BST 210
Applied Regression Analysis

Lecture 23
Plan for Today

Survival Analysis continued:

- Recall: new outcome variable in town!
- Final notes on Kaplan-Meier
- CIs for S(t) using Greenwood
- Log-rank tests to compare S(t) for 2 or more groups
- Hazard function
- Semiparametric survival analysis with Cox proportional hazards model
- Cox model interpretation and assessing PH assumption

Recall - Survival Data: It's just a matter of 'time'

- In some studies, the response variable of interest is the ***length of time* between an initial observation and the occurrence of a subsequent event**
- The event is often called a *failure*
- The time from the initial observation until failure is called the ***survival time***
- Examples: Time from birth until death, time from start of treatment until serious adverse event, time from randomization to relapse or death, time from entry in a cohort study until myocardial infarction, time between live births, time until marriage, duration of geographic stay, etc.
- Distributions of survival times tend to be right skewed

Recall - Goals of Survival Analysis

Same themes as previous linear modeling…

- To estimate the distribution of survival times for a population
- To test the equality of survival distributions (e.g., treated vs. control group, smokers vs. nonsmokers)
- To estimate and control for the effects of other covariates when investigating the relationship between a predictor variable and survival time
- To assess model fit

Recall - Functions defining time T

- There are essentially 4 related functions, all which help to define our survival (or failure) times T; we displayed 3 in last class, and will look at 4th today

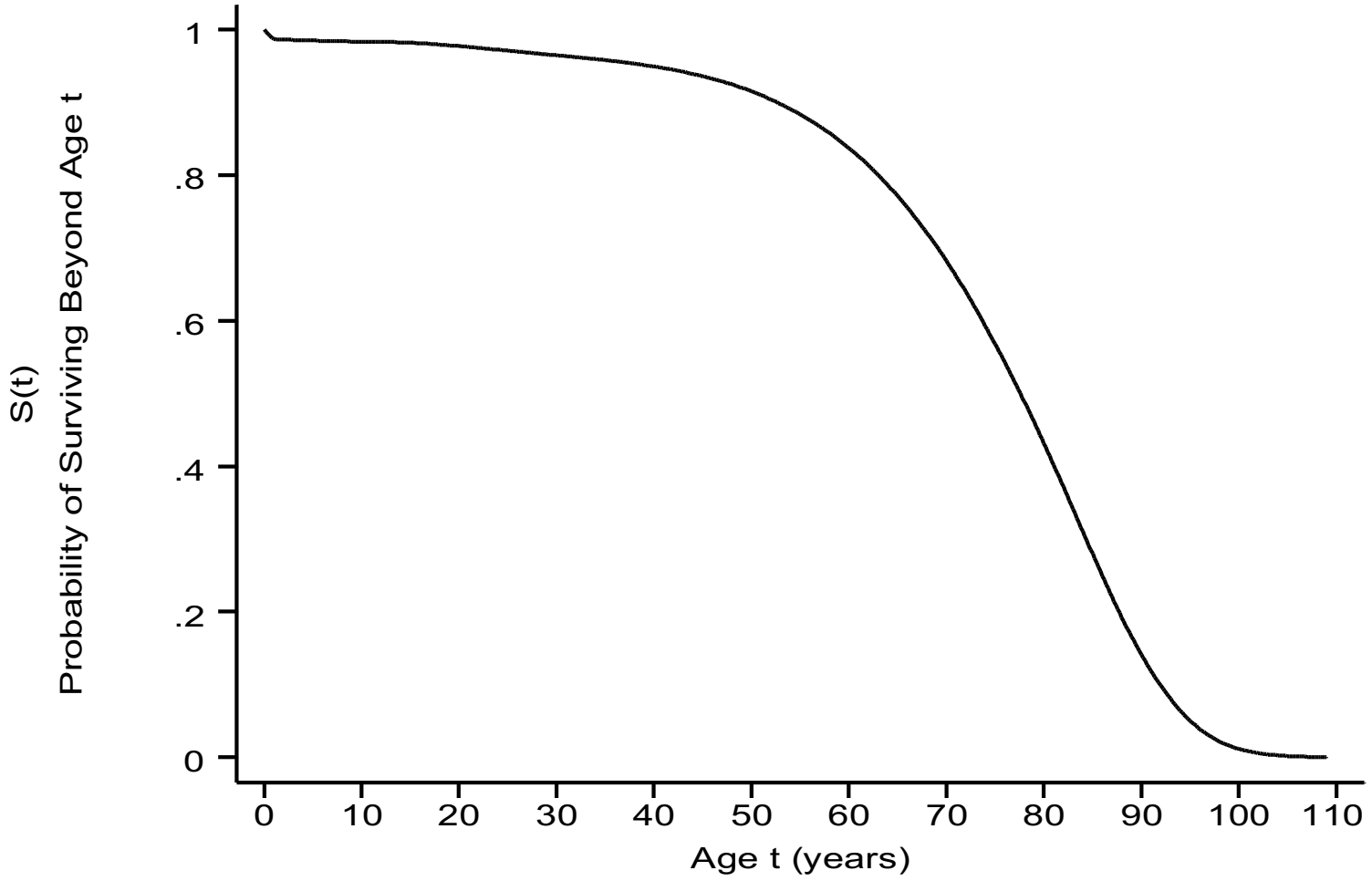
- Let random variable *T* represent the time from a start point to an event of interest (e.g., time from start of treatment to serious adverse event, time from disease remission to recurrence)

- The *survival function* S(*t*) is then defined as:

$$S(t) = P(T > t)$$

- S(t) is the proportion of individuals who are event-free at time t, or the probability of not having the event until time t.

Recall - Survival Curve (of survival probabilities)



Recall - Example: Chemotherapy for Leukemia

- There were 11 patients randomized to the maintenance chemotherapy group, and 12 patients randomized to the control group

- Time to relapse was measured for each subject

- Maintenance chemotherapy group (*n* = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (*n* = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

→ *Where the + signs represent censored observations*

Recall - Kaplan-Meier with Chemotherapy data (with censoring)

j	n_j	t_j	d_j	$S(t_j)= \prod(1 - \frac{d_i}{n_i})$ for $i=1,\ldots,j$ *(cumulative)*
0	11	0	0	1
1	11	9	1	1 – (1/11) = 10/11 = 0.909
2	10	13	1	(10/11) x (9/10) = 0.818
3	8	18	1	(10/11) x (9/10) x (7/8) = 0.716
4	7	23	1	(10/11) x (9/10) x (7/8) x (6/7) = 0.614
5	5	31	1	(10/11) x (9/10) x (7/8) x (6/7) x (4/5) = 0.491
6	4	34	1	(10/11) x (9/10) x (7/8) x (6/7) x (4/5) x (3/4) = 0.368
7	2	48	1	(10/11) x (9/10) x (7/8) x (6/7) x (4/5) x (3/4) x (1/2) = 0.184

