Chapter 4

Modeling of Survival Data

Now we will explore the relationship between survival and explanatory variables by modeling. In this class, we consider two broad classes of regression models:

- Proportional Hazards (PH) models
- Accelerated Failure Time (AFT) models

We will also distinguish between semi-parametric models and parametric models.

- **Proportional Hazards (PH) models**

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t)\boldsymbol{\psi}(\boldsymbol{Z})$$

Most commonly, we write the second term as:

$$\boldsymbol{\psi}(\boldsymbol{Z}) = e^{\boldsymbol{\beta Z}}$$

Suppose $Z = 1$ for treated subjects and $Z = 0$ for untreated subjects. Then this model says that the hazard is increased by a factor of e^{β} for treated subjects versus untreated subjects (e^{β} might be < 1).

This is an example of a **semi-parametric** model.

- **Accelerated Failure Time (AFT) models**

$$\log(T) = \mu + \boldsymbol{\beta Z} + \sigma w$$

where w is an "error distribution". Typically, we place a **parametric** assumption on w:

– exponential, Weibull, Gamma

– lognormal

Covariates:

In general, $\boldsymbol{Z}$ is a *vector* of covariates of interest.

$\boldsymbol{Z}$ may include:

- continuous factors (eg, age, blood pressure),
- discrete factors (gender, marital status),
- possible interactions (age by sex interaction)

Discrete Covariates:

Just as in standard linear regression, if we have a discrete covariate A with a levels, then we will need to include $(a-1)$ dummy variables $(U_1, U_2, \ldots, U_a)$ such that $U_j = 1$ if $A = j$. Then

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_2 U_2 + \beta_3 U_3 + \cdots + \beta_a U_a)$$

(In the above model, the subgroup with $A = 1$ or $U_1 = 1$ is the reference group.)

Interactions:

Two factors, A and B, interact if the hazard for A depends on the level of B.

We usually follow the principle of hierarchical models, and only include interactions if all of the corresponding main effects are also included.

The example I just gave was based on a proportional hazards model, but the description of the types of covariates we might want to include in our model applies to both the AFT and PH model.

We'll start out by focusing on the Cox PH model, and address some of the following questions:

- What does the term $\lambda_0(t)$ mean?
- What's "proportional" about the PH model?
- How do we estimate the parameters in the model?
- How do we interpret the estimated values?
- How can we construct tests of whether the covariates have a significant effect on the distribution of survival times?
- How do these tests compare to the logrank test or the Wilcoxon test?

4.1 The Cox Proportional Hazards model

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta Z})$$

This is the most common model used for survival data. Why?

- flexible choice of covariates
- fairly easy to fit
- standard software exists

References: Collett, Chapters 3 and 4
Hosmer, Lemeshow & May, Chapters 3-7
Allison, Chapter 5
Cox and Oakes, Chapter 7
Kleinbaum, Chapter 3
Klein and Moeschberger, Chapters 8 & 9
Kalbfleisch and Prentice
Lee

Note: some books (like HL& M and Collett) use $h(t; \boldsymbol{X})$ and $H(t)$ for the hazard and cumulative hazard, respectively, instead of $\lambda(t; \boldsymbol{Z})$ and $\Lambda(t)$.

Why do we call it proportional hazards?

Think of the first example, where $Z = 1$ for treated and $Z = 0$ for control. Then if we think of $\lambda_1(t)$ as the hazard rate for the treated group, and $\lambda_0(t)$ as the hazard for control, then we can write:

$$\begin{aligned} \lambda_1(t) &= \lambda(t; Z = 1) = \lambda_0(t) \exp(\beta Z) \\ &= \lambda_0(t) \exp(\beta) \end{aligned}$$

This implies that the ratio of the two hazards is a constant, ϕ, which does NOT depend on time, t. In other words, the hazards of the two groups remain proportional over time.

$$\phi = \frac{\lambda_1(t)}{\lambda_0(t)} = e^{\beta}$$

ϕ is referred to as the **hazard ratio**.

What is the interpretation of β here?

4.1.1 The Baseline Hazard Function

In the example of comparing two treatment groups, $\lambda_0(t)$ is the hazard rate for the control group.

In general, $\lambda_0(t)$ is called the **baseline hazard function**, and reflects the underlying hazard for subjects with all covariates $Z_1, ..., Z_p$ equal to 0 (i.e., the "reference group").

The general form is:

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p)$$

So when we substitute all of the Z_j's equal to 0, we get:

$$\begin{aligned} \lambda(t, \boldsymbol{Z} = 0) &= \lambda_0(t) \exp(\beta_1 * 0 + \beta_2 * 0 + \cdots + \beta_p * 0) \\ &= \lambda_0(t) \end{aligned}$$

In the general case, we think of the i-th individual having a set of covariates $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{pi})$, and we model their hazard rate as some multiple of the baseline hazard rate:

$$\lambda_i(t, \boldsymbol{Z}_i) = \lambda_0(t) \exp(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi})$$

4.1.2 Interpretation of parameters

This means we can write the log of the hazard ratio for the i-th individual to the reference group as:

$$\log\left(\frac{\lambda_i(t)}{\lambda_0(t)}\right) = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi}$$

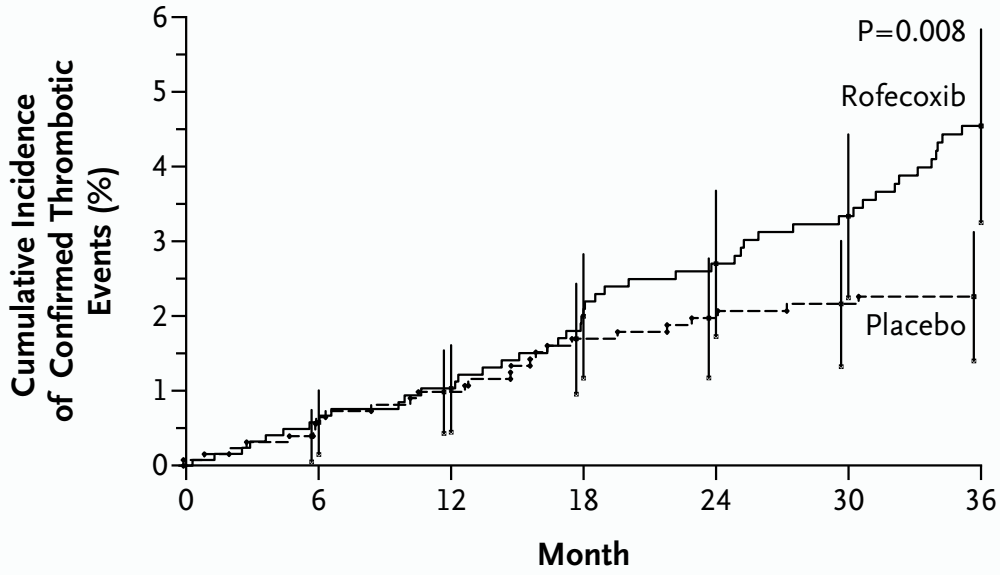
The Cox Proportional Hazards model is a linear model for the log of the hazard ratio

One of the biggest advantages of the framework of the Cox PH model is that we can estimate the parameters $\boldsymbol{\beta}$ which reflect the effects of treatment and other covariates without having to make any assumptions about the form of $\lambda_0(t)$.

In other words, we don't have to assume that $\lambda_0(t)$ follows an exponential model, or a Weibull model, or any other particular parametric model.

That's what makes the model *semi-parametric*.

Cardiovascular Events Associated with Rofecoxib (NEJM 2005)



No. at Risk							
Rofecoxib	1287	1129	1057	989	938	896	727
Placebo	1299	1195	1156	1079	1042	1001	835

Figure 2. Kaplan–Meier Estimates of the Cumulative Incidence of Confirmed Serious Thrombotic Events.

Vertical lines indicate 95 percent confidence intervals.

Hazard Ratios for Confirmed Cardiovascular Events

Table 2. Incidence of Adjudicated Thrombotic Adverse Events.*

Adverse Event	Rofecoxib Group (N=1287)		Placebo Group (N=1299)		Hazard Ratio (95% CI)
	No. of Patients (%)	Rate/100 Patient-yr	No. of Patients (%)	Rate/100 Patient-yr	
Total	46 (3.6)	1.50	26 (2.0)	0.78	1.92 (1.19–3.11)
Cardiac events	31 (2.4)	1.01	12 (0.9)	0.36	2.80 (1.44–5.45)
Myocardial infarction	21		9		
Fatal myocardial infarction	2		3		
Sudden death from cardiac causes	3		1		
Unstable angina pectoris	7		4		
Cerebrovascular events	15 (1.2)	0.49	7 (0.5)	0.21	2.32 (0.89–6.74)
Fatal ischemic stroke	1		0		
Ischemic stroke	11		6		
Transient ischemic attack	5		2		
Peripheral vascular events	3 (0.2)	0.10	7 (0.5)	0.21	0.46 (0.08–2.03)
Peripheral arterial thrombosis	1		1		
Peripheral venous thrombosis	2		4		
Pulmonary embolism	0		2		

* The total duration of follow-up was 3059 patient-years in the rofecoxib group and 3327 patient-years in the placebo group. Although a patient may have had two or more clinical adverse events, the patient was counted once within a category. The same patient may appear in different categories. CI denotes confidence interval.

So the Hazard Ratio for cardiac events was 2.80, suggesting that the risk of an event was almost three times as high for those in the Rofexocib treatment arm than in the placebo group.

Hazard Ratios for Confirmed Cardiovascular Events, continued

Questions:

1. For the confirmed cardiac events, what is the estimated β coefficient, $\hat{\beta}$?

2. Does this control for any other covariates?

3. Can you write out the Cox PH model for this analysis?

General Questions about the Cox PH Model:

1. Why don't we just model the hazard ratio, $\phi = \lambda_i(t)/\lambda_0(t)$, directly as a linear function of the covariates $\boldsymbol{Z}$?

2. Why doesn't the model have an intercept?

Example with Control for Other Covariates: NEJM, 2002

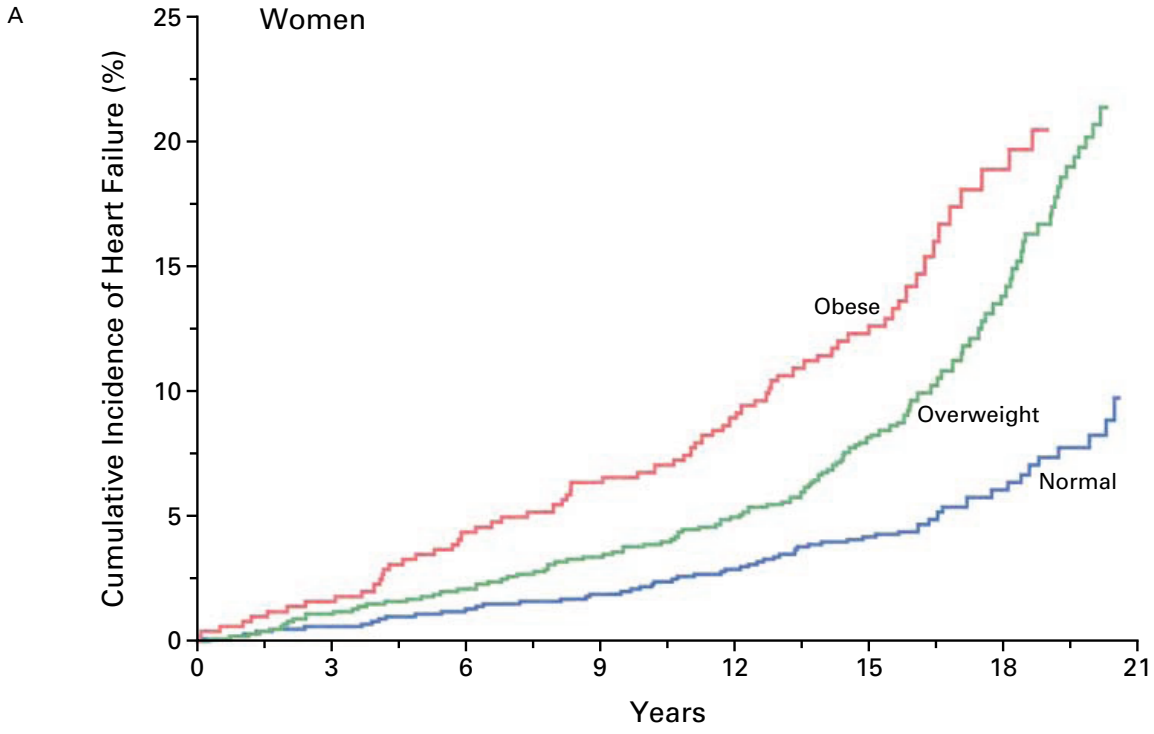
In this article entitled "Obesity and the risk of heart failure", the authors state:

The hazard ratio associated with a one unit increase in BMI is estimated at 1.07 for women and 1.05 for men, after adjusting for age, cholesterol, smoking status, gender, presence or absence of several diseases...

