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 86–94)

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 94–96)

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. **X** denotes a collection of p explanatory variables $X_1, X_2, \ldots, X_p$.

D. The model is **nonparametric** 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 96–98)

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 98–100)

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 100–104)

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

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

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

C. Typical coding identifies **X*** as unexposed group and **X** as exposed group, i.e., $X_1^* = 1$ for unexposed group and $X_1 = 0$ for exposed group; such coding allows **X*** to indicate the group with the larger hazard.

VI. Adjusted survival curves using the Cox PH model (pages 104–108)

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

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

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.

VII. The meaning of the PH assumption (pages 108–111)

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 **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

VIII. Summary (page 112)

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.

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 computer results:

Response: survival time

Variable name	Coeff	S.E.	p-value	*HR*	95%	CI	*P(PH)*
1 Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
3 Adeno cell	0.789	0.303	0.009	2.200	1.216	3.982	0.083
4 Small cell	0.457	0.266	0.086	1.579	0.937	2.661	0.080
5 Squamous cell	–0.400	0.283	0.157	0.671	0.385	1.167	0.089
7 Perf. status	–0.033	0.006	0.000	0.968	0.958	0.978	0.000
8 Disease dur.	0.000	0.009	0.992	1.000	0.982	1.018	0.919
9 Age	–0.009	0.009	0.358	0.991	0.974	1.010	0.199
10 Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.147

$-2 \ln L$: 950.359

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. Is there any suggestion from the printout that the PH assumption may not be appropriate for some of the variables in the model? Explain.

g. 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: *SPIDA manual,* Sydney, Australia, 1991; and 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 computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results starting on the next page.

Model 1:

Variable	Coeff	S.E.	p-value	*HR*	0.95	CI	*P(PH)*
Platelets	0.470	2.854	.869	1.600	0.006	429.689	0.729
Age	0.000	0.037	.998	1.000	0.930	1.075	0.417
Sex	0.183	0.725	.801	1.200	0.290	4.969	0.405
Platelets × age	–0.008	0.041	.850	0.992	0.915	1.075	0.561
Platelets × sex	–0.503	0.804	.532	0.605	0.125	2.924	0.952

–2 ln *L*: 306.080

Model 2:

Platelets	–0.725	0.401	.071	0.484	0.221	1.063	0.863
Age	–0.005	0.016	.740	0.995	0.965	1.026	0.405
Sex	–0.221	0.311	.478	0.802	0.436	1.476	0.487

–2 ln *L*: 306.505

Model 3:

Platelets	–0.706	0.401	.078	0.493	0.225	1.083	0.792
Age	–0.003	0.015	.828	0.997	0.967	1.027	0.450

–2 ln *L*: 307.018

Model 4:

Platelets	–0.705	0.397	.076	0.494	0.227	1.075	0.860
Sex	–0.204	0.307	.506	0.815	0.447	1.489	0.473

–2 ln *L*: 306.616

Model 5:

Platelets	–0.694	0.397	.080	0.500	0.230	1.088	0.793

–2 ln *L*: 307.065

a. Considering any of the above computer results, do you find any evidence that the proportional hazards assumption is not satisfied for any variable being considered?

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

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

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

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

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

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

Test

1. Consider a hypothetical two-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 two 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 output of computer results for this question is given as follows:

Model 1:

Variable	Coeff	S.E.	Chi-sq	p-value
CHR	0.8595	0.3116	7.61	.0058

$-2 \ln L = 285.74$

Model 2:

CHR	0.8051	0.3252	6.13	.0133
AGE	0.0856	0.0193	19.63	.0000

$-2 \ln L = 264.90$

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

$-2 \ln L = 264.89$