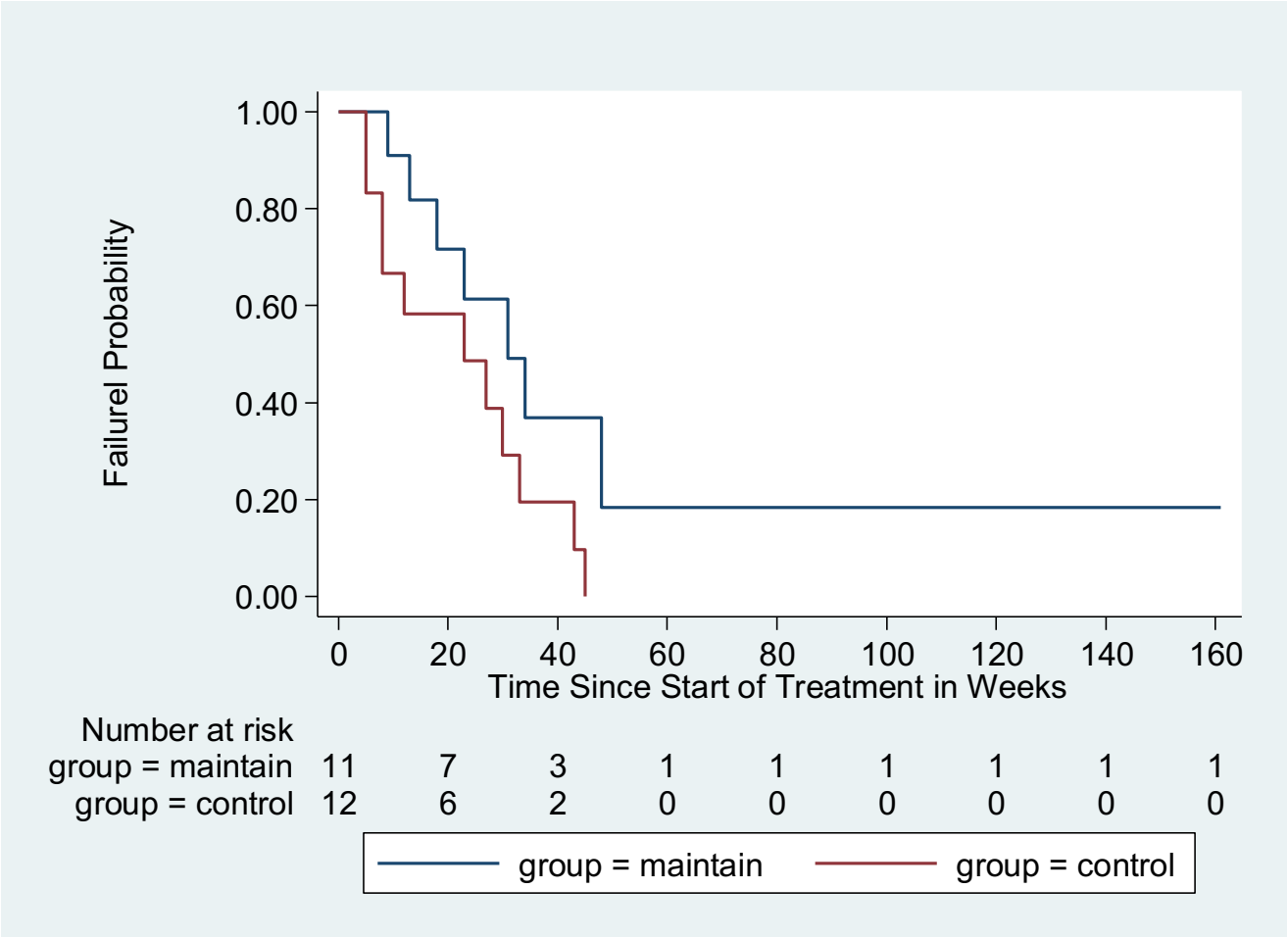
- K-M estimator is based on conditional probabilities (product limits)
- (relapses at weeks 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+)

Kaplan-Meier (with censoring)

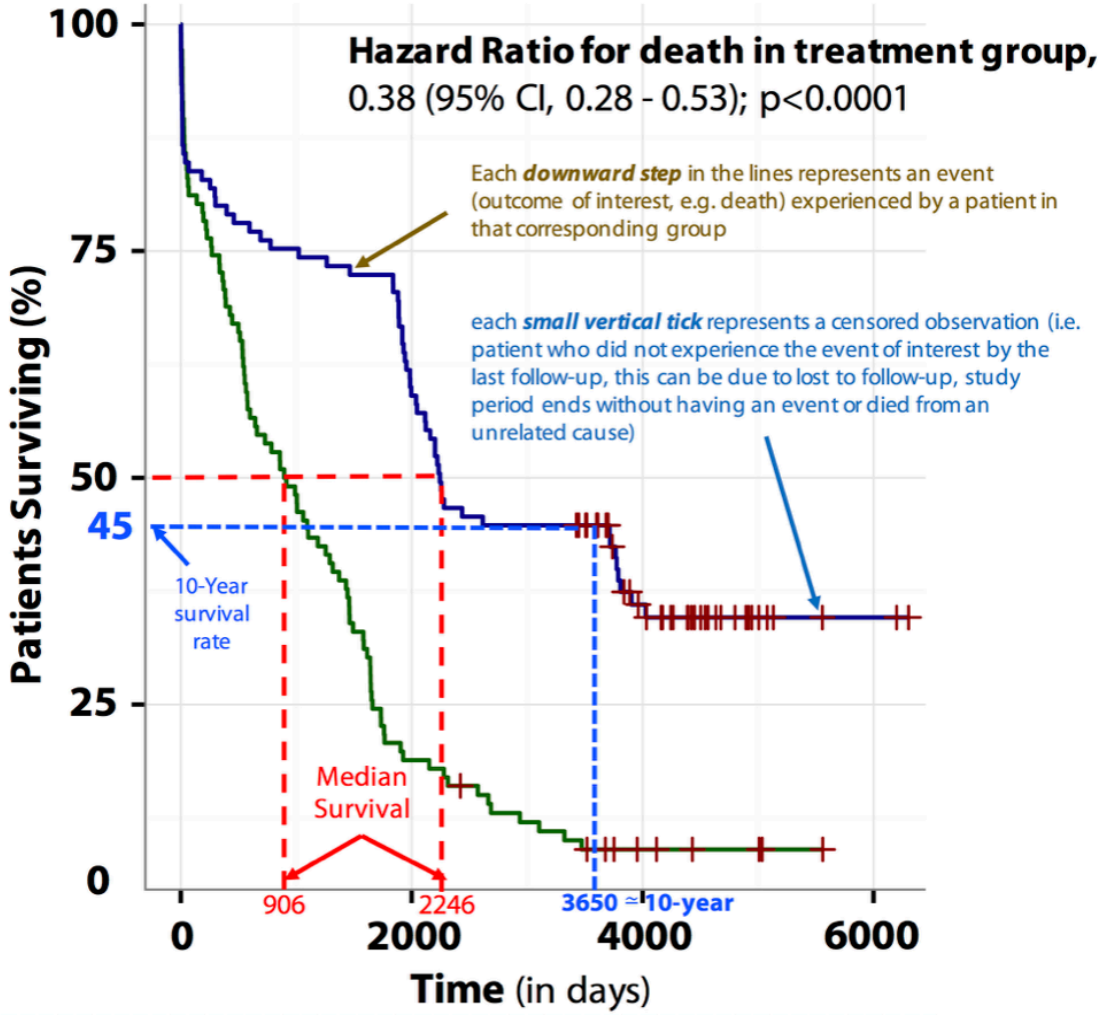
Key take aways:

- Censored observations did not get counted in the 'event' set (d_j) during the censoring interval
- Censored observations remained in the risk set for the current censoring interval, but were removed thereafter (n_j)
- In this way the censored observations don't inadvertently down-weight S(t)

Recall - Estimated Survival Curves (two groups)



Estimated Survival Curves (another example)



Next - Interval Estimation for Survival Probabilities

- The most popular method for obtaining interval estimates for survival probabilities $S(t_j)$ is to use *Greenwood's formula*

- Although survival probabilities must lie between 0 and 1, this is different than the calculation of confidence intervals for binomial proportions, since with survival data, the denominator changes over time as subjects drop out of the risk set

- Greenwood's method is based on calculating a confidence interval for $\log[S(t)]$ rather than $S(t)$ itself, since the natural log of the survival function is more closely normally distributed than the survival function itself

- We then exponentiate the lower and upper bounds to get a confidence interval for $S(t)$

Greenwood's Formula

Using the delta method, it can be shown that:

$$\text{var}\{\ln[\hat{S}(t)]\} = \sum_{\{j:t_j \le t\}} \frac{d_j}{n_j(n_j - d_j)}$$

Also, $\ln[\hat{S}(t)]$ tends to be more normally distributed than $\hat{S}(t)$.

Hence, a 100% x (1-α) CI for ln[S(t)] is given by:

$$(c_1, c_2) = \ln[\hat{S}(t)] \pm z_{1-\alpha/2}\sqrt{\text{var}\{\ln[\hat{S}(t)]\}}.$$

The corresponding 100% x (1-α) CI for S(t) is $[\exp(c_1), \exp(c_2)]$.

This is known as Greenwood's formula.

Confidence Interval for $S(t_k)$

- In the example, there are 3 survival times in the maintenance group that are ≤ 20 weeks: 9, 13, and 18 weeks, ie $\{j: t_j \leq t\}$, where t=20

- The variance of log[$S(t)$] will thus, by definition of Greenwood's formula, be summed over these 3 time points

Confidence Interval for $S(t_k)$

- $\hat{S}(20) = \hat{S}(18) = 0.716$ and $\ln[\hat{S}(20)] = -0.3342$.

  Also, from Greenwood's formula, since there were $n_i = 11$ patients in the risk set at 9 weeks, 10 patients at 13 weeks and 8 patients at 18 weeks, and $d_i = 1$ patient failed(relapsed) at each of these time points, we have:

$$\text{var}[\ln(\hat{S}(20)] = \frac{1}{11(10)} + \frac{1}{10(9)} + \frac{1}{8(7)} = 0.0381.$$

$$se[\ln(\hat{S}(20)] = \sqrt{0.0381} = 0.1951.$$

Confidence Interval for $S(t_k)$

- Thus, a 95% CI for ln[S(20)] is given by :

  $-0.3342 \pm 1.96(0.1951) = (-0.7166, 0.0482) = (c_1, c_2)$.

