the hazards cross, then a Cox PH model is not appropriate.

This example illustrates the general rule that if the hazards cross, then the PH assumption cannot be met, so that a Cox PH model is inappropriate.

Analysis when Cox PH model not appropriate? See Chapters 5 and 6.

It is natural to ask at this point, if the Cox PH model is inappropriate, how should we carry out the analysis? The answer to this question is discussed in Chapters 5 and 6. However, we will give a brier reply with regard to the surgery study example just described.

EXAMPLE: (continued)

Surgery study analysis options:

- stratify by exposure (use KM curves)
- start analysis at 3 days; use Cox PH model
- fit PH model for < 3 days and for > 3 days; get $\widehat{HR}$ (< 3 days) and $\widehat{HR}$ (> 3 days)
- include time-dependent variable (e.g., E $\times$ t); use extended Cox model

Actually for the surgery study there are several oplions available for the analysis. These include:

- analyze by stratifying on the exposure variable; that is, do not fit any model, and, instead oblain Kaplan-Meier curves for each exposure group separately;
- start the analysis at three days, and use a Cox PH model on three-day survivors;
- fit Cox model for less than three days and a different Cox model lor greater than three days to get two difterent hazard ratio estimates, one for each of these two lime periods;
- fit a modified Cox model that includes a time-dependent variable which measures the interaction of exposure with time. This model is called an **extended Cox model**.

Different options may lead to different conclusions.

Further discussion of these options is given in subsequent chapters. We point out here, that different options may lead to different conclusions, so that the investigator may have to weigh the relative merits of each option in light of the data actually obtained before deciding on any particular option as best.

Hazards cross $\Rightarrow$ PH not met

but

? $\Rightarrow$ PH met

See Chapter 4: Evaluating PH Assumption

One final comment before concluding this section: although we have shown that when the hazards cross, the PH assumption is not met, we have not shown how to decide when the PH assumption is met. This is the subject of Chapter 4 entitled, "Evaluating the PH Assumption."

IX. The Cox Likelihood

Likelihood

- Typically based on outcome distribution
- Outcome distribution not specified for Cox model
- Cox likelihood based on order of events rather than their distribution
  - Called partial likelihood

Typically, the formulation of a likelihood function is based on the distribution of the outcome. However, one of the key features of the Cox model is that there is not an assumed distribution for the outcome variable (i.e., the time to event). Therefore, in contrast to a parametric model, a full likelihood based on the outcome distribution cannot be formulated for the Cox PH model. Instead, the construction of the **Cox likelihood is based on the observed order of events** rather than the joint distribution of events. Thus the Cox likelihood is called a "partial" likelihood.

Illustration

Scenario:

- Gary, Larry, Barry have lottery tickets
- Winning tickets chosen at times t_1, t_2, ...
- Each person ultimately chosen
- Can be chosen only once

To illustrate the idea underlying the formulation of the Cox model, consider the following scenario. Suppose Gary, Larry, and Barry are each given a lottery ticket. Winning tickets are chosen at times t_j ($j = 1,2, \ldots$). Assume each person is ultimately chosen and once a person is chosen he cannot be chosen again (i.e., he is out of the risk set). What is the probability that the order each person is chosen is first Barry, then Gary, and finally Larry?

Question:
What is the probability that the order chosen is as follows?

1. Barry
2. Gary
3. Larry

Answer:

$$\text{Probability} = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{6}$$

Barry Gary Larry

The probability that Barry's ticket is chosen before Gary's and Larry's is one out of three. Once Barry's ticket is chosen, it cannot be chosen again. The probability that Gary's ticket is then chosen before Larry's is one out of two. Once Barry's and Gary's tickets are chosen, they cannot be chosen again which means that Larry's ticket must be chosen last. This yields a probability of 1/6 for this given order of events (see left).

Scenario:

Barry – 4 tickets
Gary – 1 ticket
Larry – 2 tickets

Now consider a modification of the previous scenario. Suppose Barry has 4 tickets, Gary has 1 ticket, and Larry has 2 tickets; now what is the probability that the order each person is chosen is first Barry, then Gary, and finally Larry?

Question:
What is the probability that the order chosen is as follows?

1. Barry
2. Gary
3. Larry

Barry, Gary, and Larry have 7 tickets in all and Barry owns 4 of them so Barry's probability of being chosen first is 4 out of 7. After Barry is chosen, Gary has 1 of the 3 remaining tickets and after Barry and Gary are chosen, Larry owns the remaining 2 tickets. This yields a probability of 4/21 for this order (see left).

Answer:

$$\text{Probability} = \frac{4}{7} \times \frac{1}{3} \times \frac{2}{2} = \frac{4}{21}$$

For this scenario

Subject's number of tickets affects probability

For Cox model

Subject's pattern of covariates affects likelihood of ordered events

For this scenario, the probability of a particular order is affected by the number of tickets held by each subject. For a Cox model, the likelihood of the observed order of events is affected by the pattern of covariates of each subject.

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

SURVT = Survival time (in years)
STATUS = 1 for event, 0 for censorship
SMOKE = 1 for a smoker, 0 for a nonsmoker

To illustrate this connection, consider the dataset shown on the left. The data indicate that Barry got the event at TIME = 2 years. Gary got the event at 3 years, Harry was censored at 5 years, and Larry got the event at 8 years. Furthermore, Barry and Larry were smokers whereas Gary and Harry were nonsmokers.

Cox PH model

$h(t) = h_0(t)e^{\beta_1 SMOKE}$

ID	Hazard
Barry	$h_0(t)e^{\beta_1}$
Gary	$h_0(t)e^0$
Harry	$h_0(t)e^0$
Larry	$h_0(t)e^{\beta_1}$

Individual hazards (Cox likelihood) analogous to number of tickets (lottery scenario) For example, smokers analogous to persons with extra lottery tickets

Consider the Cox proportional hazards model with one predictor, SMOKE. Under this model the hazards for Barry, Gary, Harry, and Larry can be expressed as shown on the left. The individual hazards are determined by whether the subject was a smoker or nonsmoker.

The individual level hazards play an analogous role toward the construction of the Cox likelihood as the number of tickets held by each subject plays for the calculation of the probabilities in the lottery scenario discussed earlier in this section. The subjects who smoke are analogous to persons given extra lottery tickets, thereby affecting the probability of a particular order of events.

Cox Likelihood

$$L = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

Likelihood is product of 3 terms

$$L = L_1 \times L_2 \times L_3$$

$$L_1 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right]$$

$$L_2 = \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right]$$

$$L_3 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

On the left is the Cox likelihood for these data. Notice the likelihood is a product of three terms, which correspond to the three event times. Barry got the event first at TIME = 2 years. At that time, all four subjects were at risk for the event. The first product (L_1) has the sum of the four subjects' hazards in the denominator and Barry's hazard in the numerator. Gary got the event next at 3 years when Gary, Harry, and Larry were still in the risk set. Consequently, the second product (L_2) has the sum of the three hazards for the subjects still at risk in the denominator and Gary's hazard in the numerator. Harry was censored at 5 years, which occurred between the second and third event. Therefore, when Larry got the final event at 8 years, nobody else was at risk for the event. As a result, the third product (L_3) just has Larry's hazard in the denominator and the numerator.

t_1, time = 2, four at risk (L_1)
t_2, time = 3, three at risk (L_2)
t_3, time = 8, one at risk (L_3)

For each term:

Numerator – single hazard
Denominator – sum of hazards

Baseline hazard, $h_0(t)$ cancels

$$L = \left[\frac{e^{\beta_1}}{e^{\beta_1} + e^0 + e^0 + e^{\beta_1}}\right] \times \left[\frac{e^0}{e^0 + e^0 + e^{\beta_1}}\right] \times \left[\frac{e^{\beta_1}}{e^{\beta_1}}\right]$$

Thus, L does not depend on $h_0(t)$

Data A

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

Data B

ID	TIME	STATUS	SMOKE
Barry	1	1	1
Gary	7	1	0
Harry	8	0	0
Larry	63	1	1

Comparing datasets

- TIME variable differs
- Order of events the same
- Cox PH likelihood the same

To summarize, the likelihood in our example consists of a product of three terms (L_1, L_2, and L_3) corresponding to the ordered failure times (t_1, t_2, and t_3). The denominator for the term corresponding to time t_j (j = 1, 2, 3) is the sum of the hazards for those subjects still at risk at time t_j, and the numerator is the hazard for the subject who got the event at t_j.

A key property of the Cox likelihood is that the baseline hazard cancels out in each term. Thus, the form of the baseline hazard need not be specified in a Cox model, as it plays no role in the estimation of the regression parameters. By factoring $h_0(t)$ in the denominator and then canceling it out of each term, the likelihood for Barry, Gary, and Larry can be rewritten as shown on the left.