RESULTS: During follow-up (mean, 14 years), heart failure developed in 496 subjects (258 women and 238 men). After adjustment for established risk factors, there was an increase in the risk of heart failure of 5 percent for men and 7 percent for women for each increment of 1 in body-mass index. As compared with subjects with a normal body-mass index, obese subjects had a doubling of the risk of heart failure. For women, the hazard ratio was 2.12 (95 percent confidence interval, 1.51 to 2.97); for men, the hazard ratio was 1.90 (95 percent confidence interval, 1.30 to 2.79). A graded increase in the risk of heart failure was observed across categories of body-mass index. The hazard ratios per increase in category were 1.46 in women (95 percent confidence interval, 1.23 to 1.72) and 1.37 in men (95 percent confidence interval, 1.13 to 1.67).

Obesity and Heart Disease, NEJM 2002, Figure 1 (women)

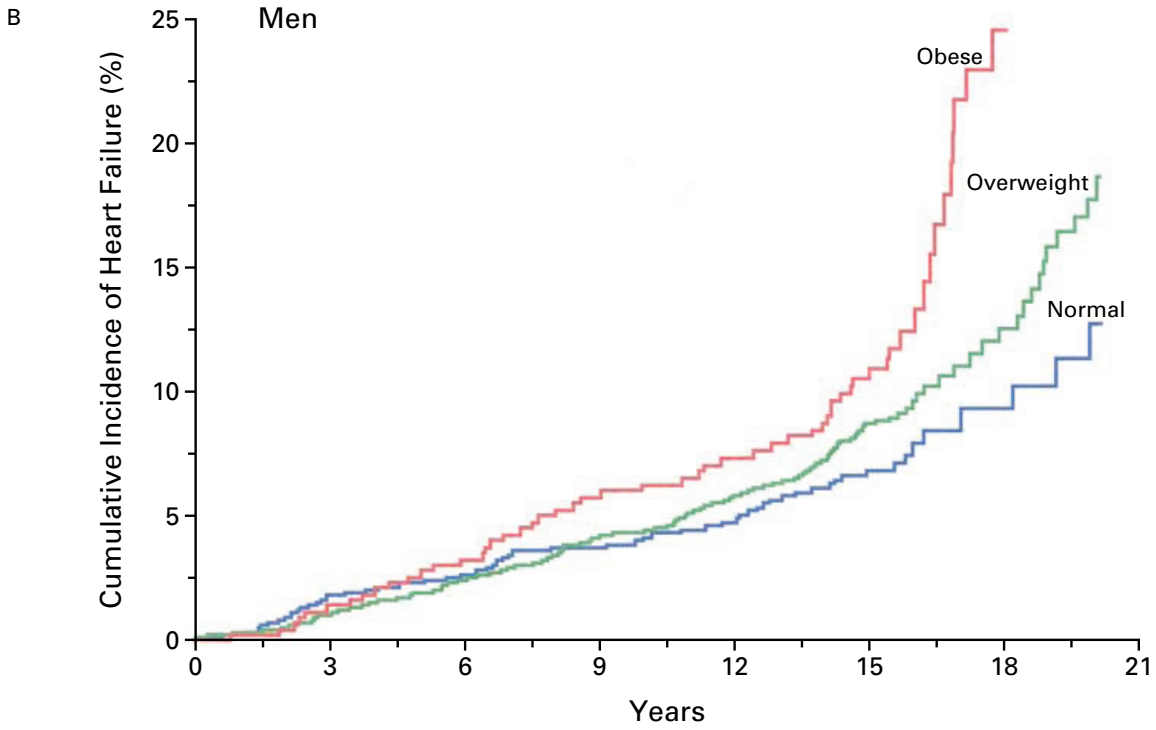
The New England Journal of Medicine



No. at Risk

Normal	1729	1688	1634	1568	1477	1227	295
Overweight	955	929	880	815	757	634	248
Obese	493	477	448	409	372	296	104

Obesity and Heart Disease, NEJM 2002, Figure 1 (men)



No. at Risk

Normal	869	822	758	690	637	512	105
Overweight	1378	1322	1254	1163	1071	871	171
Obese	457	433	403	370	342	276	51

Figure 1. Cumulative Incidence of Heart Failure According to Category of Body-Mass Index at the Base-Line Examination. The body-mass index was 18.5 to 24.9 in normal subjects, 25.0 to 29.9 in overweight subjects, and 30.0 or more in obese subjects.

Obesity and Heart Disease, NEJM 2002, portion of Table 3 (separately by sex)

TABLE 3. RESULTS OF MULTIVARIABLE COX PROPORTIONAL-HAZARDS MODELS EXAMINING THE RELATIC TO THE RISK OF HEART FAILURE.*

MODEL AND CATEGORY OF BODY-MASS INDEX	SEX-SPECIFIC ANALYSES			
	WOMEN (N=3177)		MEN (N=2704)	
	hazard ratio (95% CI)	P value	hazard ratio (95% CI)	P value
I. Models with body-mass index and all covariates defined at base line†				
A. Body-mass index as a continuous variable (per increment of 1)	1.07 (1.04–1.10)	<0.001	1.05 (1.02–1.09)	0.005
B. Body-mass index as a categorical variable				
Normal (18.5–24.9)	1.00		1.00	
Overweight (25.0–29.9)	1.50 (1.12–2.02)	0.007	1.20 (0.87–1.64)	0.27
Obese (≥30.0)	2.12 (1.51–2.97)	<0.001	1.90 (1.30–2.79)	0.001
Trend across categories	1.46 (1.23–1.72)	<0.001	1.37 (1.13–1.67)	0.002

4.2 Estimation of model parameters

The basic idea is that under PH, information about β can be obtained from the relative orderings (i.e., ranks) of the survival times, rather than the actual values. Why?

Suppose T follows a PH model:

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) e^{\boldsymbol{\beta Z}}$$

Now consider $T^* = g(T)$, where g is a monotonic increasing function. The ranks of T^* will be the same as those for T. And we can show that T^* also follows the PH model, with the same multiplier, $e^{\boldsymbol{\beta Z}}$.

Therefore, when we consider likelihood methods for estimating the model parameters, we only have to worry about the ranks of the survival times.

Likelihood Estimation for the PH Model

Kalbfleisch and Prentice derived a likelihood involving only $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ (not $\lambda_0(t)$) based on the marginal distribution of the ranks of the observed failure times (in the absence of censoring).

Cox (1972) derived the same likelihood, and generalized it for censoring, using the idea of a **partial likelihood**

Suppose we observe $(X_i, \delta_i, \boldsymbol{Z}_i)$ for individual i, where

- X_i is a censored failure time random variable
- δ_i is the failure/censoring indicator (1=fail,0=censor)
- $\boldsymbol{Z}_i$ represents a set of covariates

The covariates may be continuous, discrete, or time-varying.

Suppose there are K distinct failure (or death) times, and let $\tau_1, ...\tau_K$ represent the K ordered, distinct death times.

For now, assume there are no tied death times.

Let $\mathcal{R}(t) = \{i : x_i \geq t\}$ denote the set of individuals who are "at risk" for failure at time t.

More about risk sets:

- I will refer to $\mathcal{R}(\tau_j)$ as the risk set at the jth failure time
- I will refer to $\mathcal{R}(X_i)$ as the risk set at the failure time of individual i
- There will still be r_j individuals in $\mathcal{R}(\tau_j)$.
- r_j is a number, while $\mathcal{R}(\tau_j)$ identifies the actual subjects at risk

What is the partial likelihood?

Intuitively, it is a product over the set of observed death times of the conditional probabilities of seeing the observed deaths, given the set of individuals at risk at those times.

At each death time τ_j, the contribution to the likelihood is:

$$\begin{aligned} L_j(\boldsymbol{\beta}) &= Pr(\text{individual j fails}|\text{1 failure from } \mathcal{R}(\tau_j)) \\ &= \frac{Pr(\text{individual } j \text{ fails}|\text{ at risk at } \tau_j)}{\sum_{\ell \in \mathcal{R}(\tau_j)} Pr(\text{individual } \ell \text{ fails}|\text{ at risk at } \tau_j)} \\ &= \frac{\lambda(\tau_j; \mathbf{Z}_j)}{\sum_{\ell \in \mathcal{R}(\tau_j)} \lambda(\tau_j; \mathbf{Z}_\ell)} \end{aligned}$$

Under the PH assumption, $\lambda(t; \mathbf{Z}) = \lambda_0(t) e^{\boldsymbol{\beta Z}}$, so we get:

$$L^{partial}(\boldsymbol{\beta}) = \prod_{j=1}^{K} \frac{\lambda_0(\tau_j) e^{\boldsymbol{\beta Z}_j}}{\sum_{\ell \in \mathcal{R}(\tau_j)} \lambda_0(\tau_j) e^{\boldsymbol{\beta Z}_\ell}} = \prod_{j=1}^{K} \frac{e^{\boldsymbol{\beta Z}_j}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\boldsymbol{\beta Z}_\ell}}$$

Another derivation:
In general, the likelihood contributions for censored data fall into two categories:

- **Individual is censored at X_i:**

$$L_i(\boldsymbol{\beta}) = S(X_i) = \exp[-\int_0^{X_i} \lambda_i(u)du]$$

- **Individual fails at X_i:**

$$L_i(\boldsymbol{\beta}) = S(X_i)\lambda_i(X_i) = \lambda_i(X_i)\exp[-\int_0^{X_i} \lambda_i(u)du]$$

Thus, everyone contributes $S(X_i)$ to the likelihood, and only those who fail contribute $\lambda_i(X_i)$.

This means we get a total likelihood of:

$$L(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} \lambda_i(X_i)^{\delta_i} \;\exp[-\int_0^{X_i} \lambda_i(u)du]$$

The above likelihood holds for all censored survival data, with general hazard function $\lambda(t)$!
(In other words, we haven't used the Cox PH assumption at all yet.)

Now, let's multiply and divide by the term $\left[\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)\right]^{\delta_i}$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)}\right]^{\delta_i} \left[\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)\right]^{\delta_i} \exp[-\int_0^{X_i} \lambda_i(u)du]$$

Cox (1972) argued that the first term in this product contained almost all of the information about $\boldsymbol{\beta}$, while the second two terms contained the information about $\lambda_0(t)$, i.e., the baseline hazard.

If we just focus on the first term, then under the Cox PH assumption:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)}\right]^{\delta_i}$$

$$= \prod_{i=1}^{n} \left[\frac{\lambda_0(X_i)\exp(\boldsymbol{\beta}\boldsymbol{Z}_i)}{\sum_{j\in\mathcal{R}(X_i)} \lambda_0(X_i)\exp(\boldsymbol{\beta}\boldsymbol{Z}_j)}\right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}\boldsymbol{Z}_i)}{\sum_{j\in\mathcal{R}(X_i)} \exp(\boldsymbol{\beta}\boldsymbol{Z}_j)}\right]^{\delta_i}$$

This is the partial likelihood defined by Cox. Note that it does not depend on the underlying hazard function $\lambda_0(\cdot)$. Cox recommended treating this as an ordinary likelihood for making inferences about β in the presence of the nuisance parameter $\lambda_0(\cdot)$.

A simple example:

individual	X_i	δ_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

Now let's compile the pieces that go into the partial likelihood contributions at each failure time:

j	ordered failure time X_i	$\mathcal{R}(X_i)$	i_j	Likelihood contribution $\left[e^{\beta Z_i}/\sum_{j\in\mathcal{R}(X_i)} e^{\beta Z_j}\right]^{\delta_i}$
1	6	{1,2,3,4}	3	$e^{7\beta}/[e^{4\beta}+e^{5\beta}+e^{7\beta}+e^{3\beta}]$
2	8	{1,2,4}	2	1
3	9	{1,4}	1	$e^{4\beta}/[e^{4\beta}+e^{3\beta}]$
4	10	{4}	4	$e^{3\beta}/e^{3\beta}=1$

The partial likelihood would be the product of these four terms.