  The corresponding 95% CI for S(20) is :

  [exp(-0.7166), exp(0.0482)] = (0.49,1.05).

  Since survival estimates cannot be greater than 1,

  we truncate the upper limit and obtain a 95% CI for S(20) of (0.49,1.00).

Confidence Interval for $S(t_k)$

- Since exp(x) can be > 1, it is possible that confidence intervals based on the natural log transformation will yield upper confidence limits for $S(t)$ that are > 1, as seen in the previous example

- An alternative approach is to use the *complementary log-log transformation* instead of the log transformation when constructing confidence intervals (see appendix)

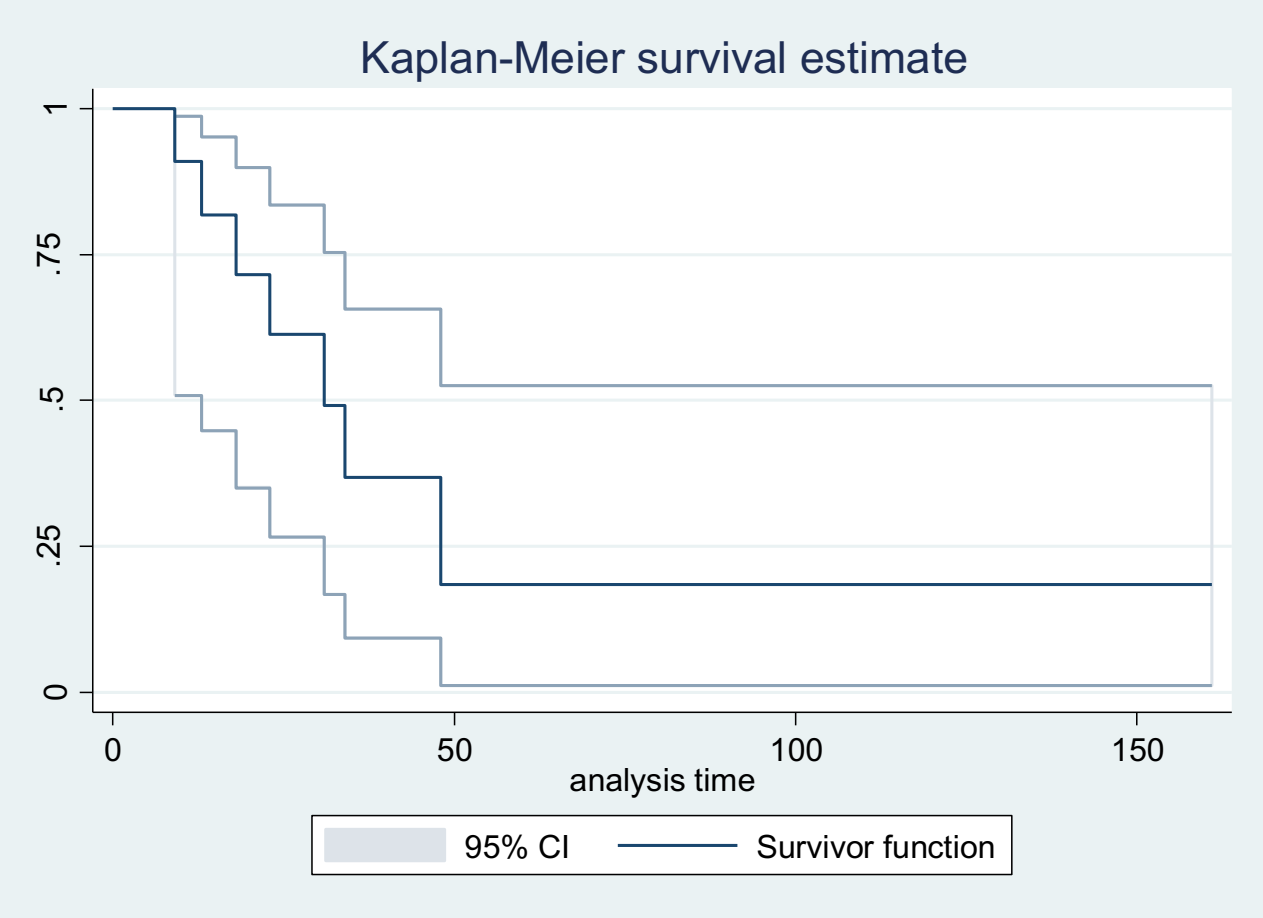
- Confidence intervals are estimated at each failure time point, and are assumed to remain constant until the next failure occurs

Confidence Bands for $S(t)$

- After calculating a confidence interval for $S(t_k)$ for each value of k, we can plot these intervals on a graph of the survival curve, connect all the lower bounds, and connect all the upper bounds

- This gives us *confidence bands* for the estimated survival curve

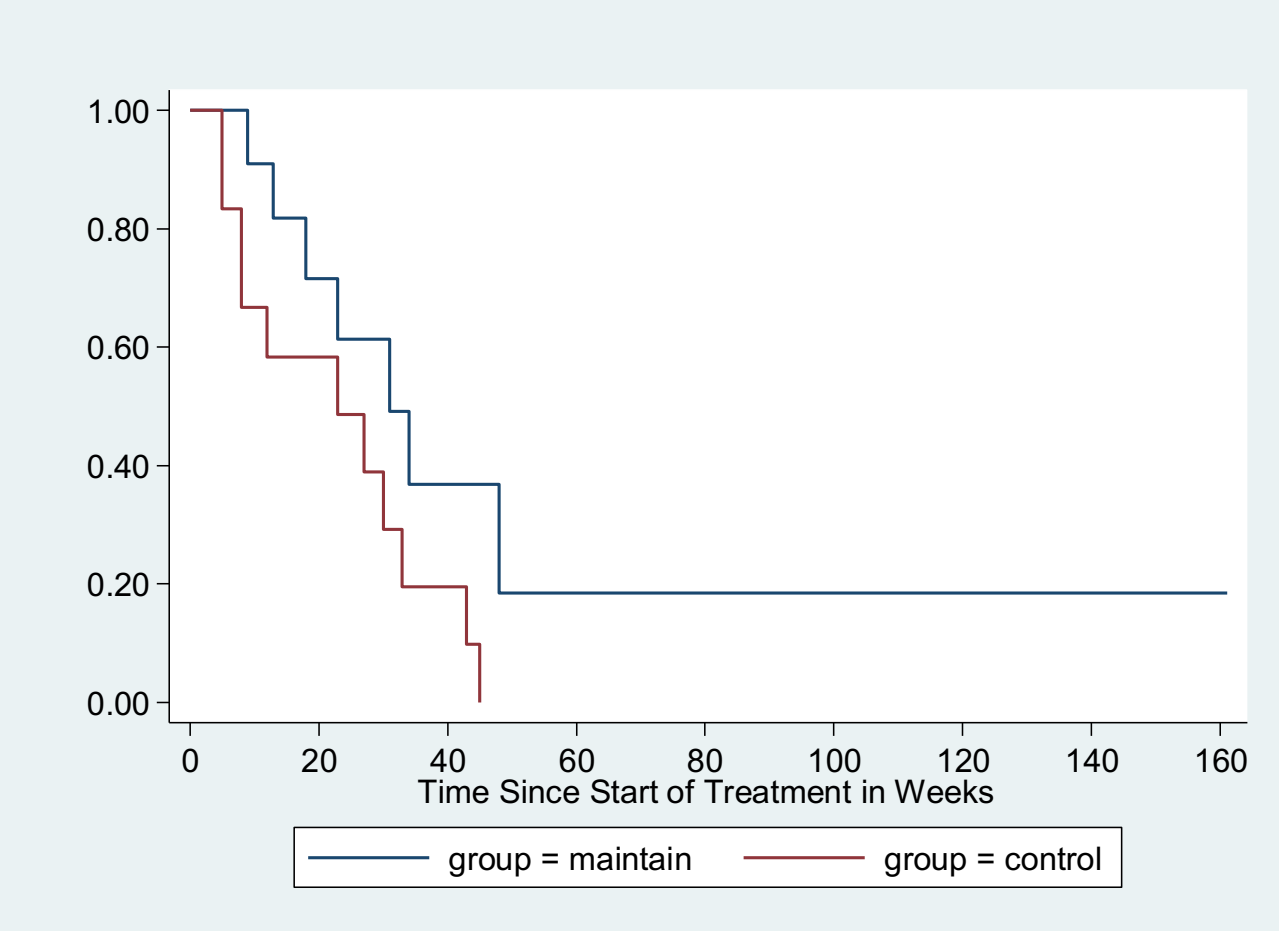
Confidence Bands for $S(t)$



Next - Comparing Survival Curves

- We often wish to compare survival in two independent groups (e.g., maintenance chemotherapy versus control)
- How might we do this?
- Should we compare survival at the time point with the largest difference between curves? Add up the differences across all times? (Weighted or not weighted?) Compare the median survival for each group?
- Let's have another look at our K-M estimates for each treatment group

Kaplan-Meier Survival Curves



Log-Rank Test

- To compare two survival curves, the most commonly used test is a generalization of the Mantel-Haenszel test for stratified 2×2 tables known as the *log-rank test*
- The log-rank test can be thought of as an application of the Mantel-Haenszel test to censored survival data

Log-Rank Test

- The log-rank test is used to test the hypothesis that two survival distributions are equal

- What are the null and alternative hypotheses?