As we mentioned earlier, the Cox likelihood is determined by the order of events and censorships and not by the distribution of the outcome variable. To illustrate this point, compare datasets A and B on the left, and consider the likelihood for a Cox PH model with smoking status as the only predictor. Although the values for the variable TIME differ in the two datasets, the Cox likelihood will be the same using either dataset because the order of the outcome (TIME) remains unchanged.

General Approach

- k failure times
- Likelihood a product of K terms
- Construction of each term similar to Barry, Gary, and Larry

$$L = L_1 \times L_2 \times L_3 \times \ldots \times L_k$$
$$= \prod_{f=1}^{k} L_f$$

We have used a small dataset (four observations with three failure times) for ease of illustration. However, the approach can be generalized. Consider a dataset with k failure times and let L_f denote the contribution to the likelihood corresponding to the f-th failure time. Then the Cox likelihood can be formulated as a product of each of the k terms as shown on the left. Each of the terms L_f is constructed in a similar manner as with the data for Gary, Larry, and Barry.

Obtaining maximum likelihood estimates

Solve system of equations

$$\frac{\partial \ln L}{\partial \beta_i} = 0, \quad i = 1, 2, 3, \ldots, p$$

$p = \#$ of parameters

Once the likelihood is formulated, the question becomes: *which values of the regression parameters would maximize L*? The process of maximizing the likelihood is typically carried out by setting the partial derivative of the natural log of L to zero and then solving the system of equations (called the score equations).

X. Using Age as the Time Scale

Outcome variable:
time until an event occurs

where "time" is measured as

time-on-study (years, months, etc., of follow-up from study entry)

or

age at follow-up

Recall that when we introduced the topic of survival analysis in Chapter 1, we wrote that the "time" variable used as the outcome variable could be measured as **time-on-study** (i.e., follow-up time since study entry) in years, months, weeks, or days from the beginning of follow-up. We also wrote that, alternatively, we might use **age as the time scale**, so that time is measured as **age at follow-up** until either an event or censorship occurs. In this section, we focus on the use of age as the time scale, and describe when such use is appropriate, provide the form of the Cox PH model in this situation, and illustrate its use.

Time 0:
starting time of the true survival time

Possible choices for time 0:

- Study entry
- Beginning of treatment
- Disease onset
- Disease diagnosis
- Surgery
- Point in calendar time
- Birth
- Conception

A key decision in any survival analysis is where to define the starting point for determining individual's "true" survival time, which we call **time 0**. Depending on the study, choices for time 0 might be: the time the subject enters the study, the time the subject begins treatment, the time of disease onset, the time of diagnosis, a point in calendar time, the time of a seminal event (e.g., surgery), birth, or conception. If we define time 0 at birth, then an individual's survival time is represented by their age.

Time 0 not necessarily equal to t_0, where

t_0 = time when subject's survival time is first observed

e.g., if survival time is measured by age and subject enters study at age 45

$\Downarrow$

t_0 = age 45 but time 0 < age 45 since time 0 = age-at-birth

Time 0 is not necessarily the time point where a subject's survival time is first observed (which we call time **t_0**). For example, if survival time is measured by age at follow-up and a subject enters the study at age 45, then t_0=45 years for this subject. In this example, the subject's survival time has been **left-truncated** at $t_0 = 45$, which we now define.

Left truncation:

- subject not observed before t_0
- if subject has event before t_0, then not included in the study
- if subject has event after t_0, then included in the study and assumed not at risk for event until t_0

Left truncation at time t_0 is defined as follows:

The subject is not observed from time 0 to t_0. If the subject has the event before time t_0, then that subject is not included in the study. If the subject has the event after time t_0, the subject is included in the study but with the caveat that the subject was not at risk to be an *observed* event until time t_0.

Two types of left truncation:

Type 1: subject has event before t_0 and not included in the study,

e.g.,

E causes death before study entry

$\Downarrow$

Bias: effect of E underestimated

We note that there are two types of left truncation at t_0. The first type of left truncation occurs if the subject has the event before t_0 and thus is not included in the study. If, for example, the exposure (E) under study causes individuals to die before they could enter the study, this could lead to a (selective) **survival bias** that would underestimate the effect of exposure.

Type 2: $t_0 > 0$ and $t > t_0$

where t = observed survival time

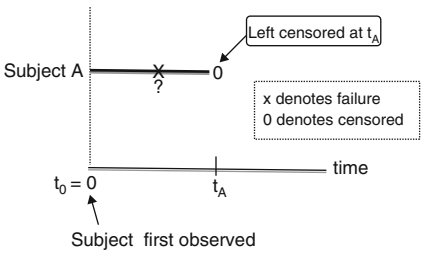
The second type of left truncation occurs if the subject survives beyond time t_0 (i.e., $t > t_0$). This is required in order for the subject to have his/her survival time observed.

Study entry $\Rightarrow$ subject survives until t_0

Type 1: subject not included in the study

Type 2: subject included in the study

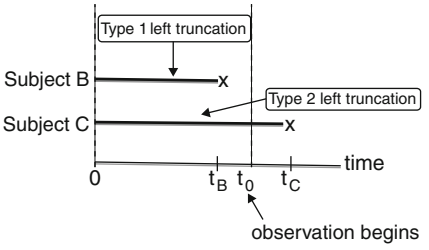
Thus, a condition of the subject's entry into the study is that they survive until time t_0. If they do not meet that condition, then their left truncation is of the first type and thus not included in the study. If they do survive past time t_0, then their left truncation is of the second type.



Left truncation (of both types) at time t is commonly confused with left censorship at time t. If a subject is **left censored at time t**, then that subject is (i) included in the study, (ii) known to be event free at time 0, (iii) known to be at risk for the event after time 0, and (iv) known to have had the event before time t but with the exact time of event being unknown.

Example: left censored data
Subject A: $t_0 = 0$
$t_A = 6$
true $t = ? < 6$

For example, for subject A in the above graph, suppose death is the outcome and a patient is first treated for some illness at the time of their first visit to the clinic (t_0=0). Further, suppose that the patient does not show up at the next scheduled clinic visit 6 months later because that patient had died in the interim. If the specific month of death is unable to be ascertained, then that patient is included in the study and **left censored** at $t_A = 6$ months.



In contrast, the diagram on the left illustrates the two types of left truncation.

In this diagram, subject B provides an example of left truncation of the first type that would occur if an individual died between disease onset (time 0) and disease diagnosis (time t_0) and thus was never included in the study. In this example, having the disease was a necessary condition for study inclusion, whereas subject B died before it was known that he/she had the disease.

Example: Type 1 left truncation
Subject B: Time $0 < t_0$
not included in the study
Example: Type 2 left truncation
Subject C: Time $0 < t_0$ but first observed at t_0

Subject C illustrates Type 2 left truncation, since he/she developed the disease (time 0) prior to being diagnosed with the disease at time t_0 and was observed after t_0.

Being observed at time t means:
If event at $t \Rightarrow$ recorded event at t

One clarifying point is that when we say a subject is observed at time t, we do not necessarily mean that the subject is observed in an active prospective manner. Rather, what we mean by a subject being observed at time t is as follows: *if that subject had an event at time t, then the subject would be recorded in the study as having had an event at time t.*

2 approaches for measuring survival time:

Time-on-study
vs.
Age-at-follow-up

We now compare two approaches of measuring survival time. One approach is to measure survival time as **time-on-study** and the other is to measure survival time as **age-at-follow-up** until either an event or censorship. The choice of approach determines the risk set at the time of each event. We illustrate this idea with hypothetical data.

- Choice determines the risk set.

Hypothetical Survival Data

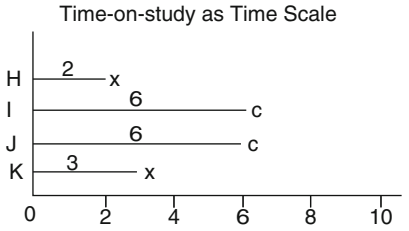
Subject	t	d	a_0	a
H	2	1	65	67
I	6	0	65	71
J	6	0	74	80
K	3	1	75	78

Consider the data shown on the left on four different subjects, for each of which we have identified time-on-follow-up (**t**), whether failed or censored (**d**), age at study entry ($\mathbf{a}_0$), and age at the end of follow-up (**a**). Note that t is simply $a - \mathbf{a}_0$, the difference between age at follow-up time and age at study entry.

Time-on-study Layout

f	$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
1	2	4	0	0	H,I,J,K
2	3	3	2	2	I,J,K

Using time-on-study (i.e., from entry into the study) as the time scale, the data layout based on ordered follow-up times is shown on the left, below which is shown a graphical representation that follows each subject from the time of study entry. There are only two failures and these occur at follow-up times 2 (subject H) and 3 (subject K).