Notes on the partial likelihood:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \left[\frac{e^{\boldsymbol{\beta} \boldsymbol{Z}_j}}{\sum_{\ell \in \mathcal{R}(X_j)} e^{\boldsymbol{\beta} \boldsymbol{Z}_\ell}} \right]^{\delta_j}$$

$$= \prod_{j=1}^{K} \frac{e^{\boldsymbol{\beta} \mathbf{z}_j}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\boldsymbol{\beta} \mathbf{z}_\ell}}$$

where the product is over the K death (or failure) times.

- contributions only at the death times
- the partial likelihood is NOT a product of independent terms, but of conditional probabilities

4.3 Inference Based on Partial Likelihood

Inference can be conducted by treating the partial likelihood as though it satisfied all the regular likelihood properties.

The **log-partial likelihood** is:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log\left[\prod_{j=1}^{n}\frac{e^{\boldsymbol{\beta}\boldsymbol{Z}_j}}{\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\boldsymbol{\beta}\boldsymbol{Z}_\ell}}\right]^{\delta_j} \\
&= \log\left[\prod_{j=1}^{K}\frac{e^{\boldsymbol{\beta}\boldsymbol{Z}_j}}{\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\boldsymbol{\beta}\boldsymbol{Z}_\ell}}\right] \\
&= \sum_{j=1}^{K}\left[\boldsymbol{\beta}\boldsymbol{Z}_j - \log[\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\boldsymbol{\beta}\boldsymbol{Z}_\ell}]\right] \\
&= \sum_{j=1}^{K} l_j(\boldsymbol{\beta})
\end{aligned}$$

where l_j is the log-partial likelihood contribution at the j-th ordered death time.

Suppose there is only one covariate ($\boldsymbol{\beta}$ is one-dimensional):

The **partial likelihood score equations** are:

$$U(\beta) = \frac{\partial}{\partial \beta}\ell(\beta) = \sum_{j=1}^{n} \delta_j \left[Z_j - \frac{\sum_{\ell \in \mathcal{R}(\tau_j)} Z_\ell e^{\beta Z_\ell}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}} \right]$$

We can express $U(\beta)$ intuitively as a sum of "observed" minus "expected" values:

$$U(\beta) = \frac{\partial}{\partial \beta}\ell(\beta) = \sum_{j=1}^{n} \delta_j (Z_j - \bar{Z}_j)$$

where $\bar{Z}_j$ is the "weighted average" of the covariate Z over all the individuals in the risk set at time τ_j. Note that β is involved through the term $\bar{Z}_j$.

The maximum partial likelihood estimator (MPLE), $\hat{\beta}$, can be found by solving $U(\beta) = 0$.

Analogous to standard likelihood theory, it can be shown (though not easily) that

$$\frac{(\widehat{\beta} - \beta)}{se(\hat{\beta})} \sim N(0, 1)$$

The variance of $\hat{\beta}$ can be obtained by inverting the second derivative of the partial likelihood,

$$var(\hat{\beta}) \sim \left[-\frac{\partial^2}{\partial \beta^2} \ \ \ell(\beta) \right]^{-1}$$

From the above expression for $U(\beta)$, we have:

$$\frac{\partial^2}{\partial \beta^2} \ell(\beta) = \sum_{j=1}^{n} \delta_j \left[-\frac{\sum_{\ell \in \mathcal{R}(\tau_j)} (Z_j - \bar{Z}_j)^2 e^{\beta Z_\ell}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}} \right]$$

<u>Note:</u> The true variance of $\hat{\beta}$ ends up being a function of β, which is unknown. We calculate the "observed" information by substituting in our partial likelihood estimate of β into the above formula for the variance

Simple Example for 2-group comparison: (no ties)

Group 0: $4^+, 7, 8^+, 9, 10^+ \implies Z_i = 0$

Group 1: $3, 5, 5^+, 6, 8^+ \implies Z_i = 1$

	ordered failure	Number at risk		Likelihood contribution
j	time X_i	Group 0	Group 1	$\left[e^{\beta Z_i} / \sum_{j \in \mathcal{R}(X_i)} e^{\beta Z_j}\right]^{\delta_i}$
1	3	5	5	$e^\beta/[5+5e^\beta]$
2	5	4	4	$e^\beta/[4+4e^\beta]$
3	6	4	2	$e^\beta/[4+2e^\beta]$
4	7	4	1	$1/[4+1e^\beta]$
5	9	2	0	$1/[2+0] = 1/2$

Again, we take the product over the likelihood contributions, then maximize to get the partial MLE for β. What does β represent in this case?

Notes:

- The "observed" information matrix is generally used because in practice, people find it has better properties. Also, the "expected" information matrix (eg., the expected value of the inverse second derivative of the log-likelihood) is very hard to calculate.

- There is a nice analogy with the score and information matrices from more standard regression problems, except that here we are summing over observed death times, rather than individuals.

- You normally cannot get a closed-form expression for the MPLE, and need to use iterative methods. The Newton-Raphson method is used by many of the computer packages to solve the partial likelihood equations.

```
install.packages("survival")
library(survival)

# Fit the Cox model
cox_model <- coxph(Surv(time, status) ~ age + sex + treatment, data = data)

# Display the summary of the model
summary(cox_model)

# Test the proportional hazards assumption
cox_zph <- cox.zph(cox_model)
print(cox_zph)

# Plot to visualize the assumption
plot(cox_zph)
```

4.4 Software for Fitting Cox Models

Fitting Cox PH model with Stata

Uses the "`stcox`" command. First, try typing "`help stcox`"

```
-------------------------------------------------------------------------------
Estimate Cox proportional hazards model
-------------------------------------
stcox [varlist] [if exp] [in range]
          [, nohr strata(varnames) robust cluster(varname) noadjust
          mgale(newvar) esr(newvars)
          schoenfeld(newvar) scaledsch(newvar)
          basehazard(newvar) basechazard(newvar) basesurv(newvar)
          {breslow | efron | exactm | exactp} cmd estimate noshow
          offset level(#) maximize-options ]

stphtest [, km log rank time(varname) plot(varname) detail
          graph-options ksm-options]

stcox is for use with survival-time data; see help st.  You must
have stset your data before using this command; see help stset.

Description
-----------
stcox estimates maximum-likelihood proportional hazards models on st data.

Options  (many more!)
-------
nohr reports the estimated coefficients rather than hazard ratios; i.e.,
    b rather than exp(b).  Standard errors and confidence intervals are
    similarly transformed.  This option affects how results are displayed,
    not how they are estimated.
```

Example: Leukemia Data

```
. stcox trt

Iteration 0:   log likelihood =  -93.98505
Iteration 1:   log likelihood = -86.385606
Iteration 2:   log likelihood = -86.379623
Iteration 3:   log likelihood = -86.379622
Refining estimates:
Iteration 0:   log likelihood = -86.379622

Cox regression -- Breslow method for ties

No. of subjects =           42                     Number of obs    =        42
No. of failures =           30
Time at risk    =          541
                                                   LR chi2(1)       =     15.21
Log likelihood  =   -86.379622                     Prob > chi2      =    0.0001
------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     trt |   .2210887   .0905501     -3.685   0.000       .0990706    .4933877
------------------------------------------------------------------------------
```

[Note: again, Stata provides the p-value as "0" when really what that means is $p < 0.001$]

Here is the same model being fit, but with the "nohr" option:

```
. stcox trt , nohr

(same iterations for log-likelihood)

Cox regression -- Breslow method for ties

No. of subjects =           42                     Number of obs    =        42
No. of failures =           30
Time at risk    =          541
                                                   LR chi2(1)       =     15.21
Log likelihood  =   -86.379622                     Prob > chi2      =    0.0001
------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     trt |  -1.509191   .4095644     -3.685   0.000      -2.311923   -.7064599
------------------------------------------------------------------------------
```

Fitting PH models in SAS - PROC PHREG

Ex. Leukemia data

```
Title 'Cox and Oakes example';
data leukemia;
    input weeks remiss trtmt;
    cards;
6     0     1
6     1     1
6     1     1
6     1     1        /* data for 6MP group */
7     1     1
9     0     1
etc
1     1     0
1     1     0        /* data for placebo group */
2     1     0
2     1     0
etc
;

proc phreg data=leukem;
  model weeks*remiss(0)=trtmt;
  title 'Cox PH Model for leukemia data';
run;
```

PROC PHREG Output:

```
The PHREG Procedure

Data Set: WORK.LEUKEM
Dependent Variable: WEEKS  Time to Relapse
Censoring Variable: REMISS
Censoring Value(s): 0
Ties Handling: BRESLOW

          Summary of the Number of
          Event and Censored Values

                                    Percent
   Total       Event    Censored    Censored

      42          30          12       28.57

                Testing Global Null Hypothesis: BETA=0

               Without        With
Criterion    Covariates    Covariates    Model Chi-Square

-2 LOG L        187.970       172.759      15.2109 with 1 DF (p<0.0001)
Score              .             .         15.9305 with 1 DF (p<0.0001)
Wald               .             .         13.5783 with 1 DF (p=0.0002)

                  Analysis of Maximum Likelihood Estimates

               Parameter   Standard    Wald       Pr >       Hazard
Variable DF     Estimate     Error  Chi-Square Chi-Square    Ratio

TRTMT     1    -1.50919    0.40956   13.5783     0.0002    0.221
```

Compare this with the logrank test from PROC LIFETEST (Using the "TEST" statement)

```
The LIFETEST Procedure

Rank Tests for the Association of WEEKS with Covariates
Pooled over Strata

Univariate Chi-Squares for the LOG RANK Test

              Test       Standard                    Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square

TRTMT         10.2505       2.5682       15.9305        0.0001
```

Notes:

- **The logrank test=score test from `PROC PHREG`!**
  In general, the score test would be for *all* of the variables in the model, but in this case, we have only "`trtmt`".
- Stata does not provide a score test in its output from the Cox model. However, the `stcox` command with the `breslow` option for ties yields the same LR test as the CMH-version logrank test from the `sts test, cox` command.

More Notes:

- The Cox Proportional hazards model has the advantage over a simple logrank test of giving us an estimate of the "hazard ratio" (i.e., $\phi = \lambda_1(t)/\lambda_0(t)$). This is more informative than just a test statistic, and we can also form confidence intervals for the hazard ratio.

- In this case, $\lambda_1(t)$ is for the treated group and $\lambda_0(t)$ is for the control group, with $\hat{\phi} = 0.221$. This can then be interpreted to mean that the hazard for relapse among patients treated with 6-MP is less than 25% of that for placebo patients.

- From the STS LIST command in Stata or PROC LIFETEST in SAS, we were able to get estimates of the entire survival distribution $\hat{S}(t)$ for each treatment group; we can't immediately get this from our Cox model without further assumptions. **Why not?**

4.5 Adjustments for ties

The proportional hazards model assumes a continuous hazard – ties are not possible. There are four proposed modifications to the likelihood to adjust for ties.

(1) **Cox's (1972) modification:** "discrete" method

(2) **Exact method (Kalbfleisch and Prentice)**

(3) **Peto-Breslow method**

(4) **Efron's (1977) method**

(5) **Exact marginal method** (stata)

Some notation:

$\tau_1, ...\tau_K$	the K ordered, distinct death times
d_j	the number of failures at τ_j
$i_{j1}, ...i_{jd_j}$	the identities of the d_j individuals who fail at τ_j

(1) Cox's (1972) modification: "discrete" method

Cox's method assumes that if there are tied failure times, they truly happened at the same time. It is based on a discrete likelihood.

The **partial likelihood** is:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{K} Pr(i_{j1}, ...i_{jd_j} \text{ fail} \mid d_j \text{ fail at } \tau_j, \text{ from } \mathcal{R})$$

$$= \prod_{j=1}^{K} \frac{Pr(i_{j1}, ...i_{jd_j} \text{ fail} \mid \text{ in } \mathcal{R}(\tau_j))}{\sum_{\ell \in s(j,d_j)} Pr(\ell_1, ...\ell_{d_j} \text{ fail} \mid \text{ in } \mathcal{R}(\tau_j))}$$

$$= \prod_{j=1}^{K} \frac{\exp(\boldsymbol{\beta Z}_{i_{j1}}) \cdots \exp(\boldsymbol{\beta Z}_{i_{jd_j}})}{\sum_{\ell \in s(j,d_j)} \exp(\boldsymbol{\beta Z}_{\ell_1}) \cdots \exp(\boldsymbol{\beta Z}_{\ell_{d_j}})} = \prod_{j=1}^{K} \frac{\exp(\boldsymbol{\beta} S_j)}{\sum_{\ell \in s(j,d_j)} \exp(\boldsymbol{\beta} S_{j\ell})}$$

where

- $s(j, d_j)$ is the set of all possible sets of d_j individuals that can possibly be drawn from the risk set at time τ_j
- S_j is the sum of the Z's for all the d_j individuals who fail at τ_j
- $S_{j\ell}$ is the sum of the Z's for all the d_j individuals in the ℓ-th set drawn out of $s(j, d_j)$

What does this all mean??!!