H_0: $S_1(t) = S_2(t)$ for all t

H_1: $S_1(t) \neq S_2(t)$

- The alternative hypothesis means that the two survival functions differ for at least one value of t, but we cannot actually be more specific than this

Log-Rank Test

- To perform the test, we construct a 2 × 2 table of treatment group by relapse yes/no at each observed failure time

- At each failure time, we then calculate:

  (a) the observed number of failures in group 1

  (b) the expected number of failures in group 1 under H_0 (if the null hypothesis is true)

  (c) the variance of the number of failures in group 1 under H_0

Log-Rank Test

- We sum the observed failures, the expected failures, and the variances over all failure times, and then construct a test statistic similar to the Mantel-Haenszel test statistic

- So, suppose there are K distinct failure times that are ordered as follows:

$$t_1 < t_2 < \ldots < t_K$$

- At the i^{th} failure time t_i, we construct a 2 × 2 table

Log-Rank Test

	Failure		
Group	Yes	No	Total
Maintenance	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Control	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Log-Rank Test

- $$O = \sum_{i=1}^{K} O_i = \sum_{i=1}^{K} d_{1i} = \text{total observed failures in group 1,}$$

  $$\text{Under } H_0,$$

  $$E = \sum_{i=1}^{K} E_i = \sum_{i=1}^{K} n_{1i} d_i / n_i = \text{total expected failures in group 1,}$$

  $$V = \sum_{i=1}^{K} V_i = \sum_{i=1}^{K} \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)} = \text{ variance of O.}$$

- All margins are considered fixed at each failure time
- Therefore, d_{1i} is a random variable that follows a hypergeometric distribution

Log-Rank Test

- The test statistic for the log-rank test is

$$X^2 = \frac{(O - E)2}{V} \sim \chi_1^2$$

if the null hypothesis is true

Example: Chemotherapy for Leukemia

- The first failure (relapse) occurred at 5 weeks.
- *Maintenance group*:

  11 subjects in the risk set and 0 failures at 5 weeks

- *Control group*:

  12 subjects in the risk set and 2 failures at 5 weeks

Example: 2 x 2 Table at 5 Weeks

	Failure		
Group	Yes	No	Total
Maintenance	0	11	11
Control	2	10	12
Total	2	21	23

Example: 2 x 2 Table at 5 Weeks

- At 5 weeks:

$$O_1 = 0,$$

$$E_1 = \frac{11(2)}{23} = 0.96,$$

$$V_1 = \frac{11(12)(2)(21)}{23^2(22)} = 0.476.$$

- For another 2x2 table at 8 weeks, and ultimately the overall set of survival probabilities S(t), see appendix

Example: Chemotherapy for Leukemia

- In summary, calculating the test statistic using tables for every relapse week we get: O = 7, E = 10.69, and V = 4.008

→ $X^2 = (7 - 10.69)^2 / 4.008 = 3.40$

- Under the null hypothesis, this test statistic has a χ_1^2 distribution and $p = 0.065$

- (Strictly speaking, V should be > 5 to use this test, based on a "rule of thumb")

Example: Chemotherapy for Leukemia

- We do not reject the null hypothesis that the two survival curves are equal (although this is a borderline result)

- The data does not provide evidence of a difference in the distribution of relapse times (but keep in mind that this is only a pilot study and sample sizes are small – with a larger sample size, we might have rejected H_0)

Summary - Log-Rank Test

- The log-rank test is based on an ordering (ranking) of event times, and is therefore a nonparametric test

- This test is most powerful when the hazard functions (we'll get to that next!) are proportional between the two groups being compared, meaning that for constant c,

$$h_2(t) / h_1(t) = c$$

- This does not mean the test is invalid for non-proportional hazards, just less powerful

- Method can be extended to > 2 groups (see appendix)

Next - Regression Models for Survival Data?

- We would like to compare survival curves between two or more groups after controlling for one or more covariates
- We'd also like to be able to assess confounding and effect modification, and perform model selection and model assessment efforts
- We can accomplish this using a regression model for survival data
- This is analogous to logistic regression, except that the time when an event occurs (not just whether an event occurs) is taken into account, and there could be censoring
- There are both parametric (specific distributional assumptions) and nonparametric (or semiparametric) regression models for survival data

Recall - Functions defining time T

- Previously we presented 3 functions of T:

$S(t) = P(T > t)$ (the survival function)
$= 1 - P(t \leq T)$
$= 1 - F(t)$
$= \int_t^{\infty} f(x)dx$

- where

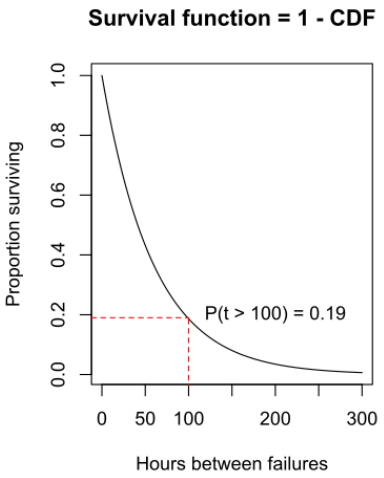
$f(t) = F'(t) = -S'(t)$

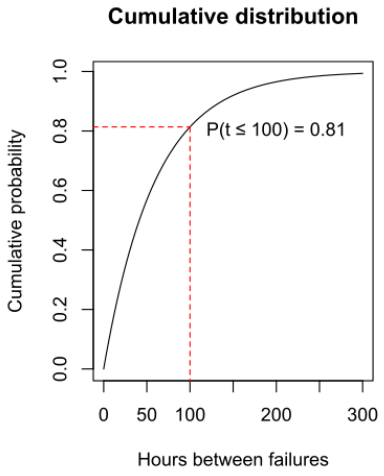
is the probability density function, and

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

is the cumulative distribution function

- *So what's the 4th function of T?*





Hazard Function

- The *hazard function* (or hazard rate) is the instantaneous rate of failure at time *t*, given that a subject has survived to time *t.* In other words, the hazard function is the short-term event rate for subjects who have not yet experienced the outcome event

- Mathematically, it can be defined by:

$$h(t) = \lim_{\Delta t \to 0} [\Pr(t \le T \le t + \Delta t) / \Delta t \,|\, T \ge t]$$

- Note that $h(t) \ge 0$ and has no upper bound; it is not a probability

- Slightly more morbid name:

"Force of Mortality" ☹

Hazard Function

- Let's show on the board what this result really means →
- Note also from earlier that F(t) is the cumulative distribution function

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

- And if we let $f(t)$ be the probability density function (for a continuous survival time), where $f(t) = F'(t) = -S'(t)$
- Then the hazard function can be expressed as

$$h(t) = f(t) / S(t) = f(t) / [1 - F(t)]$$

or

$$h(t) = \lim_{\Delta t \to 0} [\{S(t) - S(t + \Delta t)\} / \Delta t] / S(t)$$

Hazard versus Incidence Rates

- One aside: in our previous work on incidence rates (including Poisson regression), we assumed that the incidence rate (λ) remained constant over time

- This situation would be an example of a constant hazard rate

$$h(t) = \lambda \text{ for } t \geq 0$$

- The exponential survival model has a constant baseline hazard rate; the Weibull does not--> but we will get to these parametric survival distributions soon

- Before that we introduce…..