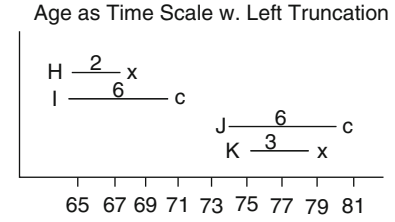
$R(t_{(1)} = 2) = \{H, I, J, K\}$
$R(t_{(2)} = 3) = \{I, J, K\}$
I and J censored after $t_{(2)} \Rightarrow q_2 = 2$
{I, J, K} contained in {H, I, J, K}

The risk set at the first failure time ($t_{(1)} = 2$) consists of all four subjects, and the risk set at the second failure time ($t_{(2)} = 3$) contains subjects I, J, and K. Subjects I and J are censored after time $t_{(2)} = 3$, i.e., the value for q_2 is 2 at this time. Notice that the risk set at time $t_{(2)} = 3$ is contained in the risk set at time $t_{(1)} = 2$, which is generally true when the outcome variable being considered is time-on-follow-up. The data layout represented here, in which the size of the risk set always decreases over time, is called a **closed cohort**.

Age as Time Scale Layout

f	$a_{(f)}$	n_f	m_f	q_f	$R(a_f)$
1	67	2	1	1	H,I
2	78	2	1	1	J,K

Now let's consider the data layout that would result if we used age as the time scale, which is shown on the left. Below this layout is a graphical representation that follows each subject from age at study entry.



First failure: $R(a_{(1)} = 67) = \{H, I\}$

I still at risk at $a_{(1)} = 67$ but
J and K not in study at $a_{(1)} = 67$

Using age as the time scale, the first failure time is at age $a_{(1)} = 67$ (for subject H), and there are two subjects (H and I) in the risk set at this time; subject I is in the risk set at $a_{(1)} = 67$ because (s)he entered the study at age 65 and was still at risk when subject H failed. However, because subjects J and K did not enter the study until ages 74 and 75, these subjects are not in the risk set at $a_{(1)} = 67$.

Second failure: $R(a_{(2)} = 78) = \{J, K\}$

H and J no longer at risk at $a_{(2)} = 78$
I censored between $a_{(1)} = 67$ and $a_{(2)} = 78 \Rightarrow q_1 = 1$
J censored after $a_{(2)} = 78 \Rightarrow q_2 = 1$

{J, K} not contained in {H, I}

The second failure time is at age $a_{(2)} = 78$ (for subject K). The only two subjects in the risk set at $a_{(2)} = 78$ are subjects J and K, since subject H failed at age 67 and subject I was censored at age 71. The values in the q column are 1 at failure age 67 (for subject I) and 1 at failure age 72 (for subject J). In contrast to the previous data layout, the risk set at the later failure age (containing J, K) is not a subset of the risk set at the first failure age (containing H, I), but, rather, is a mutually exclusive subset. This data layout, in which the size of the risk set may increase or decrease over time, is called an **open cohort**.

Time-on-Study vs. Age as Time Scale

- Closed cohort vs. Open cohort
- How we decide which to use?

We thus see that using time-on-study as the time scale can give a different view of the survival data (i.e., a closed cohort) than found when using age as the time scale (i.e., an open cohort). So which time scale should be used and how do we make such a decision in general?

Key issue:
Did all subjects first become at risk at their study entry?

To answer this question, a key issue is to determine whether all subjects in the study first begin to be at risk for the outcome at the time they enter the study.

Clinical trial:

- Subjects start to be followed for the outcome after random allocation

Suppose the study is a **clinical trial** to compare, say, treatment and placebo groups, and subjects start to be followed shortly after random allocation into one of these two groups.

- Reasonable to assume subjects start to be at risk upon study entry

⇓

Time-on-Study typically used as the outcome

(Covariates may also be controlled)

Then, it may be reasonable to assume that study subjects begin to be at risk for the outcome upon entry into the study. In such a situation, using time-on-study as the time scale is typically appropriate. Further, covariates of interest may be controlled for by stratification and/or being entered into a regression model (e.g., Cox PH model) as predictors in addition to the treatment status variable.

Observational study:

- Subjects already at risk prior to study entry
- Unknown time or age when first at risk

Suppose, instead of the above scenario, the **study is observational** (i.e., not a clinical trial) and subjects are already at risk for the outcome prior to their study entry. Also, suppose the time or age at which subjects first became at risk is unknown.

- Example: Subjects with high blood pressure enter study, but unknown date or age when first diagnosed (prior to study entry).

For example, the subjects may all have high blood pressure when the study begins and are then followed until a coronary event occurs (or censorship); such subjects already had high blood pressure when recruited for the study, but the date or their age when their high blood pressure condition was first diagnosed is assumed unknown.

- Reasonable to assume that

$$T = t_r + t$$

where
T = true survival time
t_r = time at risk prior to study entry
t = observed time-on-study

In this situation, it seems reasonable that the time at risk prior to study entry ($\mathbf{t_r}$), which is unknown, contributes to the true survival time (**T**) for the individual, although only the observed time-on-study (**t**) is actually available to analyze. The individual's true (i.e., total) survival time is therefore underestimated by the time-on-study information (obtained from study entry), i.e., the true survival time is **left-truncated**.

Left-truncated survival data
⇓
Time-on-study questionable

So, for the situation where we have left-truncated survival data, the use of time-on-study follow-up times that ignores unknown delayed entry time may be questioned.

Subject	**t**	**d**	$\mathbf{a_0}$	**a**
H	2	1	65	67
I	6	0	65	71
J	6	0	74	80
K	3	1	75	78

Recall that although both subjects I and J were censored at follow-up time 6 (say, in weeks) from study entry, subject I entered the study at age 65, whereas subject J entered the study at age 74. Because subject J is 9 years older than subject I upon study entry, and recognizing that age is a well-known risk factor for most diseases, e.g., coronary disease, we would expect subject J to have higher potential for failing (i.e., higher hazard rate) than subject I at study entry. However, if we just use time-on-study follow-up times in our analysis, we are not taking into account the increased failure potential for subject J over subject I at study entry.

Subject J is 9 years older than Subject I
⇓
h(t| subject J) > h(t| subject I).
But, using time-on-study approach does not account for this difference.

One modified approach:
Use time-on-study, but control for a_0, e.g.,

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp[\sum \beta_i X_i + \gamma a_0]$$

⇓
OK provided model correctly specified but not always appropriate.

One way to account for the age difference at entry would simply to control for age at entry (i.e., $\mathbf{a_0}$) as a covariate in one's survival analysis by adding the variable $\mathbf{a_0}$ to a Cox PH model. This approach is reasonable provided the model is specified correctly (e.g., proportional hazards assumption is met for age).

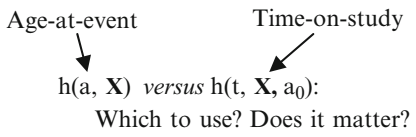
Alternatively, may consider using age as the time scale.

Alternatively, considering subjects I and J, who have entered at the same time but are 9 years different in age, we might consider using age as the time scale to represent a subject's potential for failure, which we now describe.

$h(a, \mathbf{X}) = h_0(a) \exp[\sum \beta_i X_i]$

X denotes set of covariates,
e.g., **X** = (Rx, BMI, SMK)
$h_0(a)$ = baseline hazard

In a Cox PH model that uses age as the time scale (as shown on the left), the outcome variable will be age-at-event (a) rather than time-on-study (t). **X** denotes the set of covariates in the model, e.g., **X** = (Rx, BMI, SMK). The baseline hazard function $h_0(a)$ is an unspecified function of **a** (rather than **t**).



At this point, we might again ask, when, if at all, would using a model based on h(a, **X**) be preferable to simply using a model of the form h(t, **X**, a_0) where t denotes time-on follow-up, and a_0 denotes age at entry?

It depends!
And, it might not matter!
(often same results, if model well-specified)

The answer is that "it depends". Moreover, in many situations, it might not matter, since use of either model form will often lead to essentially the same results, provided the model is well-specified in each case.

Prefer h(a, **X**) provided

- age is stronger determinant of outcome than time-on-study
- $h_0(a)$ is unspecified, so that age is not modeled as a covariate i.e., avoids mispecifying the model as linear in a_0 when a_0^2 also needed

On one hand, using h(a, **X**) may be preferable if age is a much stronger determinant of the outcome than time-on-study, i.e., age at event may have a larger effect on the hazard than time-on-study (Korn et al. 1997). Also, because age is taken into account in an unspecified baseline hazard $h_0(a)$, a more effective control of age may result that avoids the possibility of mispecifying the way the age at entry ($\mathbf{a_0}$) might be entered into a time-on-study model, e.g., using only a linear term when a quadratic term such as is also required for model adequacy.

Prefer h(t, **X**, a_0) provided

- time-on-study is stronger determinant of outcome than age
- age at entry ($\mathbf{a_0}$) is effectively controlled in the model using a linear and/ or possibly higher/ order term (or age is controlled by stratification)

On the other hand, h(t, **X**, a_0) may be preferable if time-on-study is a stronger determinant of the outcome than age at the event, as in a randomized clinical trial. Also, a time-on-study model would seem appropriate if age at entry ($\mathbf{a_0}$) is "effectively controlled" (e.g., using a quadratic term if necessary) or is stratified in the model.