Let's modify our previous simple example to include ties.

Simple Example (with ties)

Group 0: $4^+, 6, 8^+, 9, 10^+ \implies Z_i = 0$
Group 1: $3, 5, 5^+, 6, 8^+ \implies Z_i = 1$

j	Ordered failure time X_i	Number at risk: Group 0	Number at risk: Group 1	**Likelihood Contribution** $e^{\beta S_j} / \sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}$
1	3	5	5	$e^\beta/[5+5e^\beta]$
2	5	4	4	$e^\beta/[4+4e^\beta]$
3	6	4	2	$e^\beta/[6+8e^\beta+e^{2\beta}]$
4	9	2	0	$e^0/2 = 1/2$

The tie occurs at $t = 6$, when $\mathcal{R}(\tau_j) = \{Z = 0 : (6, 8^+, 9, 10^+), Z = 1 : (6, 8^+)\}$. Of the $\binom{6}{2} = 15$ possible pairs of subjects at risk at t=6, there are 6 pairs formed where both are from group 0 ($S_j = 0$), 8 pairs formed with one in each group ($S_j = 1$), and 1 pairs formed with both in group 1 ($S_j = 2$). The denominator can be difficult to calculate with large numbers of ties.

(2) Exact method (Kalbfleisch and Prentice):

The "discrete" option that we discussed in (1) is an exact method based on a discrete likelihood (assuming that tied events truly ARE tied).

This second exact method is based on the continuous likelihood, under the assumption that if there are tied events, that is due to the imprecise nature of our measurement, and that there must be some *true* ordering.

All possible orderings of the tied events are calculated, and the probabilities of each are summed.

Example with 2 tied events (1,2) from riskset (1,2,3,4):

$$= \frac{e^{\beta Z_1}}{e^{\beta Z_1} + e^{\beta Z_2} + e^{\beta Z_3} + e^{\beta Z_4}} \times \frac{e^{\beta Z_2}}{e^{\beta Z_2} + e^{\beta Z_3} + e^{\beta Z_4}}$$

$$+ \frac{e^{\beta Z_2}}{e^{\beta Z_1} + e^{\beta Z_2} + e^{\beta Z_3} + e^{\beta Z_4}} \times \frac{e^{\beta Z_1}}{e^{\beta Z_1} + e^{\beta Z_3} + e^{\beta Z_4}}$$

(3) Breslow method: (default)

Breslow and Peto suggested replacing the term in the denominator by the term $\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}\right)^{d_j}$, so that the following modified partial likelihood would be used:

$$L(\beta) \approx \prod_{j=1}^{K} \frac{e^{\beta S_j}}{\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}\right)^{d_j}}$$

Justification:

Suppose individuals 1 and 2 fail from $\{1, 2, 3, 4\}$ at time τ_j. Let $\phi(i)$ be the hazard ratio for individual i (compared to baseline).

$$\frac{\phi(1)}{\phi(1)+\phi(2)+\phi(3)+\phi(4)} \times \frac{\phi(2)}{\phi(2)+\phi(3)+\phi(4)} + \frac{\phi(2)}{\phi(1)+\phi(2)+\phi(3)+\phi(4)} \times \frac{\phi(1)}{\phi(1)+\phi(3)+\phi(4)}$$

$$\approx \frac{2\phi(1)\phi(2)}{[\phi(1)+\phi(2)+\phi(3)+\phi(4)]^2}$$

The Peto (Breslow) approximation will break down when the number of ties are large relative to the size of the risk sets, and then tends to yield estimates of β which are biased toward 0.

(4) Efron's (1977) method:

Efron suggested an even closer approximation to the discrete likelihood:

$$L(\beta) = \prod_{j=1}^{K} \frac{e^{\beta S_j}}{\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell} + \frac{j-1}{d_j} \sum_{\ell \in \mathcal{D}(\tau_j)} e^{\beta Z_\ell} \right)^{d_j}}$$

Like the Breslow approximation, Efron's method will yield estimates of β which are biased toward 0 when there are many ties.

However, (1995) Allison recommends the Efron approximation since it is much faster than the exact methods and tends to yield much closer estimates than the default Breslow approach.

Bottom Line: Implications of Ties (See Allison (1995), p.127-137)

(1) **When there are no ties**, all options give *exactly* the same results.

(2) **When there are only a few ties**, it won't make much difference which method is used. However, since the exact methods won't take much extra computing time, you might as well use one of them.

(3) **When there are many ties** (relative to the number at risk), the Breslow option (default) performs poorly (Farewell & Prentice, 1980; Hsieh, 1995). Both of the approximate methods, Breslow and Efron, yield coefficients that are attenuated (biased toward 0).

(4) **The choice of which exact method to use** should be based on substantive grounds - are the tied event times truly tied? ...or are they the result of imprecise measurement?

(5) **Computing time of exact methods** is much longer than that of the approximate methods. However, in most cases it will still be less than 30 seconds even for the exact methods.

(6) **Best approximate method** - the Efron approximation nearly always works better than the Breslow method, with no increase in computing time, so use this option if exact methods are too computer-intensive.

Example: The fecundability study

Women who had recently given birth (or had tried to get pregnant for at least a year) were asked to recall how long it took them to become pregnant, and whether or not they smoked during that time. The outcome of interest is time to pregnancy (measured in menstrual cycles).

Inputting the data into SAS:

```
data fecund;
  input   smoke      cycle     status    count;
  cards;
0          1          1        198
0          2          1        107
0          3          1         55
0          4          1         38
0          5          1         18
0          6          1         22
.....................................

1         10          1          1
1         11          1          1
1         12          1          3
1         12          0          7
;
```

Cox models for evaluating effect of smoking on fecundity:

```
proc phreg;
  model cycle*status(0) = smoke /ties=breslow;   /* default */
  freq count;

proc phreg;
  model cycle*status(0) = smoke /ties=discrete;
  freq count;

proc phreg;
  model cycle*status(0) = smoke /ties=exact;
  freq count;

proc phreg;
  model cycle*status(0) = smoke /ties=efron;
  freq count;
```

SAS Output for Fecundability study: Accounting for Ties

```
***********************************************************************
Ties Handling: BRESLOW

                Parameter   Standard       Wald         Pr >      Risk
Variable    DF  Estimate     Error     Chi-Square    Chi-Square   Ratio
 SMOKE      1   -0.329054   0.11412      8.31390      0.0039     0.720

***********************************************************************
Ties Handling: DISCRETE

                Parameter   Standard       Wald         Pr >      Risk
Variable    DF  Estimate     Error     Chi-Square    Chi-Square   Ratio
 SMOKE      1   -0.461246   0.13248     12.12116      0.0005     0.630

***********************************************************************
Ties Handling: EXACT

                Parameter   Standard       Wald         Pr >      Risk
Variable    DF  Estimate     Error     Chi-Square    Chi-Square   Ratio
 SMOKE      1   -0.391548   0.11450     11.69359      0.0006     0.676

***********************************************************************
Ties Handling: EFRON

                Parameter   Standard       Wald         Pr >      Risk
Variable    DF  Estimate     Error     Chi-Square    Chi-Square   Ratio
 SMOKE      1   -0.387793   0.11402     11.56743      0.0007     0.679

***********************************************************************
```

For this particular dataset, does it seem like it would be important to consider the effect of tied failure times? Which method would be best?

Stata Commands for PH Model with Ties:

Stata also offers four options for adjustments with tied data:

- `breslow` (default)
- `efron`
- `exactp` (same as the "discrete" option in SAS)
- `exactm` - an exact marginal likelihood calculation (different than the "exact" option in SAS)

Fecundability Data Example:

```
. stcox smoker, efron nohr

        failure _d:  status
  analysis time _t:  cycle

Iteration 0:   log likelihood = -3113.5313
Iteration 1:   log likelihood = -3107.3102
Iteration 2:   log likelihood = -3107.2464
Iteration 3:   log likelihood = -3107.2464
Refining estimates:
Iteration 0:   log likelihood = -3107.2464
Cox regression -- Efron method for ties

No. of subjects =          586                     Number of obs   =       586
No. of failures =          567
Time at risk    =         1844
                                                   LR chi2(1)      =     12.57
Log likelihood  =   -3107.2464                     Prob > chi2     =    0.0004

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  smoker |  -.3877931   .1140202     -3.401   0.001      -.6112685   -.1643177
------------------------------------------------------------------------------
```

4.6 A special case: the two-sample problem

Previously, we derived the logrank test from an intuitive perspective, assuming that we have $(X_{01}, \delta_{01}) \ldots (X_{0n_0}, \delta_{0n_0})$ from group 0 and $(X_{11}, \delta_{11}), \ldots, (X_{1n_1}, \delta_{1n_1})$ from group 1.

Just as a χ^2 test for binary data can be derived from a logistic model, we will see here that the logrank test can be derived as a special case of the Cox Proportional Hazards model.

First, let's re-define our notation in terms of (X_i, δ_i, Z_i):

$$\begin{aligned}(X_{01}, \delta_{01}), \ldots, (X_{0n_0}, \delta_{0n_0}) &\implies (X_1, \delta_1, 0), \ldots, (X_{n0}, \delta_{n0}, 0) \\ (X_{11}, \delta_{11}), \ldots, (X_{1n_1}, \delta_{1n_1}) &\implies (X_{n0+1}, \delta_{n0+1}, 1), \ldots, (X_{n0+n1}, \delta_{n0+n1}, 1)\end{aligned}$$

In other words, we have $n0$ rows of data $(X_i, \delta_i, 0)$ for the group 0 subjects, then $n1$ rows of data $(X_i, \delta_i, 1)$ for the group 1 subjects.

Using the proportional hazards formulation, we have

$$\lambda(t; Z) = \lambda_0(t)\, e^{\beta Z}$$

Group 0 hazard: $\lambda_0(t)$

Group 1 hazard: $\lambda_0(t)\, e^{\beta}$

The log-partial likelihood is:

$$\begin{aligned} logL(\beta) &= \log\left[\prod_{j=1}^{K}\frac{e^{\beta Z_j}}{\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\beta Z_\ell}}\right] \\ &= \sum_{j=1}^{K}\left[\beta Z_j - \log[\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\beta Z_\ell}]\right] \end{aligned}$$

Taking the derivative with respect to β, we get:

$$\begin{aligned} U(\beta) &= \frac{\partial}{\partial\beta}\ell(\beta) \\ &= \sum_{j=1}^{n}\delta_j\left[Z_j - \frac{\sum_{\ell\in\mathcal{R}(\tau_j)}Z_\ell e^{\beta Z_\ell}}{\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\beta Z_\ell}}\right] \\ &= \sum_{j=1}^{n}\delta_j(Z_j - \bar{Z}_j) \qquad \text{where } \bar{Z}_j = \frac{\sum_{\ell\in\mathcal{R}(\tau_j)}Z_\ell e^{\beta Z_\ell}}{\sum_{\ell\in\mathcal{R}(\tau_j)}e^{\beta Z_\ell}} \end{aligned}$$

$U(\beta)$ is called the **"score"**.

As we discussed earlier in the class, one useful form of a likelihood-based test is the **score test**. This is obtained by using the score $U(\beta)$ evaluated at H_o as a test statistic.

Let's look more closely at the form of the score:

$\delta_j Z_j$ **observed** number of deaths in group 1 at τ_j

$\delta_j \bar{Z}_j$ **expected** number of deaths in group 1 at τ_j

Why? Under $H_0 : \beta = 0$, $\bar{Z}_j$ is simply the number of individuals from group 1 in the risk set at time τ_j (call this r_{1j}), divided by the total number in the risk set at that time (call this r_j). Thus, $\bar{Z}_j$ approximates the probability that given there is a death at τ_j, it is from group 1.

Thus, the score statistic is of the form:

$$\sum_{j=1}^{n} (O_j - E_j)$$

When there are ties, the likelihood has to be replaced by one that allows for ties.