Cox Proportional Hazards Regression

Consider the situation where we wish to

- Adjust for one or more possible confounding variables or independent predictors
- Consider effect modification
- Select a good model when you have multiple covariates
- Assess the proportional hazards assumption with multiple covariates

Cox Proportional Hazards Model

- The general proportional hazards model is:

$$h(t|X_1 = x_1, ..., X_p = x_p) = h_0(t)\exp(\beta_1 x_1 + ... + \beta_p x_p)$$

which we can rewrite in the form:

$$\log[h(t|X_1 = x_1, ..., X_p = x_p)] = \log[h_0(t)] + \beta_1 x_1 + ... + \beta_p x_p,$$

where $h_0(t)$ is the baseline hazard function and the intercept term is included in $\log[h_0(t)]$.

Example: Chemotherapy for Leukemia

- Return to the leukemia treatment example that we discussed previously, having some censored observations
- We used the Kaplan-Meier (or product limit) method to estimate the survival curve S(t) in each group (maintenance chemotherapy versus control), where the failure event was relapse of leukemia
- We then used the log-rank test to compare the survival curves between groups

42

Example: Chemotherapy for Leukemia

- Times in weeks are:

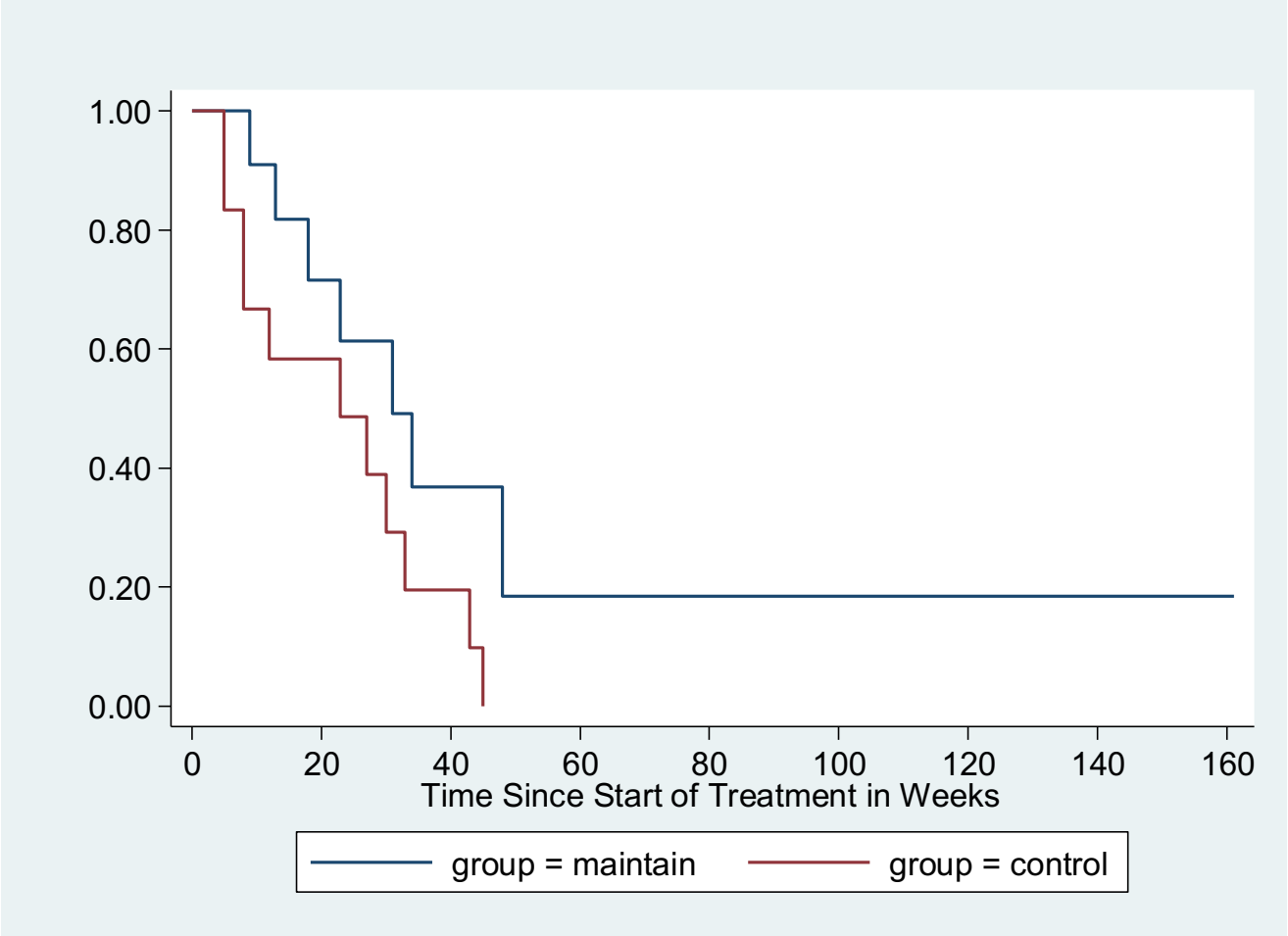
- Maintenance chemotherapy group (*n* = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (*n* = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Kaplan-Meier Survival Curves



Log rank test gave $P = 0.065$ comparing the two groups

Cox Proportional Hazards Model

- Alternatively, we could compare the survival curves using the Cox proportional hazards model given by:

$$h_i(t) = h_0(t)\exp(\beta\, x_i),$$

where

$x_i = 1,$ if subject i is in the maintained group,

$= 0,$ otherwise.

Interpretation of Model Parameters

- We notice that:

$$h(t|X=1) = h_0(t)\exp(\beta) \text{ if a patient is in the maintained group,}$$
$$\mathrm{h}(t|X=0) = h_0(t) \text{ if a subject is in the control group.}$$

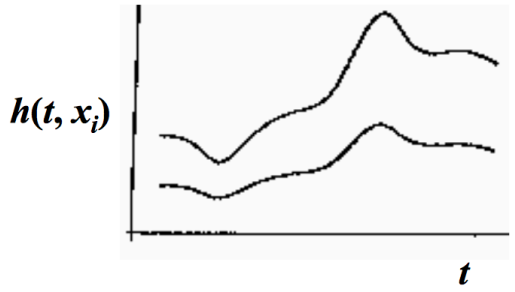
- No assumptions are made about the shape of the control group hazard function $h_0(t)$, which can vary over time in an arbitrary manner

Interpretation of Model Parameters

- Note that the hazard ratio (but not the hazard functions themselves) is assumed constant over time under a proportional hazards model

$$\frac{h(t|X=1)}{h(t|X=0)} = \exp(\beta) = \text{hazard ratio},$$

$$\text{or } \beta = \log(\text{hazard ratio}).$$



47

Cox Proportional Hazards Model

- Here β_i is the log hazard ratio associated with a one unit increase in x_i, holding all the other variables constant
- Also, $\exp(\beta_i)$ is the hazard ratio associated with a one unit increase in x_i, holding all the other variables constant
- Sounds and looks familiar!

Example: Chemotherapy for Leukemia

	Estimate	95% CI	*p* Value Wald Test
Hazard Ratio	0.405	(0.148, 1.105)	0.078
β Coefficient	–0.904	(–1.908, 0.100)	

- Here exp(–0.904) = 0.405 and log(0.405) = –0.904

Semiparametric Regression

- The Cox model is a *semiparametric* regression model
- The form of the baseline hazard $h_0(t)$ is not specified
- Because of this, it is impossible to write out the full likelihood, so maximum likelihood estimation is not possible

Partial Likelihood Estimation

- However, we can estimate the regression coefficients β_1, β_2, ..., β_p based on the *partial likelihood*, which is determined by the relative ordering of the survival times (i.e., their ranks) rather than their actual values

- The partial likelihood is constructed using the information at each event time, given the risk set of subjects at each distinct failure time t_k

Partial Likelihood Estimation

- Given that an event occurs at time t_k, we calculate the conditional probability that it is subject *i* in the risk set who fails at time t_k, given by:

$$PL(t_k) = \frac{h_0(t_k)\exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})\Delta t_k}{\sum_{j \text{ in } R(t_k)} h_0(t_k)\exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp})\Delta t_k},$$

- x_{jp} = value of the p^{th} covariate for the j^{th} subject in the risk set at time t_k, and $R(t_k)$ = risk set at time t_k

- We are starting with no tied survival outcomes

52

Partial Likelihood Estimation

- Note that the term $h_0(t_k) \times \Delta t_k$ appears in both the numerator and denominator of the partial likelihood, and thus cancels out