Alternative Cox PH models for age-truncated survival data:

Let

t = follow-up time,

a = attained age at event or censorship

$\mathbf{a}_0$ = age at enrollment into study (Note: $t = a - a_0$)

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$, vector of predictors, not including a_0

β_i = regression coeff. corresponding to X_i.

γ = regression coeff. if a_0 included in model

We now specify several alternative forms that a Cox PH model might take to account for risk-truncated survival data. As previously introduced, our notation uses **t** to denote time-on-study follow-up time, **a** to denote attained age at the event or censorship, $\mathbf{a_0}$ to denote age at study entry, **X** to denote the vector of predictor variables, not including age, β_i to denote the vector of Cox model coefficients corresponding to **X**, and γ_1 to denote the coefficient of a_0 if model includes a_0.

time on study

Model 0:
$h(t,\mathbf{X}) = h_0(t)\exp[\Sigma\beta_i X_i]$,
unadjusted for a_0

Model 1:
$h(t,\mathbf{X},a_0) = h_0(t)\exp[\Sigma\beta_i X_i + \gamma_1 a_0]$,
adjusted for a_0 as linear covariate

Model 2:
$h(t,\mathbf{X},a_0) = h_0(t)\exp[\Sigma\beta_i X_i + \gamma_1 a_0 + \gamma_2 a_0^2]$
adjusted for a_0 with quadratic covariate

Model 3:
$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\Sigma\beta_i X_i]$,
stratified by a_0 or birth cohort, g= 1, …, s

On the left, we provide seven different Cox PH models that might be considered to analyze risk-truncated survival data.

Models 0–3 use an analysis based on time-on-study follow-up, whereas Models 4–6 consider age as the time scale.

age as time scale

Model 4:
$h(a,\mathbf{X}) = h_0(a)\exp[\Sigma\beta_i X_i]$,
unadjusted for left truncation at a_0

Model 5:
$h(a,\mathbf{X}) = h_0(a|a_0)\exp[\Sigma\beta_i X_i]$,
adjusted for left truncation at a_0

Model 6:
$h_g(a,\mathbf{X}) = h_{0g}(a|a_0)\exp[\Sigma\beta_i X_i]$,
adjusted for left truncation at a_0 and stratified by birth cohort, g= 1, …, s

Of all these models, Model 0 is the least appropriate since this model uses time-on-study as the outcome and does not adjust for age at entry ($\mathbf{a_0}$) in any way.

Models 1–3 control for a_0 differently
- Model 1: linear effect of $\mathbf{a_0}$
- Model 2: quadratic effect of $\mathbf{a_0}$
- Model 3: stratifies on $\mathbf{a_0}$ or on birth cohorts defined from $\mathbf{a_0}$ (uses Stratified Cox PH model)

Models 1–3 are time-on-study models that control for age at entry ($\mathbf{a_0}$), but do so differently. Model 1 controls for $\mathbf{a_0}$ as a covariate and assumes a **linear effect of a_0**. Model 2, in contrast, assumes that $\mathbf{a_0}$ has **both linear and quadratic effects**. Model 3 **stratifies on either $\mathbf{a_0}$ or on birth cohort** defined from $\mathbf{a_0}$. Model 3 is called a Stratified Cox (SC) PH model, which we describe in detail in Chapter 5.

Models 1–3 reasonable

- if all study subjects begin risk at study entry
- if models provide effective control of $\mathbf{a_0}$

Models 1–3 are all reasonable if we assume that study subjects begin to be at risk upon study entry, as in a randomized clinical trial. Moreover, even for an observational study design in which subjects have different ages at entry, these models may appear justifiable if they provide effective control of $\mathbf{a_0}$.

Model 3
$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\sum \beta_i X_i]$

- alternative method of control
- may account for advances in medical management if stratified on birth cohort
- stratifying on either $\mathbf{a_0}$ or on birth cohort likely to give similar results unless enrollment over long time period

Model 3 controls for entry age by stratifying either on age at entry ($\mathbf{a_0}$) or on birth cohort based on $\mathbf{a_0}$. Model 3 provides an alternative way to control for age without explicitly putting $\mathbf{a_0}$ as a covariate in the model (as was done in Models 1 and 2). If we stratify by birth cohort instead of by $\mathbf{a_0}$, we can account for possible advances in medical management in later birth cohorts. Nevertheless, stratifying by age at entry or stratifying by birth cohort would likely give similar results unless enrollment happens over a long period of time. In the latter case, we recommend stratifying on birth cohort.

Models 4–6:

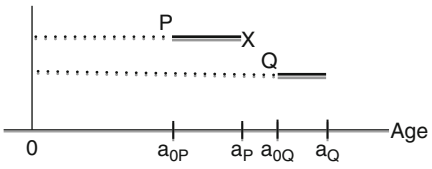
- outcome is age-at-event
- differ in baseline hazard

Models 4–6 use age-at-event or censorship rather than time-on-study as the outcome variable. These models differ in the way the baseline hazard function is specified.

Model 4: $h(a, \mathbf{X}) = h_0(a) \exp[\sum \beta_i X_i]$

- does not adjust for left-truncation at a_0
- assumes risk starts at birth
- data layout describes **closed cohort**

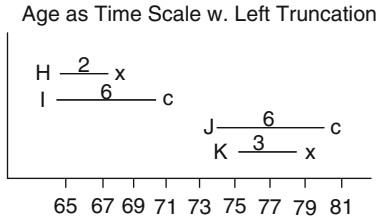
Model 4 uses $h_0(a)$ to indicate that, although age is the outcome, the model does not adjust for left truncation at the entry age (a_0). In effect, this baseline hazard assumes that each subject's *observed risk period* started at birth.



R(a) = {P, Q} using Model 4 even though Q enrolled after P failed

In other words, Model 4 allows keeping in the risk set R($\mathbf{a_P}$) any subject (e.g., subject Q in the figure at left) who was not under study at age $\mathbf{a_P}$ but who enrolled later (at age $\mathbf{a_{0Q}}$). Here, subject Q is in the risk set R($\mathbf{a_P}$) because we assume he is at risk from birth (Age=0) when subject P fails at $\mathbf{a_P}$. The data layout with ordered failure ages is thus a **closed cohort** that starts with all subjects in the risk set at birth.

Previous example:

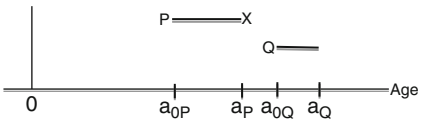


R(a = 67) = {H, I, J, K} using Model 4 since all four subjects at risk from birth x until H fails at age a = 67

If Model 4 were applied to our previous example involving four subjects, subjects J and K would be incorrectly included in the risk set R (**a**=67) when subject H failed, even though both these subjects were enrolled after age 67. This model ***inappropriately*** assumes that all subjects were at risk from birth; it does not adjust for age-truncation.

Model 5: $h(a, \mathbf{X}) = h_0(a|a_0) \exp[\sum \beta_i X_i]$

- adjusts for left-truncation at a_0
- data layout describes **open cohort**



R(a) = {P} and R(a*) = {Q} using Model 5 because Q enrolled after P failed

Model 5, on the other hand, accounts for left truncation by age at entry. The baseline hazard $h_0(a \mid a_0)$ is used to indicate that the data layout with ordered failure ages is an **open cohort**. For this model, the risk set R(**a**) at time **a** contains only those subjects who are under study at age **a**.

Previous example with H, I, J, K: R(a=67) = {H, I} and R(a=78) = {J, K} using Model 5 since J and K had not enrolled when H failed at 67 and {H, I} were not used in study when K failed at 78.

If Model 5 was applied to our previous example, subjects J and K, who had not enrolled when subject H failed at 67, would not be in the risk set R(**a**=67). Also, subjects H and I, who were no longer in the study when subject K failed at age 78, would not be in the risk set R(**a**=78).

	Model 3	Model 6
Stratifies on birth cohort?	Yes	Yes
Adjusts for age-truncation?	No	Yes

Model 6 is similar to Model 3 stratified on birth cohort. However, Model 6 adjusts for age truncation, whereas Model 3 does not. As with Model 3, Model 6 is intended to account for possible advances in medical management in later birth cohorts. Model 6 would not be necessary if we are considering a study in which everyone is enrolled within a short period of time.

Summary about Models 0–6:
Models 0 and 4:

- Both inappropriate
- Model 0 does not adjust for age
- Model 4 incorrectly assumes that all subjects are at risk from birth

Models 1–3, 5, 6

- All adjust for age-at-entry (a_0)
- Question: Do they differ in practice?