In SAS or Stata:

discrete/exactp → Mantel-Haenszel logrank test

breslow → linear rank version of the logrank test

I already showed you the equivalence of the linear rank logrank test and the Breslow (default) Cox PH model in SAS (p.24-25)

Here is the output from SAS for the leukemia data using the **method=discrete** option:

```
Logrank test with proc lifetest - strata statement

Test of Equality over Strata
                                 Pr >
Test      Chi-Square    DF  Chi-Square

Log-Rank     16.7929     1      0.0001
Wilcoxon     13.4579     1      0.0002
-2Log(LR)    16.4852     1      0.0001

************************************************************
The PHREG Procedure

Data Set: WORK.LEUKEM
Dependent Variable: WEEKS  Time to Relapse
Censoring Variable: REMISS
Censoring Value(s): 0
Ties Handling: DISCRETE

                Testing Global Null Hypothesis: BETA=0

               Without        With
Criterion    Covariates    Covariates    Model Chi-Square

-2 LOG L        165.339       149.086      16.252 with 1 DF (p=0.0001)
Score              .             .         16.793 with 1 DF (p=0.0001)
Wald               .             .         14.132 with 1 DF (p=0.0002)
```

4.7 Confidence Intervals and Tests for the HR

(see Collet 3.4; Hosmer, Lemeshow & May 4.2)

Many software packages provide estimates of β, but the hazard ratio HR$=\exp(\beta)$ is usually the parameter of interest.

We can use the delta method to get standard errors for $\exp(\hat{\beta})$:

$$\widehat{Var(HR)} \;=\; Var(\exp(\hat{\beta})) = \exp(2\hat{\beta})Var(\hat{\beta})$$

Example: Effect of combination ARV therapy on time to death in AIDS patients:

$$\hat{\beta} = -0.497; \quad se(\beta) = 0.244 \quad \text{so } Var(\beta) = (0.244)^2$$

$$\begin{aligned} \widehat{Var(HR)} &= Var(\exp(\hat{\beta})) \\ &= \exp(2 * -0.497) \cdot (0.244)^2 \;=\; (0.3701) * (0.0576) \;=\; 0.0213 \end{aligned}$$

$$\text{so we get:} \quad \widehat{se(HR)} \;=\; \sqrt{0.0213} \;=\; 0.146$$

4.7.1 Constructing confidence intervals for $\exp(\beta)$

Two options: (assuming that β is a scalar)

I. Using $se(\exp\hat{\beta})$ obtained above via the delta method as $se(\exp\hat{\beta}) = \sqrt{[Var(\exp(\hat{\beta}))]}$, calculate the endpoints as:

$$[L, U] \;=\; [\widehat{HR} - 1.96\, se(\widehat{HR}), \widehat{HR} + 1.96\, se(\widehat{HR})]$$

II. Form a confidence interval for $\hat{\beta}$, and then exponentiate the endpoints.

$$[L, U] \;=\; [e^{\hat{\beta} - 1.96 se(\hat{\beta})}, e^{\hat{\beta} + 1.96 se(\hat{\beta})}]$$

Which approach do you think would be the most preferable?

Let's work these out for the combination therapy/MAC study example:

The estimated HR is $\widehat{HR} = e^{(-0.497)} = 0.608$.

I. CI for HR directly:

$$L = 0.608 - (1.96)(0.146) = \qquad U = 0.608 + (1.96)(0.146) =$$

$$(L, U) = 0.608 \pm 0.286$$

$$\text{so the 95\% CI is:} \quad (L, U) = (0.322, 0.894)$$

II. CI for β, then exponentiate:

$$L = \exp\left[-0.497 - (1.96)(0.244)\right] \qquad U = \exp\left[-0.497 + (1.96)(0.244)\right]$$

$$L = \exp(-0.975) \qquad U = \exp(-0.0188)$$

$$\text{so the 95\% CI is:} \quad (L, U) = (0.377, 0.981)$$

Basically, you always want to use the 2nd approach!

4.7.2 Hypothesis Tests: Wald tests

For each covariate of interest, the null hypothesis is

$$H_o : HR_j = 1 \Leftrightarrow \beta_j = 0$$

Wald tests: A Wald test of the above hypothesis is constructed as:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad \text{or} \quad \chi^2 = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right)^2$$

The first form (Z) follows a normal distribution, and the second (χ^2) follows a Chi-square distribution with 1 df. STATA gives the Z statistic, while SAS gives the χ^2_1 test statistic. For both, the p-values are also given, and don't depend on which form, Z or χ^2, is provided.

Notes:

- This test for $\beta_j = 0$ assumes that all other terms in the model are held fixed.
- This test also assumes that the covariate is either binary (a dummy variable or indicator of characteristic) or continuous, so that it represents only a single coefficient β_j.

Wald tests for categorical variables with multiple levels:

If we have a factor A with a levels, then we would need to construct a χ^2 test with $(a-1)$ df, using a test statistic based on a quadratic form:

$$\chi^2_{(a-1)} = \widehat{\boldsymbol{\beta}}'_A Var(\widehat{\boldsymbol{\beta}}_A)^{-1} \widehat{\boldsymbol{\beta}}_A$$

where $\boldsymbol{\beta}_A = (\beta_2, ..., \beta_a)'$ are the $(a-1)$ coefficients corresponding to $Z_2, ..., Z_a$ (or $Z_1, ..., Z_{a-1}$, depending on the reference group).

Notes:

- This test statistic follows a χ^2 distribution with $(a-1)$ df under the null hypothesis, which is $H_0 : \beta_1 = \beta_2 = \cdots \beta_{(a-1)} = 0$.
- Fitting the model will typically give us separate tests for each of the $(a-1)$ coefficients, plus an overall test of the categorical variable.
- Depending on how we include the categorical variable in the model (as separate indicators or as a class variable), some additional code might be required to obtain the overall test.

4.7.3 Hypothesis Tests: Likelihood Ratio Tests

Likelihood Ratio (LR) tests can be conducted to evaluate the addition of q covariates to a model which already includes p covariates.

Consider the following models:

- **Model 1:** (contains only the first p covariates)

$$\frac{\lambda_i(t, \boldsymbol{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p)$$

- **Model 2:** (contains all $(p+q)$ covariates)

$$\frac{\lambda_i(t, \boldsymbol{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_{p+q} Z_{p+q})$$

These are *nested* models. For such nested models, we can construct a **likelihood ratio** test of

$$H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0$$

as 2 times the difference in log-likelihoods:

$$\chi^2_{LR} \;=\; -2\left[\log(\hat{L}(1)) - \log(\hat{L}(2))\right]$$

Under H_o, this test statistic is approximately distributed as χ^2 with q df.

4.7.4 Some Examples of CIs and Tests

Example: MAC Prevention Clinical Trial

ACTG 196 was a randomized clinical trial to study the effects of combination regimens on prevention of MAC (*mycobacterium avium complex*), one of the most common Opportunistic Infections in patients with HIV infection.

The **treatment regimens** were:

- clarithromycin (new)
- rifabutin (standard)
- clarithromycin plus rifabutin (combination)

Other characteristics of trial:

- Patients enrolled between April 1993 and February 1994
- Follow-up ended August 1995

The main intent-to-treat analysis compared the 3 treatment arms with respect to time to MAC.

For additional details, see Benson CA, Williams PL, Cohn DL, et al and the ACTG 196/CPCRA 009 Protocol Team. Clarithromycin or rifabutin alone or in combination for primary prophylaxis of *mycobacterium avium* complex disease in patients with AIDS: A randomized, double-blind, placebo-controlled trial. *Journal of Infectious Disease* 2000; **181**:1289-1297.

Calculating test statistics and CIs using Stata `stcox`

Model 1: Includes 3 covariates (karnof, rif, clari)

```
. use mac

. stset mactime macstat

. stcox karnof rif clari, nohr

        failure _d:  macstat
  analysis time _t:  mactime

Cox regression -- Breslow method for ties

No. of subjects =         1151                 Number of obs    =      1151
No. of failures =          121
Time at risk    =       489509
                                               LR chi2(3)       =     32.01
Log likelihood  =   -754.52813                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |  -.0448295   .0106355    -4.215  0.000    -.0656747   -.0239843
     rif |   .8723819   .2369497     3.682  0.000     .4079691    1.336795
   clari |   .2760775   .2580215     1.070  0.285    -.2296354    .7817903
------------------------------------------------------------------------------
```

Model 2: Adds one additional covariate (CD4)

```
. stcox karnof rif clari cd4, nohr

         failure _d:  macstat
   analysis time _t:  mactime


Cox regression -- Breslow method for ties

No. of subjects =         1151                 Number of obs    =      1151
No. of failures =          121
Time at risk    =       489509
                                               LR chi2(4)       =     63.74
Log likelihood  =   -738.66225                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |  -.0368538   .0106652    -3.456   0.001    -.0577572   -.0159503
     rif |    .880338   .2371111     3.713   0.000     .4156089    1.345067
   clari |   .2530205   .2583478     0.979   0.327     -.253332    .7593729
     cd4 |  -.0183553   .0036839    -4.983   0.000    -.0255757   -.0111349
------------------------------------------------------------------------------
```

Notes:

- If we omit the **nohr** option, we will get the estimated hazard ratio along with 95% confidence intervals using Method II (i.e., forming a CI for the log HR (β), and then exponentiating the bounds)

```
------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |   .9638171   .0102793    -3.456   0.001     .9438791    .9841762
     rif |   2.411715   .5718442     3.713   0.000     1.515293    3.838444
   clari |    1.28791   .3327287     0.979   0.327     .7762102    2.136936
     cd4 |   .9818121   .0036169    -4.983   0.000     .9747486    .9889269
------------------------------------------------------------------------------
```

- We can also compute the hazard ratio ourselves, by exponentiating the coefficients:

$$HR_{cd4} = \exp(-0.01835) = 0.98$$

Why is this HR so close to 1, and yet still highly significant?

What is the interpretation of this HR?

Computing the Wald test:

This is done for us, for any covariate which is a 1 df term in the model.
The wald test for CD4 is $Z = -4.98$ $(p < 0.001)$.

Computing the LR test:

Stata gives us an overall LR test of 32.01 for Model 1 and 63.74 for Model 2, but these LR tests are overall tests for all of the covariates in the model (eg., with 3 df and 4 df, respectively).

To get a LR test for just the single additional covariate we added (`cd4`), we can calculate the likelihood ratio test as twice the difference in minus log-likelihoods between the two models:

$$\chi^2_{LR} = 2 * (754.528 - (738.662)) = 31.73$$

(Note that this is actually the same as the difference in the overall LR tests for all covariates in the model; 63.74-32.01=31.73)

How does this test statistic compare to the Wald χ^2 test?

Obtaining LR statistic from Stata directly

Can we obtain this LR test statistic from the software, including the p-value, without having to perform the calculations manually?

Yes!!

After fitting the first model, store the parameter estimates (and log-likelihood) using any name. Then fit the 2nd model, and again store the estimates:

```
.stcox karnof rif clari            [output suppressed, same as above]

.estimates store model1

.stcox karnof rif clari cd4        [output suppressed, same as above]

.estimates store model2

.lrtest model1 model2

Likelihood-ratio test                                 LR chi2(1)  =     31.73
(Assumption: model1 nested in model2)                 Prob > chi2 =    0.0000
```

Computing the Wald test for the 2df Treatment Effect:

- In the MAC study, there were three treatment arms (rif, clari, and the rif+clari combination). Because we have only included the **rif** and **clari** effects in the model, the combination therapy is the "reference" group.
- We can conduct a 2df Wald test of whether there is any difference among the three treatment arms using the **test** command in Stata:

```
. test rif clari

 ( 1)  rif = 0.0
 ( 2)  clari = 0.0

           chi2(  2) =   17.01
         Prob > chi2 =    0.0002
```

This is a test of whether $\beta_{rif} = 0$ and $\beta_{clari} = 0$, but both β's reflect the log(HR) relative to the combination group.

- The **test** command can also be used to test whether there is a difference between the **rif** and **clari** treatment arms:

```
.  test rif=clari

 ( 1)  rif - clari = 0.0

           chi2(  1) =    8.76
         Prob > chi2 =    0.0031
```

Calculating test statistics and CIs using SAS `proc phreg`

```
proc phreg data=alloi;
  model dthtime*dthstat(0)=mlogrna cd4grp1 cd4grp2 combther
        / risklimits;
  cd4level: test cd4grp1, cd4grp2;
  title1 'Proportional hazards regression model for time to Death';
  title2 'Model 1: Baseline viral load and CD4 predictors';
```

```
proc phreg data=alloi;
  model dthtime*dthstat(0)=mlogrna cd4grp1 cd4grp2 combther decrs8 incrs8
        / risklimits;
  cd4level: test cd4grp1, cd4grp2;
  wk8resp: test decrs8, incrs8;
  title2 'Model 2: Changes in CD4 and Viral load';
```

Class variables in SAS PROC PHREG:

A categorical variable can now be included in SAS PROC PHREG using a `class` statement.

```
data alloi2;
  set alloi;
  if cd4grp1=1 then cd4group=1;
    else if cd4grp2=1 then cd4group=2;
    else if cd4grp1^=. and cd4grp2^=. then cd4group=3;

proc format;
  value cd4group
    1='<100'
    2='100-200'
    3='>200';

proc phreg data=alloi2;
  class cd4group;
  model dthtime*dthstat(0)=mlogrna cd4group combther decrs8 incrs8
        / risklimits;
  format cd4group cd4group.;
  title2 'Model 3: Class variable for CD4 categories';
```

This syntax will give a 2df χ^2 test for all three CD4 categories adjusting for the other covariates in the model, without having to use a `test` statement.