- Therefore we can rewrite $PL(t_k)$ in the form:

$$PL(t_k) = \frac{\exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{\sum_{j \text{ in } \mathrm{R}(t_k)} \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp})}$$

Partial Likelihood Estimation

- The overall partial likelihood is given by:

$$PL = \prod_{k=1}^{K} PL(t_k)$$

- We do not need to know the baseline hazard $h_0(t)$ to calculate the partial likelihood *PL*

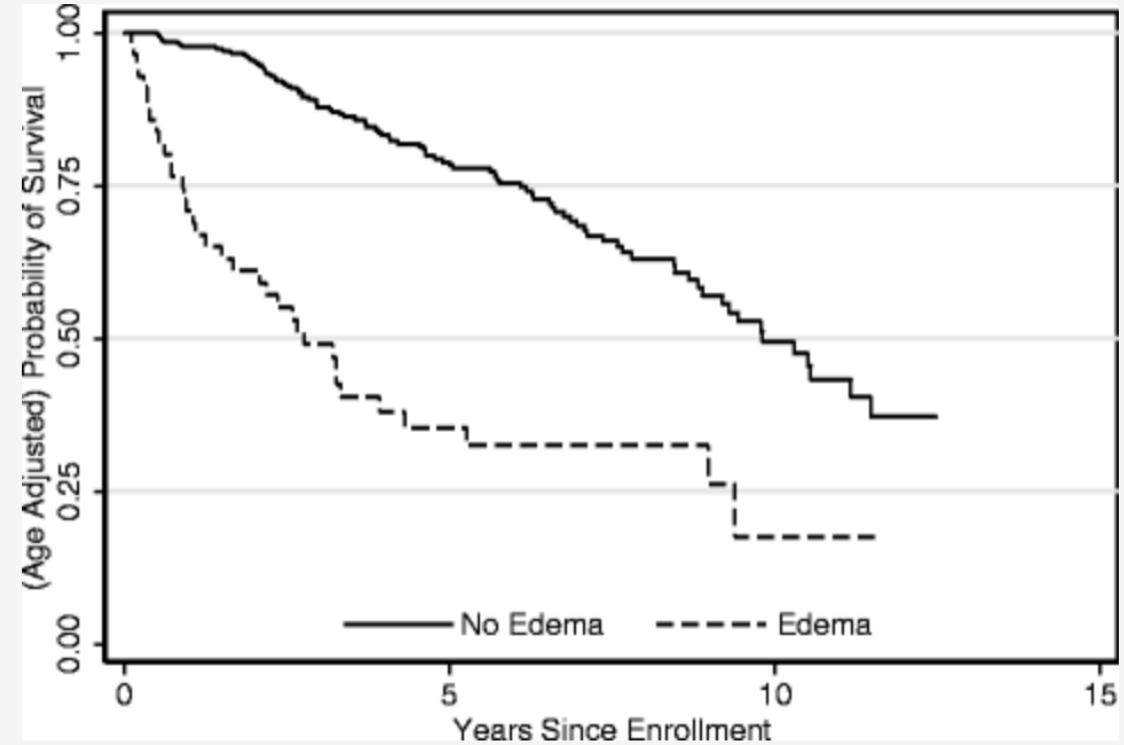
- We find $\beta_1, \ldots, \beta_p$ which maximize the partial likelihood *PL* by numerical iterative methods; the product is over the distinct failure times

What about the proportional hazards assumption?

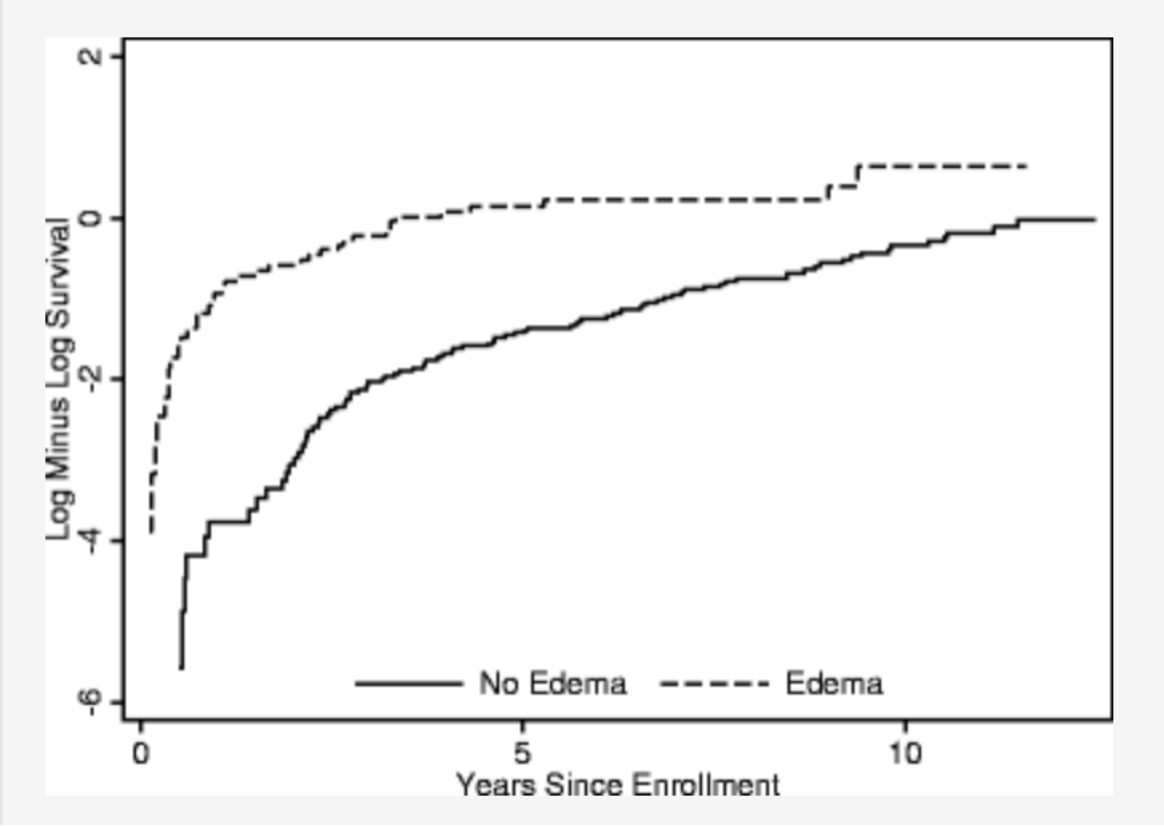
- An assumption of the Cox model is that the hazards of risk factors are proportional over time

- Specifically, when comparing two groups (e.g., with a binary covariate) and holding all the other covariates constant, the hazard ratio of exposed versus control is assumed to be the same across all times

Proportional?



Proportional?



Checking the Proportional Hazards Assumption

- Another helpful result regarding S(t) gives us:

$$S(t) = \exp[-\Lambda(t)], \text{ where}$$

$$\Lambda(t) = \text{cumulative hazard } = \int_{u=0}^{t} h(u)\, du.$$

If we denote the cumulative hazard for the maintained and control groups by $\Lambda_1(t)$ and $\Lambda_0(t)$, then under the proportional hazards model,

$$\Lambda_1(t) = \int_{u=0}^{t} h_0(u) \exp(\beta)\, du = \exp(\beta) \int_{u=0}^{t} h_0(u)\, du$$

$$= \exp(\beta)\Lambda_0(t).$$

58

Checking the Proportional Hazards Assumption

- Therefore,

$$S_1(t) = \exp[-\Lambda_1(t)] = \exp[-\Lambda_0(t)\exp(\beta)]$$

- We can take natural logs of both sides of this equation and multiply by –1, to get

$$-\log[S_1(t)] = -[-\Lambda_0(t)\exp(\beta)] = \Lambda_0(t)\exp(\beta)$$

59

Checking the Proportional Hazards Assumption

- Taking natural logs of both sides a second time, we get

$$\log[-\log[S_1(t)]] = \log[\Lambda_0(t)] + \beta$$

- Similarly, for the control group we would obtain

$$\log[-\log[S_0(t)]] = \log[\Lambda_0(t)]$$

- What separates these 2 equations?