Pencina et al. (2007):

- Compare estimated regression coefficients for Models 1–6
- Consider Model 5 (age-truncated age scale) most appropriate conceptually
- Consider Models 1 and 2 (covariate adjusted for a_0) "attempts to approximate Model 5"
- Used numerical simulations and practical examples from Framingham

Conclusions:

- correct adjustment for the age at entry is crucial
- Model 1 inferior (and biased)
- Little practical or meaningful difference between Models 2 through 6

Cox PH Regression Coefficients (± se) for two CHD risk factors among men- Framingham Heart Disease Study (Pencina et al, 2007)

Time-on-study	Model					Age-time-scale
	1	2	3	**4**	5	6
	linear	quad	strat	unadj	age-trunc	strat
Diabetic *versus* non-diabetic (n=2439)						
	0.48* ± 0.21	0.49* ± 0.21	0.48* ± 0.21	**0.23*** ± 0.20	0.47* ± 0.21	0.45* ± 0.21
Education: post– HS *versus* HS or less (n = 2177)						
	–0.43* ± 0.15	–0.40* ± 0.15	–0.43* ± 0.15	**0.18** ± 0.15	–0.43* ± 0.16	–0.38* ± 0.15

*The coefficient is significantly different from zero at the 0.05 level.

In summary, of the seven models we have presented, Models 0 and 4 are inappropriate because Model 0 does not account for age at all and Model 4 ignores age truncation by incorrectly assuming that all study subjects were observed for the outcome from birth.

The other five models (i.e., 1–3, 5, 6) all adjust for age at study entry in some way. A logical question at this point is whether in practice, it makes a difference which model is used to analyze age-truncated survival data?

The above question was actually addressed by Pencina et al. (Statist. Med., 2007) by comparing Models 1–6 above in terms of the estimated regression coefficients they produce. These authors consider Model 5, the age-truncated age scale model, to be "possibly the most appropriate refinement" to account for age-truncation. They also view time-on-study Models 1 and 2, which use linear and/or quadratic terms to adjust for entry age as a covariate as "attempts to approximate" Model 5.

Nevertheless, by considering numerical simulations as well as four practical examples from the Framingham Heart Study, Pencina et al. conclude that **correct adjustment for the age at entry is crucial** in reducing bias of the estimated coefficients. The unadjusted age-scale model (Model 1) is inferior to any of the five other models considered, regardless of their choice of time scale. Moreover, if correct adjustment for age at entry is made when considering Models 2–6, their analyses suggest that **there exists little if any practical or meaningful difference in the estimated regression coefficients depending on the choice of time scale**.

To illustrate, we show on the left results from Pencina et al. corresponding to Models 1–6 applied to 12-year follow-up Framingham Heart Study data. The outcome considered here is coronary heart disease (**CHD**) in men.

These results focus on two risk factors measured at baseline: diabetes mellitus status and education status, the latter categorized into two groups defined by post-high-school education (yes/no). The estimated regression coefficients (separately) relating these two risk factors to **CHD** outcome are presented in the table.

Summary of Framingham results from Pencina et al.:

- Model 4 inferior to other models
- Results for Models 1–3, 5, and 6 are similar
- Directions of estimated coefficients were as anticipated conceptually, e.g., + diabetes and smoking − education
- Quadratic terms (Model 2) were significant, suggesting that Model 1 is mispecified but
- Did not materially influence magnitude or significance of exposure variables (e.g., diabetes, smoking, education)

As expected, the table shows a substantial difference in the coefficient of the risk group variable estimated by the unadjusted age-scale model (Model 4) and the five other models. Moreover, the results for Models 1–3, 5, and 6 are all quite similar.

Pencina also point out that the directions of coefficients for these five models are in the directions anticipated conceptually, e.g., diabetes coefficients are positive, whereas education coefficients are negative.

The quadratic baseline age term (Model 2) was significant for both CHD risk factors. This suggests potential misspecification in the modeling of the relationship between CHD and age introduced by Model 1, which treats entry age as linear. However, its inclusion in the time-on-study model did not materially influence the magnitude or significance of the estimated exposure variable (diabetes or education status) coefficient.

Data Layout for Age-as-Time Scale: **CP format for age-truncated survival data (Model 5)**

Subj#	**d**	$\mathbf{a_0}$	**a**	$\mathbf{X_1}$	...	$\mathbf{X_p}$
1	d_1	a_{01}	a_1	X_{11}	...	X_{1p}
2	d_2	a_{02}	a_2	X_{21}	...	X_{2p}
3	d_3	a_{03}	a_3	X_{31}	...	X_{3p}
⋮	⋮	⋮	⋮	⋮	...	⋮
n	d_n	a_{0n}	a_n	X_{n1}	...	X_{np}

When using age as the time scale and accounting for age truncation (i.e., using Model 5 above), the data layout requires the counting process (**CP**) in start–stop format previously introduced in Section VI of Chapter 1 with a_0 as the start variable and a as the stop variable. However, since we are not considering recurrent events data here, the CP format for age-truncated survival data has a simpler form, involving only one line of data for each study subject, as shown on the left. The computer code needed to program the analysis is described in the Computer Appendix for STATA, SAS, SPSS, or R packages.

Model 4 layout: Set $a_0 = 0$ for all subjects or use "standard" layout (w/0 a_0 column)

Subj#	**d**	**a**	$\mathbf{X_1}$	...	$\mathbf{X_p}$
1	d_1	a_1	X_{11}	...	X_{1p}
2	d_2	a_2	X_{21}	...	X_{2p}
3	d_3	a_3	X_{31}	...	X_{3p}
⋮	⋮	⋮	⋮	...	⋮
n	d_n	a_n	X_{n1}	...	X_{np}

Note that the CP format corresponding to Model 4, which assumes the starting time is birth, would modify the Model 4 layout by letting $a_0 = 0$ in the a_0 column for all subjects. Nevertheless, this layout would be equivalent to the "standard" layout that omits **a_0 column** and simply treats the **a column** data as time-on-study information. Again, since Model 4 appears to be inferior to the other models, we caution the reader not to use this format unless the risk period was observed since birth.

XI. Summary

In this section we briefly summarize the content covered in this presentation.

1. Review: $S(t)$, $h(t)$, data layout, etc.
2. Computer example of Cox model:
   - estimate *HR*
   - test hypothesis about *HR*
   - obtain confidence intervals
3. Cox model formula:

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

4. Why popular: Cox PH model is "robust"

5. ML estimation: maximize a partial likelihood $L = L(\beta)$ = joint probability of observed data

6. Hazard ratio formula:

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

7. Interval estimation-interaction: HR=exp[ℓ],

where $\ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_k W_K$
β_1 = coeff. of X_1, and
δ_j = coeff. of $X \times W_j$, $j = 1, \ldots, k$

95% CI for HR=exp[ℓ]:
$\exp[\hat{\ell} \pm 1.96\sqrt{\widehat{\text{Var}}\hat{\ell})}]$

Most computer packages, e.g., SAS, STATA, compute $\widehat{\text{Var}}\hat{\ell}$ as part of the program options (see Computer Appendix).

- We began with a computer example that uses the Cox PH model. We showed how to use the output to estimate the *HR*, and how to test hypotheses and obtain confidence intervals about the hazard ratio.
- We then provided the formula for the hazard function for the Cox PH model and described basic features of this model. The most important feature is that the model contains two components, namely, a baseline hazard function of time and an exponential function involving X's but not time.
- We discussed reasons why the Cox model is popular, the primary reason being that the model is "robust" for many different survival analysis situations.
- We then discussed ML estimation of the parameters in the Cox model, and pointed out that the ML procedure maximizes a "partial" likelihood that focuses on probabilities at failure times only.
- Next, we gave a general formula for estimating a hazard ratio that compared two specifications of the X's, defined as $\mathbf{X}^*$ and $\mathbf{X}$. We illustrated the use of this formula when comparing two exposure groups adjusted for other variables.
- We then described how to obtain a 95% CI for the HR when the hazard model contains interaction terms of the form $X_1 \times W_j$, where X_1 is a (0,1) exposure variable and W_j is an effect modifier of exposure. The formula is shown at the left. In this formula, the $\widehat{\text{Var}}\,\hat{\ell}$ is difficult to calculate without the use of a computer program.

  Fortunately, most computer packages have procedures for calculating this formula as part of the program options, e.g., SAS's "contrast" option and STATA's "lincom" option.

8. Adjusted survival curves: 0 or 1
   Comparing E groups:
   $$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1 E + \sum_{i\neq 1}\hat{\beta}_i \bar{X}_i\right]}$$
   Single curve:
   $$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp[\sum \hat{\beta}_i \bar{X}_i]}$$

9. PH assumption:
   $$\frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \hat{\theta} \text{ (a constant over } t)$$
   i.e., $\hat{h}(t,\mathbf{X}^*) = \hat{\theta}\hat{h}(t,\mathbf{X})$
   Hazards cross $\Rightarrow$ PH not met

10. Derivation of Cox PH Likelihood

11. Using "age-as-the-time scale" instead of "time-on-follow-up"
    Reason: account for left truncation of age
    Cox PH model that adjusts for age truncation:
    $$h(a,X) = h_0(a|a_0)\exp[\sum \beta_i X_i]$$
    where $\mathbf{a}$ = age at event or censorship
    $\mathbf{a_0}$ = age at study entry
    Data Layout: CP (start–stop) format

- We then defined an adjusted survival curve and presented formulas for adjusted curves comparing two groups adjusted for other variables in the model and a formula for a single adjusted 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.
- We described the PH assumption as meaning that the hazard ratio is constant over time, or equivalently that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. We also showed that for study situations in which the hazards cross, the PH assumption is not met.
- We then showed how the Cox likelihood is derived using ordered failure times.
- Finally, we considered the use of "age as the time scale" instead of "time-on-follow-up" as the outcome variable, described why such use is appropriate to account for left truncation of age, provided the form of the Cox PH model in this situation, illustrate its use, and described the data layout required using a "stop-start" Counting Process (CP) format.