Notes:

- The "risklimits" option (or "rl" for short) on the model statement provides 95% confidence intervals using the preferred approach (i.e., forming a CI for the log HR (β), and then exponentiating the bounds)

- The "test" statement has the following form:

  ```
  Label: test varname1, varname2, ..., varnamek;
  ```

  for a k df Wald chi-square test of whether the k coefficients are all equal to 0.

OUTPUT FROM PROC PHREG: (Model 1)

Proportional hazards regression model for time to Death
Baseline viral load and CD4 predictors

Data Set: WORK.ALLOI
Dependent Variable: DTHTIME Time to death (days)
Censoring Variable: DTHSTAT Death status (1=died,0=censored)
Censoring Value(s): 0
Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
690	89	601	87.10

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	1072.543	924.167	148.376 with 4 DF (p=0.0001)
Score	.	.	189.702 with 4 DF (p=0.0001)
Wald	.	.	127.844 with 4 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
MLOGRNA	1	0.833237	0.17808	21.89295	0.0001
CD4GRP1	1	2.364612	0.32436	53.14442	0.0001
CD4GRP2	1	1.171137	0.34434	11.56739	0.0007
COMBTHER	1	-0.497161	0.24389	4.15520	0.0415

OUTPUT FROM PROC PHREG, Model 1 continued

Output from "risklimits" and "test" statements

Analysis of Maximum Likelihood Estimates

Conditional Risk Ratio and 95% Confidence Limits

Variable	Risk Ratio	Lower	Upper	Label
MLOGRNA	2.301	1.623	3.262	log baseline rna (roche assay)
CD4GRP1	10.640	5.634	20.093	CD4<=100
CD4GRP2	3.226	1.643	6.335	100<CD4<=200
COMBTHER	0.608	0.377	0.981	Combination therapy with AZT/ddI/ddC/Nvp

Linear Hypotheses Testing

Label	Wald Chi-Square	DF	Pr > Chi-Square
CD4LEVEL	55.0794	2	0.0001

OUTPUT FROM PROC PHREG: (Model 2)

Proportional hazards regression model for time to Death
Baseline viral load and CD4 predictors

Data Set: WORK.ALLOI
Dependent Variable: DTHTIME Time to death (days)
Censoring Variable: DTHSTAT Death status (1=died,0=censored)
Censoring Value(s): 0
Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
690	89	601	87.10

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	1072.543	912.009	160.535 with 6 DF (p=0.0001)
Score	.	.	198.537 with 6 DF (p=0.0001)
Wald	.	.	132.091 with 6 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
MLOGRNA	1	0.893838	0.18062	24.48880	0.0001
CD4GRP1	1	2.023005	0.33594	36.26461	0.0001
CD4GRP2	1	1.001046	0.34907	8.22394	0.0041
COMBTHER	1	-0.456506	0.24687	3.41950	0.0644
DECRS8	1	-0.410919	0.26383	2.42579	0.1194
INCRS8	1	-0.834101	0.32884	6.43367	0.0112

OUTPUT FROM PROC PHREG, Model 2 continued

Output from "risklimits" and "test" statements

Analysis of Maximum Likelihood Estimates

Conditional Risk Ratio and
95% Confidence Limits

Variable	Risk Ratio	Lower	Upper	Label
MLOGRNA	2.444	1.716	3.483	log baseline rna (roche assay)
CD4GRP1	7.561	3.914	14.606	CD4<=100
CD4GRP2	2.721	1.373	5.394	100<CD4<=200
COMBTHER	0.633	0.390	1.028	Combination therapy with AZT/ddI/ddC/Nvp
DECRS8	0.663	0.395	1.112	Decrease>=0.5 log rna at week 8?
INCRS8	0.434	0.228	0.827	Increase>=50 CD4 cells, week 8?

Linear Hypotheses Testing

Label	Wald Chi-Square	DF	Pr > Chi-Square
CD4LEVEL	37.6833	2	0.0001
WK8RESP	10.4312	2	0.0054

Calculation of LR Test for 8-week Response Measures:

Since Model 1 is nested in Model 2, we can compare the $-2\log L$ values for the two models:

- **Model 1:** $-2\log L = 924.167$
- **Model 2:** $-2\log L = 912.009$
- **Difference:** $924.167 - 912.009 = 12.158$

The 2 df LR Test for the 8-week response measures is slightly larger than the Wald test

$$\chi^2_{LR} = 12.1584$$

$$\chi^2_{wald} = 10.4312, \quad p = 0.0054$$

so the p-value would be something less than 0.0054.

Note of Caution: when comparing two models such as above, you have to make sure that there are the exact same number of observations in the two models. If some participants were missing the covariate values which are in one model but not the other (eg., the 8-week response measures could be missing for some subjects if they discontinued the study), then the models would no longer be nested.

Calculating Percent of Treatment Effect:

- We can use an approach described by Freedman (1982) to assess the effects of intermediate endpoints (incrs8, decrs8) on the treatment effect (i.e., assess their use as surrogate markers). The percentage of treatment effect explained, γ, is estimated by:

$$\hat{\gamma} \;=\; 1 - \frac{\hat{\beta}_{trt,M2}}{\hat{\beta}_{trt,M1}}$$

  where M1 is the model without the intermediate endpoint and M2 is the model with the marker.

- The percentage of treatment effect explained by including the RNA and CD4 response to treatment by Week 8 is:

$$\hat{\gamma} \;=\; 1 - \frac{-0.456}{-0.497} \approx 0.08$$

  or 8%. The percentage of treatment effect on time to first opportunistic infection or death is much higher (about 24%).

- Causal inference has come a long way since 1982, and the above measure has no clear causal interpretation. More sophisticated approaches have been developed to account for intermediate covariates, particularly those which are both affected by treatment and affect future outcomes.

4.7.5 Hypothesis tests: Score tests (preview)

We have already seen that if there is only one covariate in a Cox model, and it is a binary one, then the Score test (with 1df) is identical to the logrank test comparing the two survival distributions.

Similarly, if there is just one covariate (whether binary, continuous, ordinal, etc), the Score Statistic is given automatically in SAS Proc Phreg.

The examples above also show that it is easy (in SAS, at least), to get a Score test for ALL of the covariates in the model.

Example: For Model 1 fit to the "ALLOI" dataset, the overall Score statistic for testing all 4 covariates (`mlogrna`, `cd4grp1`, `cd4grp2`, and `combther`) is:

$$\chi^2_{score} \quad = \quad 189.702$$

This χ^2 statistic has 4 df since it is testing whether four β's are simultaneously equal to 0.

But what about a Score statistic for 1 or more covariates, adjusted for the other covariates in the model?

Score tests adjusted for other covariates

It turns out that a score test for a single covariate adjusted for other covariates can be obtained in SAS by fitting the model with the covariates we want to adjust for, and then using a *forward selection* approach to force the covariate of interest to be added to the model.

For example, the following code could be used:

```
proc phreg data=alloi;
  class cd4group;
  model dthtime*dthstat(0)= cd4group sex / risklimits
    selection=forward include=1 slentry=0.20;
  title2 'Obtaining score test for sex adjusted for other covariates';
run;
```

We will return to Score tests again later once we discuss model selection procedures.

We will also discuss the efficiency of the three different types of tests (Wald, Likelihood Ratio, and Score).

Partial output from forward selection procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
cd4group	1	1	3.04847	0.28835	111.7704	<.0001
cd4group	2	1	1.68279	0.29599	32.3220	<.0001
SEX		1	-0.68434	0.37024	3.4164	0.0646

Analysis of Maximum Likelihood Estimates

Parameter		Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
cd4group	1	21.083	11.981	37.100	cd4group 1
cd4group	2	5.381	3.012	9.611	cd4group 2
SEX		0.504	0.244	1.042	sex (1=male, 2=female)

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Effect Label
1	SEX	1	2	3.5488	0.0596	sex (1=male, 2=female)

An easier way to obtain Score tests

In the latest version of SAS (9.3), one can also obtain Score tests by including a `type3` option on the model statement:

```
proc phreg data=alloi;
  class cd4group;
  model dthtime*dthstat(0)= cd4group sex / risklimits type3(score);
  title2 'Obtaining score test adjusted for other covars using type3 option';
run;
```

You can also specify options for `wald` and `lr` tests, or to get all three chi-square tests, you can use the "`all`" option (equivalent to `type3(wald lr score)`):

```
proc phreg data=alloi;
  class cd4group;
  model dthtime*dthstat(0)= cd4group sex / risklimits type3(all);
  title2 'Obtaining Score, Wald, and LR tests adjusted for other covars using type3 option';
run;
```

Obtaining Score, Wald, and LR tests adjusted for other covars using type3 option

Type 3 Tests

LR Statistics

Effect	DF	Chi-Square	Pr > ChiSq
cd4group	2	127.9274	<.0001
SEX	1	4.1311	0.0421

Score Statistics

Effect	DF	Chi-Square	Pr > ChiSq
cd4group	2	168.2736	<.0001
SEX	1	3.5488	0.0596

Wald Statistics

Effect	DF	Chi-Square	Pr > ChiSq
cd4group	2	113.2070	<.0001
SEX	1	3.4164	0.0646

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
cd4group	1	1	3.04847	0.28835	111.7704	<.0001
cd4group	2	1	1.68279	0.29599	32.3220	<.0001
SEX		1	-0.68434	0.37024	3.4164	0.0646

4.8 Predicted Survival using PH

The Cox PH model says that $\lambda_i(t, \boldsymbol{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta Z})$. What does this imply about the survival function, $S_z(t)$, for the i-th individual with covariates $\boldsymbol{Z}_i$?

For the baseline (reference) group, we have:

$$S_0(t) \;=\; e^{-\int_0^t \lambda_0(u)du} = e^{-\Lambda_0(t)}$$

This is by definition of a survival function (see intro notes).

For the i-th patient with covariates $\boldsymbol{Z}_i$, we have:

$$\begin{aligned}
S_i(t) &= e^{-\int_0^t \lambda_i(u)du} = e^{-\Lambda_i(t)} \\
&= e^{-\int_0^t \lambda_0(u)\exp(\boldsymbol{\beta Z}_i)du} \\
&= e^{-\exp(\boldsymbol{\beta Z}_i)\int_0^t \lambda_0(u)du} \\
&= \left[e^{-\int_0^t \lambda_0(u)du}\right]^{\exp(\boldsymbol{\beta Z}_i)} \text{ since } \ [\mathrm{e}^{\mathrm{b}}]^{\mathrm{a}} = \mathrm{e}^{\mathrm{ab}} \\
&= [S_0(t)]^{\exp(\boldsymbol{\beta Z}_i)} \\
&= [S_0(t)]^{HR}
\end{aligned}$$

So after calculating $\hat{\beta}$ using the MPLE, we can obtain an estimate for $S_i(t)$ as

$$\hat{S}_i(t) = \left[\hat{S}_0(t)\right]^{\widehat{HR}}$$

Say we are interested in the survival pattern for single males in the nursing home study. Based on the previous formula, if we had an estimate for the survival function in the reference group, i.e., $\hat{S}_0(t)$, we could get estimates of the survival function for any set of covariates $\boldsymbol{Z}_i$.

How can we estimate the survival function, $S_0(t)$?

We could use the KM estimator, but there are a few disadvantages of that approach:

- It would only use the survival times for observations contained in the reference group, and not all the rest of the survival times.
- It would tend to be somewhat choppy, since it would reflect the smaller sample size of the reference group.
- It's possible that there are no subjects in the dataset who are in the "reference" group (ex. say covariates are age and sex; there is no one of age=0 in our dataset).