60

Checking the Proportional Hazards Assumption

- If we plot the complementary log-log transformation of $S(t)$ versus $\log(t)$ for each group, the graphs should be roughly parallel and separated by a constant β (some softwares plot this with one more multiplication by -1)
- This is often called a *log-log plot*
- It is used to *visually evaluate* the assumption of proportional hazards

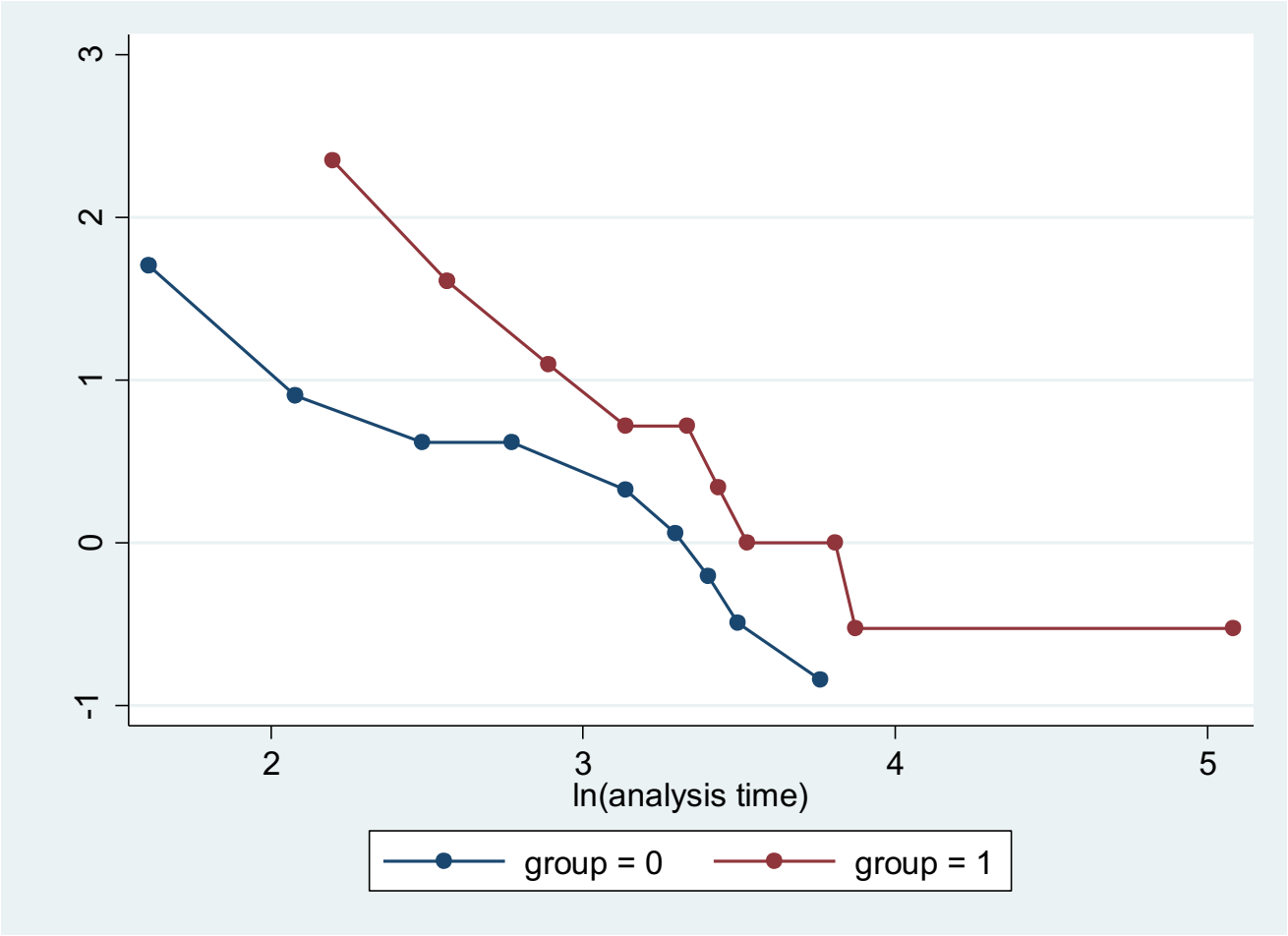
61

Checking the Proportional Hazards Assumption

In Stata, use the following command:

```
. stphplot, by group
```

Checking the Proportional Hazards Assumption (Leukemia Example)



Checking the Proportional Hazards Assumption

- Could use the following:

```
. stphplot, by(group)
```

- We can also assess the PH assumption graphically *adjusting for other factors*, using:

```
. stphplot, by(group) adjust(age sex)
```

Checking the Proportional Hazard Assumption

- The lines being roughly parallel is consistent with the proportional hazards assumption, though this is only a visual check
- The average distance between the two curves should be log(hazard ratio) = β (comparing exposed with $x = 1$ vs. control with $x = 0$), taking absolute values
- If a violation of the proportional hazards assumption is found, perhaps separate analyses should be performed for different time periods
- If the Kaplan-Meier survival curves cross each other, then the hazards are *not* proportional

Test of Proportional Hazards Assumption

- To further test the proportional hazards assumption (for the first covariate, x_1), we can enhance the previous proportional hazards model as follows:

$$\log(h(t|X)) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \log(t)(\gamma_1 x_1)$$

- The log hazard ratio for a one unit increase in x_1 at time t is given by $\beta_1 + \gamma_1 \log(t)$, which can be exponentiated to estimate the HR

Test of Proportional Hazards Assumption

- We can test multiple factors at one time using this approach
- Note that the main effects (based on the β terms) are not directly interpretable alone when the interaction terms with log(t) (the γ terms) are included
- But we can formally test whether the γ terms are needed or not (i.e., get a p-value)
- Choosing log(t) is commonly done, though other parameterizations of time are possible

Checking Proportional Hazards

- To fit this model, we use the **tvc** and **texp** options of the **stcox** command of Stata, where

tvc is a set of possibly time-dependent covariates, and

texp is a function of time

- Note that Stata creates a variable _t to represent follow-up time --- regardless of the user-specified variable for time (e.g., log(*t*))

Extensions

Survival analysis methods can be extended in a number of directions:

- More on parametric survival models (exponential, Weibull, gamma, or other models, with covariates, and assessment of fit)
- Time-varying covariates (allowing the covariates, and hence the hazard, to change over time)
- Multiple (recurrent) events per subject (e.g., time to next migraine)
- Accelerated failure time models
- Competing risks (having more than one competing outcome possible)

Appendix

Confidence Interval for $S(t_k)$

- The complementary log-log transformation is given by:

$$y(t) = \log\{-\log[S(t)]\}$$

- The advantage of this transformation is that if we solve for $S(t)$ as a function of $y(t)$, we obtain:

$$S(t) = e^{-e^{y(t)}}, \text{where} -\infty < y(t) < \infty.$$

Confidence Interval for $S(t_k)$

- Since $S(t) = 1$ when $y(t) = -\infty$, and $S(t) = 0$, when $y(t) = \infty$, it implies that if we obtain confidence limits for $y(t)$ and transform back to the $S(t)$ scale, the corresponding confidence limits for $S(t)$ will always be between 0 and 1.

- The variance formula is more complicated than for the log transformation

Confidence Interval for $S(t_k)$

Using the delta method to estimate the variance, a 100% x (1-α) confidence interval for $y(t)$ is given by

$$y(t) \pm z_{1-\alpha/2} \sqrt{\frac{\sum_{\{j:t_j \le t\}} \frac{d_j}{n_j(n_j - d_j)}}{\left[\sum_{\{j:t_j \le t\}} \ln\left(\frac{n_j - d_j}{n_j}\right)\right]^2}} = [y_1(t), y_2(t)].$$

The corresponding 100% x (1-α) CI for S(t) is given by :

$$[e^{-e^{y_2(t)}}, e^{-e^{y_1(t)}}].$$

Using Stata to Estimate CIs for $S(t_k)$

Using this approach Stata obtains a 95% confidence interval for $S(20)$ in the maintenance group of (0.35, 0.90)

Confidence intervals are estimated at each failure time point, and are assumed to remain constant until the next failure occurs

Using Stata to Estimate CIs for $S(t_k)$

Stata reports the standard error from Greenwood's formula, but the 95% confidence interval is based on the complementary log-log transformation

With large sample sizes, confidence intervals based on $\log[S(t)]$ and $\log\{-\log[S(t)]\}$ are generally similar

With small sample sizes, the complementary log-log approach is preferable

Log-Rank Test

n_{1i} = number of maintenance subjects in the risk set at time t_i

n_{2i} = number of control subjects in the risk set at t_i

n_i = total number of subjects in the risk set at time t_i

d_{1i} = number of failures in the maintenance group at time t_i

d_{2i} = number of failures in the control group at t_i

d_i = number of failures in both groups combined at time t_i

Example: 2 x 2 Table at 8 Weeks

The next failure (relapse) occurred at 8 weeks

Maintenance group:

11 people in the risk set and 0 failures at 8 weeks

Control group:

10 people in the risk set and 2 failures at 8 weeks

Example: 2 x 2 Table at 8 Weeks

	Failure		
Group	Yes	No	Total
Maintenance	0	11	11
Control	2	8	10
total	2	19	21

Example: 2 x 2 Table at 8 Weeks

At 8 weeks,

$$O_2 = 0,$$

$$E_2 = \frac{11(2)}{21} = 1.05,$$

$$V_2 = \frac{11(10)(2)(19)}{21^2(20)} = 0.474.$$

Comparison of Survival Curves in the Leukemia Example

t_j	d_1j	n_1j	d_2j	n_2j	d_j	n_j	E_1j	V_1j
5	0	11	2	12	2	23	0.96	0.476
8	0	11	2	10	2	21	1.05	0.474
9	1	11	0	8	1	19	0.58	0.244
12	0	10	1	8	1	18	0.56	0.247
13	1	10	0	7	1	17	0.59	0.242
18	1	8	0	6	1	14	0.57	0.245
23	1	7	1	6	2	13	1.08	0.456
27	0	6	1	5	1	11	0.55	0.248
30	0	5	1	4	1	9	0.56	0.247
31	1	5	0	3	1	8	0.63	0.234
33	0	4	1	3	1	7	0.57	0.245
34	1	4	0	2	1	6	0.67	0.222
43	0	3	1	2	1	5	0.60	0.240
45	0	3	1	1	1	4	0.75	0.188
48	1	2	0	0	1	2	1.00	0.000
Tot	7						10.69	4.008

Log-Rank Test

- With the standard version of the Mantel-Haenszel test (for proportions), we consider the *K* strata to have come from a confounding variable(s) with *K* different levels
- The strata are independent of each other
- For the log-rank test, each successive table consists of a subset of subjects at risk from previous tables, so the strata are *not* independent

Log-Rank Test

To derive the distribution of d_{1i}, the number of failures in group 1 in the i^{th} stratum, we condition on:

(a) the total number of subjects in the risk set in each group

(b) the total number of failures in the two groups combined

Hence the d_{1i} are independent across strata, since subjects who previously failed are no longer in the risk set

Log-Rank Test

The log-rank test weights each event failure time equally in its test statistic

Other tests have been proposed which assign different weights to the failure times

For example, the Gehan-Wilcoxon test weights the data from each 2 × 2 table according to the total number of subjects in the risk set at that time

Alternative Tests for Survival Curves

Therefore, the Gehan-Wilcoxon test is more sensitive to early differences in survival, and less sensitive to late differences compared to the log-rank test. The Gehan-Wilcoxon test is an extension of the Wilcoxon rank sum test to possibly censored observations.

Other tests (e.g., Peto-Prentice, Fleming-Harrington, Tarone-Ware) assign different weights

Log-Rank Test

The log-rank test (and some of the others) can be used to compare survival distributions *for more than two groups*. If the null hypothesis of equality is rejected, then at least two of the survival functions are different. We must perform pairwise comparisons to see where the differences lie.

In addition, *stratified versions* of log-rank tests can be performed, to adjust for strata. An alternative is to use proportional hazards regression.

Accounting for Tied Event Times

The partial likelihood uses information about the relative ordering of event times to obtain estimates of the β coefficients

This is ambiguous in the case of tied event times

Several different approaches to this issue have been proposed

Accounting for Tied Event Times

Tied event times occur frequently in survival data, especially when time is recorded to the nearest week, month, or year

There were 2 subjects in the control group who failed at 5 weeks, and 2 subjects who failed at 8 weeks

If there were no tied survival times, the *score test* of the binary covariate in the Cox model and the log-rank test would provide identical results (in terms of *P* values). Also, the score test is *asymptotically equivalent* to the Wald and LR tests.

Treatment of Tied Event Times

Suppose there are 2 subjects (i_1, i_2) who both fail at time t_k

The *exact partial likelihood* at time t_k is given by:

$$PL(t_k) = \Pr(\text{subjects } i_1 \text{ and } i_2 \text{ fail at time } t_k \mid \text{exactly 2 subjects fail at time } t_k)$$

$$= \frac{\exp(\beta_1 x_{i_1 1} + \ldots + \beta_p x_{i_1 p}) \exp(\beta_1 x_{i_2 1} + \ldots + \beta_p x_{i_2 p})}{\sum_{\{j_1, j_2 \text{ in } R(t_k)\}} \exp(\beta_1 x_{j_1 1} + \ldots + \beta_p x_{j_1 p}) \exp(\beta_1 x_{j_2 1} + \ldots + \beta_p x_{j_2 p})}$$

88

Treatment of Tied Event Times

The sum in the denominator is over all combinations of two people in the risk set R_k, who might have failed at time t_k

The number of terms in the denominator of PL(t_k) is $n_k(n_k - 1)/2$, where n_k = size of the risk set at time t_k

Treatment of Tied Event Times

The problem is that the number of terms in the denominator will be large if n_k is large, especially if there are many tied event times – it is very computationally intensive. This method is sometimes called using the exact partial likelihood.

90

Breslow Treatment of Ties

The default option for handling tied event times in many statistical packages is called the *Breslow* option

The assumption with the Breslow method is that risk sets are generally large and the effect of using a finer categorization of time to "break the ties" is small

Therefore, under the Breslow approach, the same risk set is used for all events occurring at time t_k

Breslow Treatment of Ties

Specifically, $PL_{Breslow}$ is given by:

$$PL_{Breslow} = \frac{\exp(\beta_1 x_{i_1 1} + \ldots + \beta_p x_{i_1 p}) \exp(\beta_1 x_{i_2 1} + \ldots + \beta_p x_{i_2 p})}{\left[\sum_{\{j \text{ in } R_k\}} \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp}) \right]^2}$$

This approach is computationally faster, but may be more inaccurate if there are many tied event times (e.g., > two tied times). Notice that the numerator is the same as for the exact partial likelihood contribution, but the denominator contains n_k^2 terms.

92

Efron Treatment of Ties

The *Efron* approach assumes that if we break the ties, then the risk set for the first failure at time t_k is R_k

For the second failure time, there is a 50% probability that subject i_1 will be in the risk set, and a 50% probability that subject i_2 will be in the risk set

Efron Treatment of Ties

$$PL_{Efron} = A/(BC), \text{where}$$

$$\mathrm{A} = \exp(\beta_1 x_{i_1 1} + \ldots + \beta_p x_{i_1 p})\exp(\beta_1 x_{i_2 1} + \ldots + \beta_p x_{i_2 p}),$$

$$B = \sum_{\mathrm{j\ in\ R_k}} \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp}),$$

$$C = \sum_{j \text{ in } \mathrm{R_k}, j \neq i_1, j \neq i_2} \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp})$$

$$+ \frac{\exp(\beta_1 x_{i_1 1} + \ldots + \beta_p x_{i_1 p}) + \exp(\beta_1 x_{i_2 1} + \ldots + \beta_p x_{i_2 p})}{2}$$

This method takes longer to compute than the Breslow method, but is more accurate

94

Example: Chemotherapy for Leukemia

Method	Hazard Ratio	95% CI	*p* Value Wald Test
Breslow	0.405	(0.148, 1.105)	0.078
Efron	0.400	(0.147, 1.092)	0.074
Exact	0.398	(0.145, 1.094)	0.074

Example: Chemotherapy for Leukemia

There are only minor differences when using different methods for handling ties with this data set

Differences may be greater in data sets with more tied event times, especially large numbers of ties as the same distinct event time

Kaplan-Meier Estimate

$$\hat{S}(t) = 1 \; for \; 0 \le t < t_1,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in \; risk \; set \; at \; time \; t_1)$$

$$= 1 - d_1 / n_1 = \hat{q}_1 \; for \; t_1 \le t < t_2,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in \; risk \; set \; at \; time \; t_1) \; \text{x} \; \hat{\Pr}(T > t_2 \,|\, in \; risk \; set \; at \; time \; t_2)$$

$$= (1 - d_1 / n_1)(1 - d_2 / n_2) = \prod_{j=1}^{2} (1 - d_j / n_j) \; for \; t_2 \le t < t_3.$$

In general, the survival function is estimated as:

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - d_j / n_j) = \prod_{j=1}^{k} \hat{q}_j \; for \; t_k \le t < t_{k+1}.$$

Coming Up

- Parametric survival distributions and models

- Sample size and power for studies
- Basics of missing data analysis