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

This presentation is now complete. We recommend that the reader review the detailed outline that follows and then do the practice exercises and test.

The next Chapter (4) describes how to evaluate the PH assumption. Chapters 5 and 6 describe methods for carrying out the analysis when the PH assumption is not met.

Detailed Outline

I. A computer example using the Cox PH model (pages 100–108)

A. Printout shown for three models involving leukemia remission data.

B. Three explanatory variables of interest: treatment status, log WBC, and product term; outcome is time until subject goes out of remission.

C. Discussion of how to evaluate which model is best.

D. Similarity to classical regression and logistic regression.

II. The formula for the Cox PH model (pages 108–110)

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

B. $h_0(t)$ is called the **baseline hazard function**.

C. $\mathbf{X}$ denotes a collection of p explanatory variables $X_1, X_2, \ldots, X_p$.

D. The model is semiparametric because $h_0(t)$ is unspecified.

E. Examples of the Cox model using the leukemia remission data.

F. Survival curves can be derived from the Cox PH model.

III. Why the Cox PH model is popular (pages 110–112)

A. Can get an estimate of effect (the hazard ratio) without needing to know $h_0(t)$.

B. Can estimate $h_0(t)$, $h(t, \mathbf{X})$, and survivor functions, even though $h_0(t)$ is not specified.

C. The e part of the formula is used to ensure that the fitted hazard is nonnegative.

D. The Cox model is "robust": it usually fits the data well no matter which parametric model is appropriate.

IV. ML estimation of the Cox PH model (pages 112–114)

A. Likelihood function is maximized.

B. L is called a partial likelihood, because it uses survival time information only on failures, and does not use censored information explicitly.

C. L makes use of the risk set at each time that a subject fails.

D. Inferences are made using standard large sample ML techniques, e.g., Wald or likelihood ratio tests and large sample confidence intervals based on asymptotic normality assumptions.

V. Computing the hazard ratio (pages 114–117)

A. Formula for hazard ratio comparing two individuals, and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

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

B. Examples are given using a (0, 1) exposure variable, potential confounders, and potential effect modifiers.

C. Typical coding identifies $\mathbf{X}^*$ as the group with the larger hazard and $\mathbf{X}$ as the group with the smaller hazard, e.g., $X_1^* = 1$ for unexposed group and $X_1 = 0$ for exposed group.

VI. Interval estimation: interaction (pages 117–119)

A. Example- Model 3 from Remission Time Data

i. $h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 logWBC + \beta_3(Rx \times logWBC)]$

ii. $HR = \exp[\beta_1 + \beta_3 logWBC]$

B. General form of HR:
$HR = \exp[\ell]$, where
$\ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_k W_k$,
$X_1 = (0, 1)$ exposure variable, β_1 = coeff of X_1, and δ_j = coeff of $X_1 \times W_j$, j=1,..., k

C. General form of 95% CI for $HR = \exp[\ell]$:

$$\exp\left[\hat{\ell} \pm 1.96\sqrt{\hat{Var}\,\hat{\ell})}\right], \text{ where}$$

$$\text{Var}\left(\hat{\ell}\right) = \text{Var}\left(\hat{\beta}_1 + \hat{\delta}_1 W_1 + \ldots + \hat{\delta}_k W_k\right)$$

D. Computation of var complicated.

i. Computer programs, e.g., in SAS, STATA, can do this for the user.

ii. Otherwise, user must carry out complicated calculation using formula for var:

$$\hat{Var}\left(\hat{\ell}\right) = \hat{Var}\left(\hat{\beta}_1\right) + \sum_j W_j^2 \hat{Var}\left(\hat{\delta}_j\right) + 2\sum_j W_j \hat{Cov}\left(\hat{\beta}_1, \hat{\delta}_j\right) + 2\sum_j \sum_k W_j W_k \hat{Cov}\left(\hat{\delta}_j, \hat{\delta}_k\right)$$

iii. Variances and covariances provided in computer output

iv. User specifies W's values of interest.

v. Model 3 formula for $\hat{Var}(\hat{\ell})$:

$$\hat{Var}\left(\hat{\ell}\right) = \hat{Var}\left(\hat{\beta}_1\right) + (\log WBC)^2 \hat{Var}\left(\hat{\beta}_3\right) + 2(\log WBC)^2 Cov\left(\hat{\beta}_1, \hat{\beta}_3\right)$$

E. Example of 95% CI: Model 3

VII. Adjusted survival curves using the Cox PH model (pages 120–123)

A. Survival curve formula can be obtained from hazard ratio formula:

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

where $S_0(t)$ is the baseline survival function that corresponds to the baseline hazard function $h_0(t)$.

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

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

D. Examples of "adjusted" $S(t, \mathbf{X})$ using leukemia remission data.

VIII. The meaning of the PH assumption (pages 123–127)

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

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

B. Baseline hazard function not involved in the HR formula.

C. Hazard ratio for two $\mathbf{X}$'s are proportional:

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

D. An example when the PH assumption is not satisfied: hazards cross

IX. The Cox likelihood (pages 127–131)

A. Lottery Example

B. Likelihood based on order of events

X. Using age as the time scale (pages 131–142)

A. Definition of Left Truncation

i. Type I: subject has event before t_0 and not included in study

ii. Type II: $t_0 > 0$ and $t > t_0$ where t_0 time when first observed t = observed survival time

B. Left Truncation versus Left Censoring

C. Time-on-study versus Age-as-time scale: Closed cohort versus Open cohort

D. When to use Age-as-time-scale

i. It depends

a. Type of study

b. Well-defined model involving a_0 where a_0 denotes age at entry

E. Alternative Models
   i. Time-on-study
      a. outcome is t = time since first observed
      b. consider control for a_0, e.g.,

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp\left[\sum \beta_i X_i + \gamma_1 a_0\right] \text{ or}$$

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp\left[\sum \beta_i X_i + \gamma_1 a_0 + \gamma_2 a_0^2\right]$$

   ii. Age-as-time-scale
      a. outcome is a = age at event or censorship
      b. adjusting for age truncation, e.g.,

$$h(a, \mathbf{X}) = h_0(a \mid a_0) \exp\left[\sum \beta_i X_i\right] \text{ or}$$

$$h_g(a, \mathbf{X}) = h_{0g}(a \mid a_0) \exp\left[\sum \beta_i X_i\right]$$

F. Example from Pencina et al (2007)
   i. age-as-time-scale model
      a. need to adjust for age truncation subjects
      b. incorrect results if subjects assumed to be observed from birth
   ii. time-on-study model: when a_0 is controlled, results similar to age-as-time-scale model.
   iii. overall recommendation: correct adjustment for the age at entry is crucial

XI. Summary (pages 143–144)

Practice Exercises

1. In a 10-year follow-up study conducted in Evans County, Georgia, involving persons 60 years or older, one research question concerned evaluating the relationship of social support to mortality status. A Cox proportional hazards model was fit to describe the relationship of a measure of social network to time until death. The social network index was denoted as SNI, and took on integer values between 0 (poor social network) to 5 (excellent social network). Variables to be considered for control in the analysis as either potential confounders or potential effect modifiers were AGE (treated continuously), RACE (0,1), and SEX (0,1).

   a. State an initial PH model that can be used to assess the relationship of interest, which considers the potential confounding and interaction effects of the AGE, RACE, and SEX (assume no higher than two-factor products involving SNI with AGE, RACE, and SEX).

   b. For your model in part 1a, give an expression for the hazard ratio that compares a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates being controlled.

   c. Describe how you would test for interaction using your model in part 1a. In particular, state the null hypothesis, the general form of your test statistic, with its distribution and degrees of freedom under the null hypothesis.

   d. Assuming a revised model containing no interaction terms, give an expression for a 95% interval estimate for the adjusted hazard ratio comparing a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates in your model.

   e. For the no-interaction model described in part 1d, give an expression (i.e., formula) for the estimated survival curve for a person with SNI = 4, adjusted for AGE, RACE, and SEX, where the adjustment uses the overall mean value for each of the three covariates.

   f. Using the no-interaction model described in part 1d, if the estimated survival curves for persons with SNI = 4 and SNI = 2 adjusted for (mean) AGE, RACE, and SEX are plotted over time, will these two estimated survival curves cross? Explain briefly.

   g. For the (interaction) model described in Part 1a, what is the formula for the 95% CI for the HR that compares a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates being controlled?