Instead, we will use a baseline hazard estimator which takes advantage of the proportional hazards assumption to get a smoother estimate.

$$\hat{S}_i(t) \;=\; [\hat{S}_0(t)]^{\exp(\widehat{\boldsymbol{\beta}}\boldsymbol{Z}_i)}$$

Using the above formula, we substitute $\widehat{\boldsymbol{\beta}}$ based on fitting the Cox PH model, and calculate $\hat{S}_0(t)$ by one of the following approaches:

- Breslow estimator (Stata and now SAS)
- Kalbfleisch/Prentice estimator

(1) **Breslow Estimator:**

$$\hat{S}_0(t) = \exp^{-\hat{\Lambda}_0(t)}$$

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

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k\in\mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

(2) **Kalbfleisch/Prentice Estimator**

$$\hat{S}_0(t) = \prod_{j:\tau_j<t} \hat{\alpha}_j$$

where $\hat{\alpha}_j, j = 1, ...d$ are the MLE's obtained by assuming that $S(t; Z)$ satisfies

$$S(t; Z) = [S_0(t)]^{e^{\beta Z}} = \left[\prod_{j:\tau_j<t} \alpha_j \right]^{e^{\beta Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta Z}}$$

Breslow Estimator: further motivation

The Breslow estimator is based on extending the concept of the Nelson-Aalen estimator to the proportional hazards model.

Recall that for a single sample with no covariates, the **Nelson-Aalen Estimator** of the cumulative hazard is:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \frac{d_j}{r_j}$$

where d_j and r_j are the number of deaths and the number at risk, respectively, at the j-th death time.

When there are covariates and assuming the PH model above, one can generalize this to estimate the cumulative baseline hazard by adjusting the denominator:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k\in\mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

Heuristic behind Breslow estimator:

The expected number of failures in $(t, t+\delta t)$ is

$$d_j \approx \delta t \times \sum_{k \in \mathcal{R}(t)} \lambda_0(t) \exp(z_k \hat{\beta})$$

Hence,

$$\delta t \times \lambda_0(t_j) \approx \frac{d_j}{\sum_{k \in \mathcal{R}(t)} \exp(z_k \hat{\beta})}$$

Kalbfleisch/Prentice Estimator: further motivation

This method is analogous to the Kaplan-Meier Estimator. Consider a discrete time model with hazard $(1-\alpha_j)$ at the j-th observed death time.

(Note: we use $\alpha_j = (1-\lambda_j)$ to simplify the algebra!)

Thus, for someone with Z=0, the survivorship function is

$$S_0(t) = \prod_{j:\tau_j<t} (1-\lambda_j) \;=\; \prod_{j:\tau_j<t} \alpha_j$$

and for someone with $Z \neq 0$, it is:

$$S(t;Z) = S_0(t)^{e^{\beta Z}} = \left[\prod_{j:\tau_j<t} \alpha_j\right]^{e^{\beta Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta Z}}$$

So α_j represents the conditional probability of surviving to time t_j given that one has survived to time $t_{(j-1)}$, *within the reference group* (eg, Z=0).

The likelihood contributions under this model are:

- for someone censored at t: $S(t; Z)$
- for someone who fails at t_j:

$$S(t_{(j-1)}; Z) - S(t_j; Z) = \left[\prod_{k<j} \alpha_k\right]^{e^{\beta z}} [1 - \alpha_j^{e^{\beta Z}}]$$

$$\text{or equivalently } \frac{S(t_j; Z)}{S(t_{(j-1)}; Z)} = \left[\frac{S_0(t_j)}{S_0(t_{(j-1)})}\right]^{e^{\beta Z}} = \alpha_j^{e^{\beta Z}}$$

Following the usual Maximum Likelihood approach, and substituting the partial likelihood estimates $\hat{\beta}$, the solution for α_j satisfies:

$$\sum_{k \in \mathcal{D}_j} \frac{\exp(Z_k \hat{\beta})}{1 - \alpha_j^{\exp(Z_k \hat{\beta})}} = \sum_{k \in \mathcal{R}_j} \exp(Z_k \hat{\beta})$$

(Notes: the above notation allows for tied events D_j at time t_j, but the form simplifies if there are no ties. Also note what happens when $Z = 0 \Rightarrow \alpha_j = d_j / r_j$)

Obtaining $\hat{S}_0(t)$ from software packages

- Stata provides the Breslow estimator of $S_0(t; Z)$, but not predicted survivals at specified covariate values..... you have to construct these yourself

- SAS used to provide the Kalbfleisch/Prentice estimator of the baseline hazard but now gives the Breslow estimator. SAS can provide estimates of survival at arbitrary values of the covariates with a little bit of programming.

In practice, the two approaches are **incredibly** close! (see Fleming and Harrington, *Communications in Statistics* 1984)

4.8.1 Using SAS to Predict Survival

The SAS command BASELINE calculates the predicted survival values at the event times for a given set of covariate values.

(1) To get the estimated baseline survival $\hat{S}_0(t)$, create a dataset with 0's for values of all covariates in the model

(2) To get the estimated survival $\hat{S}_i(t)$ for any other subgroup (i.e., not the reference or baseline group), create a data set which inputs the baseline values of the covariates for the subgroup of interest.

For either case, we then supply the corresponding dataset name to the BASELINE command under PROC PHREG.

By giving the input dataset several lines, each corresponding to a different combination of covariate values, we can compute predicted survival values for more than one group at once.

(1) Baseline Survival Estimate

(note that the baseline survival function does not correspond to any observations in our sample, since health status values range from 2-5)

```
*** Estimating Baseline Survival Function under PH;
data inrisks;
  input married health;
  cards;
0 0
;

proc phreg data=pop out=survres;
  model los*fail(0)=married health;
  baseline covariates=inrisks out=outph survival=ps/nomean;

proc print data=outph;
title1 'Nursinghome data: Baseline Survival Estimate';
```

Estimating the Baseline Survival with SAS

Nursinghome data: Baseline Survival Estimate

OBS	MARRIED	HEALTH	LOS	PS
1	0	0	0	1.00000
2	0	0	1	0.99253
3	0	0	2	0.98672
4	0	0	3	0.98363
5	0	0	4	0.97776
6	0	0	5	0.97012
7	0	0	6	0.96488
8	0	0	7	0.95856
9	0	0	8	0.95361
10	0	0	9	0.94793
11	0	0	10	0.94365
12	0	0	11	0.93792
13	0	0	12	0.93323
14	0	0	13	0.92706
15	0	0	14	0.92049
16	0	0	15	0.91461
17	0	0	16	0.91017
18	0	0	17	0.90534
19	0	0	18	0.90048
20	0	0	19	0.89635
.				
.				
.				
287	0	0	364	0.49934
288	0	0	365	0.49872

(2) Predicted Survival Estimate for Subgroups

The following SAS commands will generate the predicted survival probability for each combination of covariates, at every observed event time in the dataset.

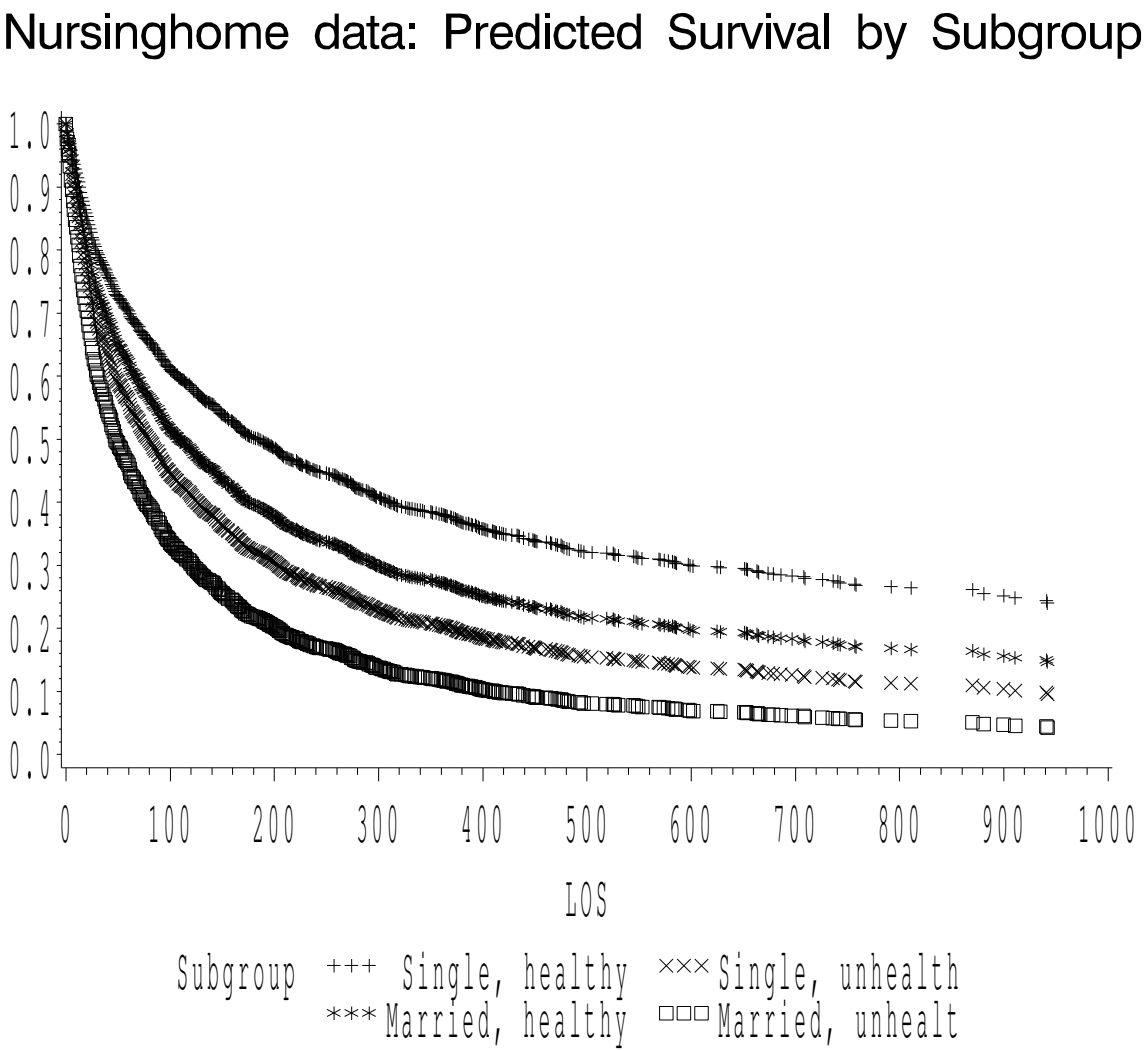
```
*** Estimating Baseline Survival Function under PH;
data inrisks;
  input married health;
  cards;
0 2
0 5
1 2
1 5
;

proc phreg data=pop out=survres;
  model los*fail(0)=married health;
  baseline covariates=inrisks out=outph survival=ps/nomean;

proc print data=outph;
title1 'Nursinghome data: predicted survival by subgroup';
```

Survival Estimates by Marital and Health Status

Nursinghome data: Predicted Survival by Subgroup

OBS	MARRIED	HEALTH	LOS	PS
1	0	2	0	1.00000
2	0	2	1	0.98961
3	0	2	2	0.98156
171	0	2	184	0.50104
172	0	2	185	0.49984
396	0	5	0	1.00000
397	0	5	1	0.98300
398	0	5	2	0.96988
474	0	5	78	0.50268
475	0	5	80	0.49991
791	1	2	0	1.00000
792	1	2	1	0.98605
793	1	2	2	0.97527
897	1	2	108	0.50114
898	1	2	109	0.49986
1186	1	5	0	1.00000
1187	1	5	1	0.97719
1188	1	5	2	0.95969
1233	1	5	47	0.50519
1234	1	5	48	0.49875

We can get a visual picture of what the proportional hazards assumption implies by looking at these four subgroups



4.8.2 Using Stata to Predict Survival

The Stata command `basesurv` calculates the predicted survival values for the reference group, i.e., those subjects with all covariates=0.

(1) **Baseline Survival:**
To obtain the estimated baseline survival $\hat{S}_0(t)$, follow the example below (for the nursing home data):

```
. use nurshome

. stset los fail

. stcox married health, basesurv(prsurv)

. sort los

. list los prsurv
```

Estimating the Baseline Survival with Stata

```
            los      prsurv

   1.         1   .99252899
   2.         1   .99252899
   3.         1   .99252899
   4.         1   .99252899
   5.         1   .99252899
.
.
.
  22.         1   .99252899
  23.         2   .98671824
  24.         2   .98671824
  25.         2   .98671824
  26.         2   .98671824
  27.         2   .98671824
  28.         2   .98671824
  29.         2   .98671824
.
.
.
  40.         3   .98362595
  41.         3   .98362595
.
.
.
```

Stata creates a predicted baseline survival estimate for every observed event time in the dataset, even if there are duplicates.