2. For this question, we consider the survival data for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their book (*The Statistical Analysis of Survival Time Data, Wiley*, 1980). The variables in this dataset are listed as follows:

	Variable#	Variable name	Coding
	1	Treatment	Standard = 1, test = 2
Four indicator variables for cell type	2	Cell type 1	Large = 1, other = 0
	3	Cell type 2	Adeno = 1, other = 0
	4	Cell type 3	Small = 1, other = 0
	5	Cell type 4	Squamous = 1, other = 0
	6	Survival time	(Days) integer counts
	7	Performance status	0 = worst, . . . , 100 = best
	8	Disease duration	(Months) integer counts
	9	Age	(Years) integer counts
	10	Prior therapy	None = 0, some = 10
	11	Status	0 = censored, 1 = died

For these data, a Cox PH model was fitted yielding the following edited computer results:
Response: survival time

Variable name	Coef.	Std. Err.	$p > \|z\|$	Haz. Ratio	[95% Conf.	interval]
1 Treatment	0.290	0.207	0.162	1.336	0.890	2.006
3 Adeno cell	0.789	0.303	0.009	2.200	1.216	3.982
4 Small cell	0.457	0.266	0.086	1.579	0.937	2.661
5 Squamous cell	−0.400	0.283	0.157	0.671	0.385	1.167
7 Perf. status	−0.033	0.006	0.000	0.968	0.958	0.978
8 Disease dur.	0.000	0.009	0.992	1.000	0.982	1.018
9 Age	−0.009	0.009	0.358	0.991	0.974	1.010
10 Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054

Log likelihood = −475.180

a. State the Cox PH model used to obtain the above computer results.

b. Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with large cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.

c. Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with squamous cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.

d. Based on the computer results, is there an effect of treatment on survival time? Explain briefly.

e. Give an expression for the estimated survival curve for a person who was given the test treatment and who had a squamous cell type, where the variables to be adjusted are performance status, disease duration, age, and prior therapy.

f. Suppose a revised Cox model is used which contains, in addition to the variables already included, the product terms: treatment × performance status; treatment × disease duration; treatment × age; and treatment × prior therapy. For this revised model, give an expression for the hazard ratio for the effect of treatment, adjusted for the other variables in the model.

3. The data for this question contain survival times of 65 multiple myeloma patients (references Krall et al., "A Step-up Procedure for Selecting Variables Associated with Survival Data," *Biometrics*, vol. 31, pp. 49–57, 1975). A partial list of the variables in the dataset is given below:

   Variable 1: observation number
   Variable 2: survival time (in months) from time of diagnosis
   Variable 3: survival status (0 = alive, 1 = dead)
   Variable 4: platelets at diagnosis (0 = abnormal, 1 = normal)
   Variable 5: age at diagnosis (years)
   Variable 6: sex (1 = male, 2 = female)

Below, we provide edited computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results.

Model 1:

Variable	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Platelets	0.470	2.854	.869	1.600	0.006	429.689
Age	0.000	0.037	.998	1.000	0.930	1.075
Sex	0.183	0.725	.801	1.200	0.290	4.969
Platelets × age	−0.008	0.041	.850	0.992	0.915	1.075
Platelets × sex	−0.503	0.804	.532	0.605	0.125	2.924
	Log likelihood = −153.040					

Model 2:

Platelets	−0.725	0.401	.071	0.484	0.221	1.063
Age	−0.005	0.016	.740	0.995	0.965	1.026
Sex	−0.221	0.311	.478	0.802	0.436	1.476
	Log likelihood = −153.253					

Model 3:

Platelets	−0.706	0.401	.078	0.493	0.225	1.083
Age	−0.003	0.015	.828	0.997	0.967	1.027
	Log likelihood = −153.509					

Model 4:

Platelets	−0.705	0.397	.076	0.494	0.227	1.075
Sex	−0.204	0.307	.506	0.815	0.447	1.489
	Log likelihood = −153.308					

Model 5:

Platelets	−0.694	0.397	.080	0.500	0.230	1.088
	Log likelihood = −153.533					

a. For model 1, give an expression for the hazard ratio for the effect of the platelet variable adjusted for age and sex.

b. Using your answer to part 3a, compute the estimated hazard ratio for a 40-year-old male. Also compute the estimated hazard ratio for a 50-year-old female.

c. Carry out an appropriate test of hypothesis to evaluate whether there is any significant interaction in model 1. What is your conclusion?

d. Considering models 2–5, evaluate whether age and sex need to be controlled as confounders?

e. Which of the five models do you think is the best model and why?

f. Based on your answer to part 3e, summarize the results that describe the effect of the platelet variable on survival adjusted for age and sex.

g. Why might you consider using age-as-the-time-scale instead of time-on-follow-up as the outcome to analyze these data?

Test

1. Consider a hypothetical 2-year study to investigate the effect of a passive smoking intervention program on the incidence of upper respiratory infection (URI) in newborn infants. The study design involves the random allocation of one of three intervention packages (A, B, C) to all healthy newborn infants in Orange County, North Carolina, during 1985. These infants are followed for 2 years to determine whether or not URI develops. The variables of interest for using a survival analysis on these data are:

   T = time (in weeks) until URI is detected or time until censored
   s = censorship status (= 1 if URI is detected, = 0 if censored)
   PS = passive smoking index of family during the week of birth of the infant
   DC = daycare status (= 1 if outside daycare, = 0 if only daycare is in home)
   BF = breastfeeding status (= 1 if infant is breastfed, = 0 if infant is not breastfed)
   T_1 = first dummy variable for intervention status (= 1 if A, = 0 if B, = −1 if C)
   T_2 = second dummy variable for intervention status (= 1 if B, = 0 if A, = −1 if C).

   a. State the Cox PH model that would describe the relationship between intervention package and survival time, controlling for *PS*, *DC*, and *BF* as confounders and effect modifiers. In defining your model, use only two factor product terms involving exposure (i.e., intervention) variables multiplied by control variables in your model.

   b. Assuming that the Cox PH model is appropriate, give a formula for the hazard ratio that compares a person in intervention group A with a person in intervention group C, adjusting for *PS*, *DC*, and *BF*, and assuming interaction effects.

   c. Assuming that the PH model in part 1a is appropriate, describe how you would carry out a chunk test for interaction; i.e., state the null hypothesis, describe the test statistic and give the distribution of the test statistic and its degrees of freedom under the null hypothesis.

d. Assuming no interaction effects, how would you test whether packages A, B, and C are equally effective, after controlling for *PS*, *DC*, and *BF* in a Cox PH model without interaction terms (i.e., state the two models being compared, the null hypothesis, the test statistic, and the distribution of the test statistic under the null hypothesis).

e. For the no-interaction model considered in parts 1c and 1d, give an expression for the estimated survival curves for the effect of intervention A adjusted for *PS*, *DC*, and *BF*. Also, give similar (but different) expressions for the adjusted survival curves for interventions B and C.

2. The data for this question consists of a sample of 50 persons from the 1967–1980 Evans County Study. There are two basic independent variables of interest: AGE and chronic disease status (CHR), where CHR is coded as 0 = none, 1 = chronic disease. A product term of the form AGE × CHR is also considered. The dependent variable is time until death, and the event is death. The primary question of interest concerns whether CHR, considered as the exposure variable, is related to survival time, controlling for AGE. The edited output of computer results for this question is given as follows:

Model 1:

Variable	Coef.	Std. Err.	Chi-sq	$p > \|z\|$
CHR	0.8595	0.3116	7.61	.0058
	Log likelihood = −142.87			
Model 2:				
CHR	0.8051	0.3252	6.13	.0133
AGE	0.0856	0.0193	19.63	.0000
	Log likelihood = −132.45			
Model 3:				
CHR	1.0009	2.2556	0.20	.6572
AGE	0.0874	0.0276	10.01	.0016
CHR × AGE	−0.0030	0.0345	0.01	.9301
	Log likelihood = −132.35			

a. State the Cox PH model that allows for main effects of CHR and AGE as well as the interaction effect of CHR with AGE.

b. Carry out the test for significant interaction; i.e., state the null hypothesis, the test statistic, and its distribution under the null hypothesis. What are your conclusions about interaction?

c. Assuming no interaction, should AGE be controlled? Explain your answer on the basis of confounding and/or precision considerations.

d. If, when considering plots of various hazard functions over time, the hazard function for persons with CHR = 1 crosses the hazard function for persons with CHR = 0, what does this indicate about the use of any of the three models provided in the printout?

e. Using model 2, give an expression for the estimated survival curve for persons with CHR = 1, adjusted for AGE. Also, give an expression for the estimated survival curve for persons with CHR = 0, adjusted for AGE.

f. What is your overall conclusion about the effect of CHR on survival time based on the computer results provided from this study?

3. The data for this question contain remission times of 42 multiple leukemia patients in a clinical trial of a new treatment. The variables in the dataset are given below:

Variable 1: survival time (in weeks)
Variable 2: status (1 = in remission, 0 = relapse)
Variable 3: sex (1 = female, 0 = male)
Variable 4: log WBC
Variable 5: *Rx* status (1 = placebo, 0 = treatment)

Below, we provide computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results starting below.

Model 1:

Variable	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Rx	0.894	1.815	.622	2.446	0.070	85.812
Sex	−1.012	0.752	.178	0.363	0.083	1.585
log WBC	1.693	0.441	.000	5.437	2.292	12.897
Rx × Sex	1.952	0.907	.031	7.046	1.191	41.702
Rx × log WBC	−0.151	0.531	.776	0.860	0.304	2.433
	Log likelihood = −69.515					

Model 2:

Rx	0.405	0.561	.470	1.500	0.499	4.507
Sex	−1.070	0.725	.140	0.343	0.083	1.422
log WBC	1.610	0.332	.000	5.004	2.610	9.592
Rx × Sex	2.013	0.883	.023	7.483	1.325	42.261
	Log likelihood = −69.555					

Model 3:

Rx	0.587	0.542	.279	1.798	0.621	5.202
Sex	−1.073	0.701	.126	0.342	0.087	1.353
Rx × Sex	1.906	0.815	.019	6.726	1.362	33.213
	Log likelihood = −83.475					

Model 4:

Rx	1.391	0.457	.002	4.018	1.642	9.834
Sex	0.263	0.449	.558	1.301	0.539	3.139
log WBC	1.594	0.330	.000	4.922	2.578	9.397
	Log likelihood = −72.109					

a. Use the above computer results to carry out a chunk test to evaluate whether the two interaction terms in model 1 are significant. What are your conclusions?

b. Evaluate whether you would prefer model 1 or model 2. Explain your answer.

c. Using model 2, give an expression for the hazard ratio for the effect of the *Rx* variable adjusted for SEX and log WBC.

d. Using your answer in part 3c, compute the hazard ratio for the effect of *Rx* for males and for females separately.

e. By considering the potential confounding of log WBC, determine which of models 2 and 3 you prefer. Explain.

f. Of the models provided which model do you consider to be best? Explain.

Answers to Practice Exercises

1. a. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1\,\text{SNI} + \beta_2\,\text{AGE} + \beta_3\,\text{RACE} + \beta_4\,\text{SEX} + \beta_5\,\text{SNI} \times \text{AGE} + \beta_6\,\text{SNI} \times \text{RACE} + \beta_7\,\text{SNI} \times \text{SEX}]$

   b. $HR = \exp[2\beta_1 + 2(\text{AGE})\beta_5 + 2(\text{RACE})\beta_6 + 2(\text{SEX})\beta_7]$

   c. H_0: $\beta_5 = \beta_6 = \beta_7 = 0$. Likelihood ratio test statistic: $-2\ln L_R - (-2\ln L_F)$, which is approximately X^2_3 under H_0, where R denotes the reduced model (containing no product terms) under H_0, and F denotes the full model (given in part la above).

   d. 95% CI for adjusted HR:

   $$\exp\left[2\hat{\beta}_1 \pm 1.96 \times 2\sqrt{\hat{\text{Var}}\left(\hat{\beta}_1\right)}\right]$$

   e. $\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[4\hat{\beta}_1 + (\overline{\text{AGE}})\hat{\beta}_2 + (\overline{\text{RACE}})\hat{\beta}_3 + (\overline{\text{SEX}})\hat{\beta}_4\right]}$

   f. The two survival curves will **not** cross, because both are computed using the same proportional hazards model, which has the property that the hazard functions, as well as their corresponding estimated survivor functions, will not cross.

   g. 95% CI for HR = $\exp[\ell]$: $\exp[\hat{\ell} \pm 1.96\sqrt{\hat{\text{Var}}\,\hat{\ell}}]$ where $\ell = 2\beta_1 + 2(\text{AGE})\beta_5 + 2(\text{RACE})\beta_6 + 2(\text{SEX})\beta_7$

2. a. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \ldots + \beta_{10} X_{10}]$

   b. Adeno cell type: $\mathbf{X}^*$ = (treatment, 1, 0, 0, perfstat, disdur, age, prther)
   Large cell type: $\mathbf{X}$ = (treatment, 0, 0, 0, perfstat, disdur, age, prther)

   $$\begin{aligned} HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} &= \exp\left[\sum_{i=1}^{p}\beta_i\left(X_i^* - X_i\right)\right] \\ &= \exp\,[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) \\ &\quad + \hat{\beta}_5(0-0) + 0 + \ldots + 0] \\ &= .\exp\left[\hat{\beta}_3\right] = \exp[0.789] = 2.20 \end{aligned}$$

   c. Adeno cell type: $\mathbf{X}^*$ = (treatment, 1, 0, 0, perfstat, disdur, age, prther)
   Squamous cell type: $\mathbf{X}$ = (treatment, 0, 0, 1, perfstat, disdur, age, prther)

   $$\begin{aligned} HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} &= \exp\left[\sum_{i=1}^{p}\beta_i\left(X_i^* - X_i\right)\right] \\ &= \exp\,[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) \\ &\quad + \hat{\beta}_5(0-1) + 0 + \ldots + 0] \\ &= \exp\left[\hat{\beta}_3 - \hat{\beta}_5\right] = \exp\,[0.789 \\ &\quad - (-0.400)] = \exp\,[1.189] = 3.28 \end{aligned}$$

d. There does not appear to be an effect of treatment on survival time, adjusted for the other variables in the model. The hazard ratio is 1.3, which is close to the null value of one, the p-value of 0.162 for the Wald test for treatment is not significant, and the 95% confidence interval for the treatment effect correspondingly includes the null value.

e. $\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp[2\hat{\beta}_1+\hat{\beta}_5+\overline{(\text{perfstat})}\hat{\beta}_7+\overline{(\text{disdur})}\hat{\beta}_8+\overline{(\text{age})}\hat{\beta}_9+\overline{(\text{prther})}\hat{\beta}_{10}]}$

f. $$HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp[\beta_1 + (\text{perfstat})\beta_{11} + (\text{disdur})\beta_{12} + (\text{age})\beta_{13} + (\text{prther})\beta_{14}]$$

where β_1 is the coefficient of the treatment variable and β_{11}, β_{12}, β_{13}, and β_{14} are the coefficients of product terms involving treatment with the four variables indicated.

3. a. $\widehat{HR} = \exp[0.470 + (-0.008)\text{age} + (-0.503)\text{sex}]$

b. 40-year-old male:

$$\widehat{HR} = \exp[0.470 + (-0.008)40 + (-0.503)1] = 0.70$$

50-year-old Female:

$$\widehat{HR} = \exp[0.470 + (-0.008)50 + (-0.503)2] = 0.39$$

c. The LR (chunk) test for the significance of both interaction terms simultaneously yields the following likelihood ratio statistic which compares models 1 and 2:

$$LR = [(-2 \times -153.253) - (-2 \times -153.040)]$$
$$= 306.506 - 306.080 = 0.426$$

This statistic is approximately chi-square with 2 degrees of freedom under the null hypothesis of no interaction. This LR statistic is highly nonsignificant. Thus, we conclude that there is no significant interaction in the model (1).

d. The gold-standard hazard ratio is 0.484, which is obtained for model 2. Note that model 2 contains no interaction terms and controls for both covariates of interest. When either age or sex or both are dropped from the model, the hazard ratio (for platelets) does not change appreciably. Therefore, it appears that neither age nor sex need to be controlled for confounding.

e. Models 2–5 are all more or less equivalent, since they all give essentially the same hazards ratio and confidence interval for the effect of the platelet variable. A political choice for best model would be the gold-standard model (2), because the critical reviewer can see both age and sex being controlled in model 2.

f.
- The point estimate of the hazard ratio for normal versus abnormal platelet count is $0.484 = 1/2.07$, so that the hazard for an abnormal count is twice that for a normal count.
- There is a borderline significant effect of platelet count on survival adjusted for age and sex ($P = .071$).
- The 95% CI for the hazard ratio is given by $0.221 < HR < 1.063$, which is quite wide and therefore shows a very imprecise estimate.

4. Subjects may already at risk for the outcome prior to their study entry (at diagnosis). If so, then the time at risk prior to study entry contributes to the true survival time (say, T) for the individual, although only the observed time-on-study (t), is actually available to be analyzed. The individual's survival time is therefore underestimated by the time-on-study information (obtained from study entry), i.e., the true survival time is **left-truncated**. However, if age-as-the-time-scale as the outcome is considered, then it is possible to adjust for this left truncation by using age at entry in a hazard model of the form

$$h(a, \mathbf{X}) = h_0(a|a_0) \exp\left[\sum \beta_i X_i\right]$$

where a denotes age at follow-up, a_0 denotes age at study entry, and $h_0(a|a_0)$ is a baseline hazard that adjusts for age truncation at a_0.