(2) **Predicted Survival for Subgroups**

To obtain the estimated survival $\hat{S}_i(t)$ for any other subgroup (i.e., not the reference or baseline group), follow the Stata commands below:

```
. predict betaz, xb

. gen newterm=exp(betaz)

. gen predsurv=prsurv^newterm

. sort married health los

. list married health los predsurv
```

Predicting Survival for Subgroups with Stata

```
       married     health        los   predsurv

  1.         0          2          1   .9896138
  8.         0          2          2   .9815570
 11.         0          2          3   .9772769
 13.         0          2          4   .9691724
 16.         0          2          5   .9586483
.....................................................................
300.         0          3          1   .9877566
302.         0          3          2   .9782748
304.         0          3          3   .9732435
305.         0          3          4   .9637272
312.         0          3          5   .9513916
.....................................................................
768.         0          4          1   .9855696
777.         0          4          2   .9744162
779.         0          4          3   .9685058
781.         0          4          4   .9573418
785.         0          4          5   .9428996
.
.
.
1468.        1          4          1   .9806339
1469.        1          4          2   .9657326
1472.        1          4          3   .9578599
1473.        1          4          5   .9239448
.....................................................................
1559.        1          5          1   .9771894
1560.        1          5          2   .9596928
1562.        1          5          3   .9504684
1564.        1          5          4   .9331349
```

4.9 Predicted median survival

Predicted Medians

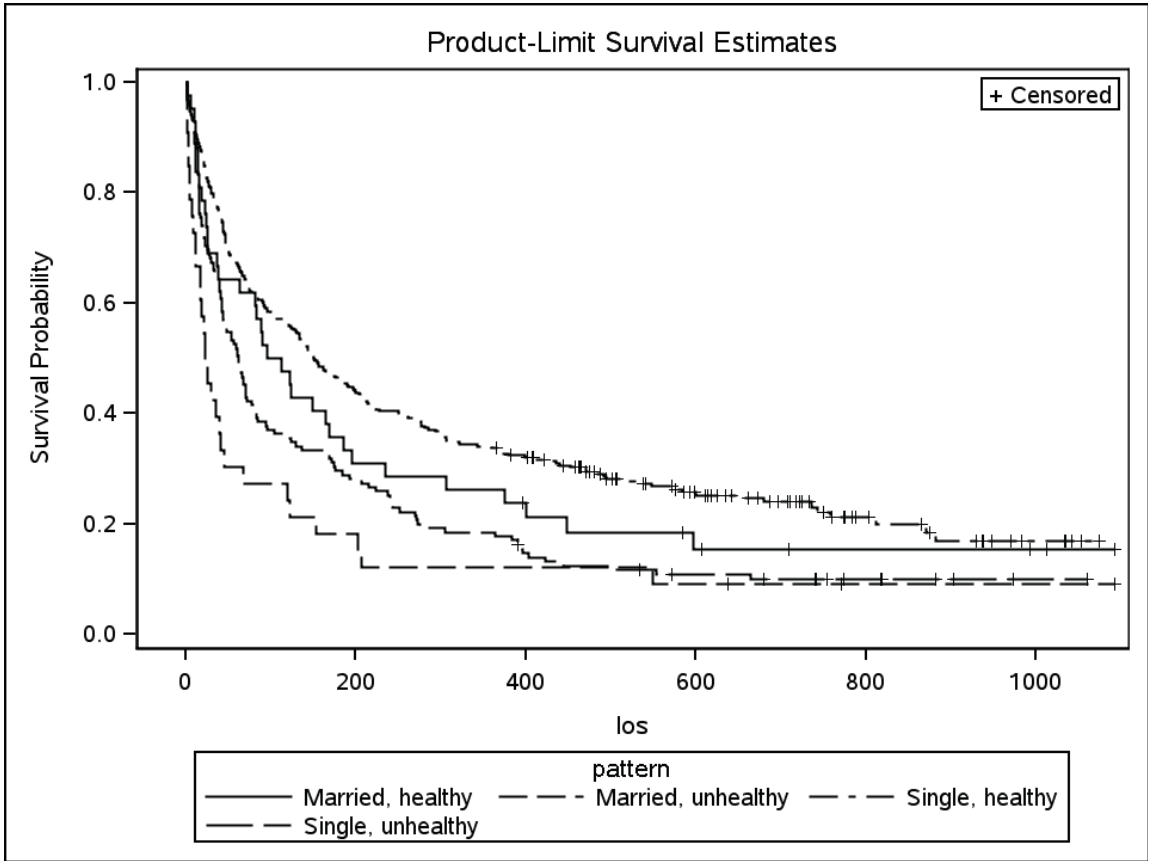
Suppose we want to find the predicted median survival for an individual with a specified combination of covariates (e.g., a single person with health status 5).

Three possible approaches:

(1) Calculate the median from the subset of individuals with the specified covariate combination (using KM approach)

- Create separate datasets, one for each combination of covariates, then calculate KM estimate for each dataset
- OR...Use a single dataset, but stratify by a variable representing combination of covariates of interest (might have to create this new variable first)

Here is the set of KM survival estimates based on a stratified analysis by `marhealth`, created as `married*health`:

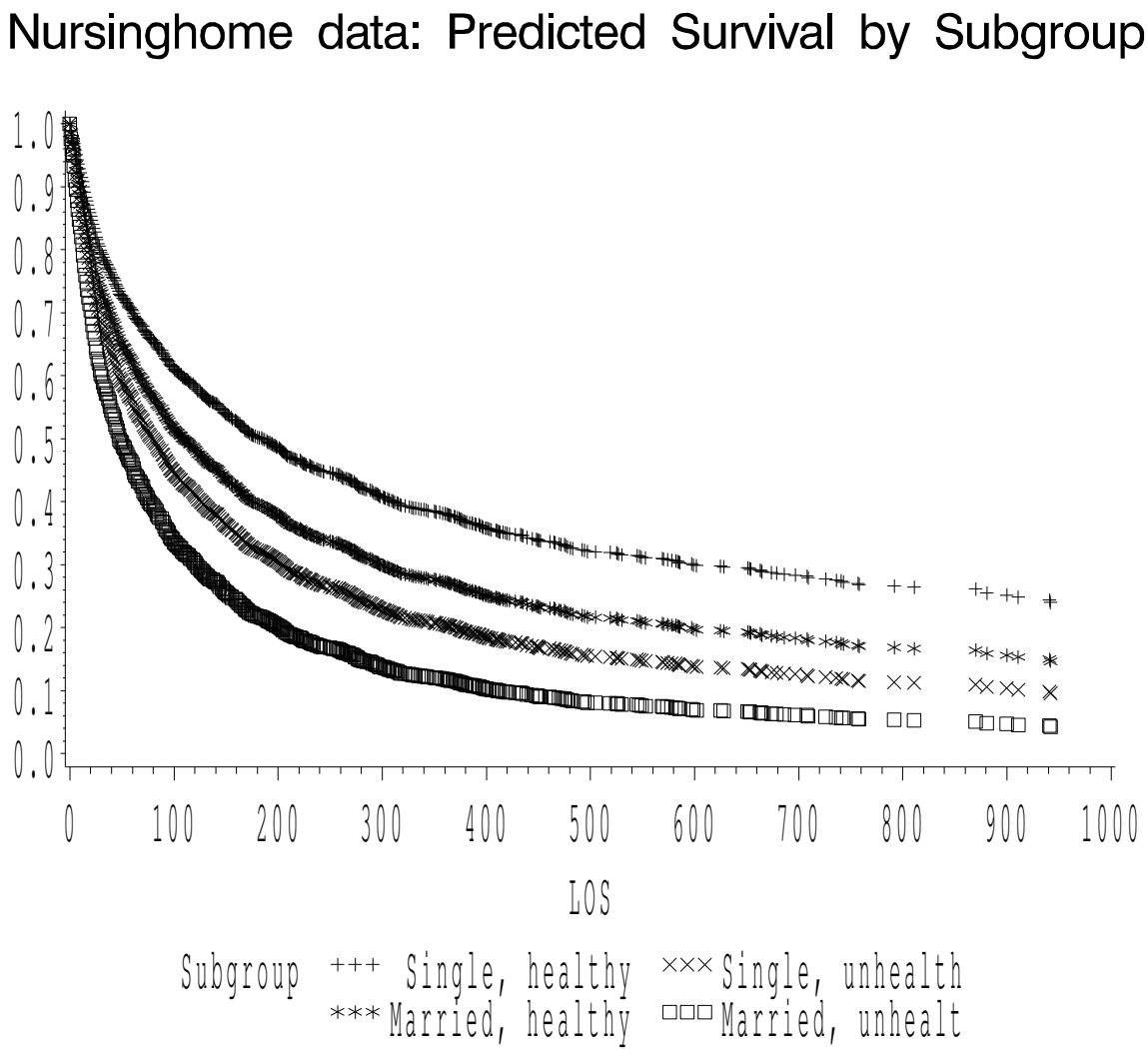


(2) Generate predicted survival curves for each combination of covariates, and obtain the medians directly

OBS	MARRIED	HEALTH	LOS	PREDSURV
171	0	2	184	0.50104
172	0	2	185	0.49984
474	0	5	78	0.50268
475	0	5	80	0.49991
897	1	2	108	0.50114
898	1	2	109	0.49986
1233	1	5	47	0.50519
1234	1	5	48	0.49875

Recall that previously we defined the median as the *smallest* value of t for which $\hat{S}(t) \leq 0.5$, so the medians from above would be 185, 80, 109, and 48 days for single healthy, single unhealthy, married healthy, and married unhealthy, respectively.

This is also what we would get by drawing a horizontal line at 0.5 for the survival probability, and seeing where it intersects predicted survival curve for each group:



(3) Generate the predicted survival curve from the estimated baseline hazard, as follows:

We want the estimated median (M) for an individual with covariates $\boldsymbol{Z}_i$. We know

$$S(M;Z) = [S_0(M)]^{e^{\beta Z_i}} = 0.5$$

Hence, M satisfies (raising both sides by $e^{-\beta Z_i}$):

$$S_0(M) = [0.5]^{e^{-\beta Z}}$$

Ex. Suppose we want to estimate the median survival for a single unhealthy subject from the nursing home data. The reciprocal of the hazard ratio for unhealthy (health=5) is: $e^{-0.165*5} = 0.4373$, (where $\hat{\beta} = 0.165$ for health status)

So, we want M such that $S_0(M) = (0.5)^{0.4373} = 0.7385$

So the median for single unhealthy subject is the 73.8^{th} percentile of the baseline group.

OBS	MARRIED	HEALTH	LOS	PREDSURV
79	0	0	78	0.74028
80	0	0	80	0.73849
81	0	0	81	0.73670

So the estimated median would still be 80 days. Note: similar logic can be followed to estimate other quantiles besides the median.

4.10 Estimating P-year survival

Suppose we want to find the P-year survival rate for an individual with a specified combination of covariates, $\hat{S}(P; \boldsymbol{Z}_i)$

For an individual with $\boldsymbol{Z}_i = 0$, the P-year survival can be obtained from the baseline survivorship function, $\hat{S}_0(P)$

For individuals with $\boldsymbol{Z}_i \neq 0$, it can be obtained as:

$$\hat{S}(P; \boldsymbol{Z}_i) = [\hat{S}_0(P)]^{e^{\widehat{\boldsymbol{\beta}}\boldsymbol{Z}_i}}$$

Notes:

- Although I say "P-year" survival, the units of time in a particular dataset may be days, weeks, or months. The answer here will be in the same units of time as the original data.
- **Example:** predicted 365-day survival for a single unhealthy subject in Nursing Home study (`MARRIED=0, HEALTH=5`):

$$\begin{aligned}
\hat{S}(P; \boldsymbol{Z}_i) &= [\hat{S}_0(365)]^{e^{\widehat{\boldsymbol{\beta}}\boldsymbol{Z}_i}} \\
&= [\hat{S}_0(365)]^{e^{0.165*5}} \\
&= [\hat{S}_0(365)]^{(2.282)} \\
&= (0.500)^{(2.282)} = 0.206 \qquad (\text{assuming } \hat{S}_0(365) = 0.50)
\end{aligned}$$

- If $\widehat{\boldsymbol{\beta}}\boldsymbol{Z}_i$ is positive, then the P-year survival rate for the i-th individual will be lower than for a baseline individual.

  Why is this true?

Contents