

Biost 517

Applied Biostatistics I

.....

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 15: Two Sample Inference with Right Censored Data

November 22, 2006

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

Lecture Outline

.....

- Comparing Independent Proportions
 - Large Samples (Censored)
 - Using Kaplan-Meier Estimates
- Comparing Hazard Functions
 - Logrank Test
 - Wilcoxon Test for Censored Data
- Comparing Quantiles
 - Parametric Accelerated Failure Time Models

2

Right Censored Data

.....

3

Right Censored Data

.....

- Recall from Lecture 6: Censored variables
 - A special type of missing data (the exact value is not always known)
 - Right censoring: for some observations it is only known that the true value exceeds some threshold
 - Left censoring: for some observations it is only known that the true value is below some threshold
 - Interval censoring: for some observations it is only known that the true value is between some thresholds

4

Examples

.....

- PSA data set
 - Subjects were followed with serial PSAs
 - Interested in time to relapse
 - Some still in remission at time of analysis
 - (Ignoring these subjects is ignoring successes)
- University salary data set
 - Interest is in sex discrimination
 - Interested in time to promotion from associate
 - Some subjects have not yet been promoted
 - (Ignoring these subjects may be ignoring discrimination)

5

Descriptive Statistics

.....

- Sample mean, sample median (and other quantiles), sample standard deviation and variance are not appropriate
- Instead, descriptive statistics must be computed from Kaplan-Meier estimates
 - Only exception: You could use binomial proportions to estimate survival to the first censoring time
 - E.g., PSA data: All subjects followed at least 24 months

6

Noninformative Censoring

.....

- Recall: Our estimation methods only appropriate if censoring is not informative about subjects who were either more or less likely to have an event in the immediate future
 - Censored subjects must look like a random sample of those at risk at time of censoring
 - (Later we shall say that they are a random sample from all subjects at risk having similar modeled covariates)

7

Comparing Independent Proportions

.....

Large Samples with Right Censored Data

8

Kaplan-Meier Estimates

- Estimate $S(t) = Pr (T^0 > c)$ for arbitrary c
 - Nonparametric
 - Works for all distributions
 - (Also works for uncensored data)
 - Consistent for true value in infinite samples
 - Can derive estimates of quantiles
 - Can only estimate mean if estimated survival curve goes to 0
 - But can define “restricted mean” up to some time

9

Approximate Distribution

- If interested in $\theta = S(c) = Pr (T^0 \geq c)$ in presence of right censoring

Kaplan - Meier estimates for i th group

$$\theta_i = \hat{S}_i(c) = \prod_{j:t_j \leq c} \left(1 - \frac{d_{ij}}{n_{ij}} \right) \sim N \left(S_i(c), \left[se(\hat{S}_i(c)) \right]^2 \right)$$

(with $se(\hat{S}_i(c))$ from Greenwood's formula)

10

Stata: Kaplan-Meier Commands

- Syntax for “setting survival data”
 - `“stset endtime eventind,
t0(entrytime)”`
 - *endtime*: name of the variable measuring the time at the end of the interval
 - *eventind*: name of an indicator (0 or 1) variable indicating event status at the end of the interval
 - *entrytime*: name of the variable specifying the time at the start of the interval
 - (does not need to be supplied)
- `“stset, clear”` resets the data set

11

Stata: Kaplan-Meier Commands

- Syntax for getting estimates, plots
 - Plotting survival curves
 - `“sts graph”`
 - `“sts graph, atrisk”`
 - `“sts graph, cens(s)”`
 - Listing survival estimates
 - `“sts list”`
 - Saving survival estimates
 - `“sts gen newvar = s”`

12

Two Group Comparisons

.....

- To compare survival probabilities, we would compute SE for each group individually, then use methods for combining estimates

For independent $\hat{\theta}_1 \sim N(\theta_1, se_1^2)$, $\hat{\theta}_2 \sim N(\theta_2, se_2^2)$

$$\hat{\theta}_1 + \hat{\theta}_2 \sim N(\theta_1 + \theta_2, se_1^2 + se_2^2)$$

$$\hat{\theta}_1 - \hat{\theta}_2 \sim N(\theta_1 - \theta_2, se_1^2 + se_2^2)$$

$$\hat{\theta}_1 / \hat{\theta}_2 \sim N\left(\frac{\theta_1}{\theta_2}, \frac{1}{\theta_2^2} \left(se_1^2 + \frac{\theta_1^2}{\theta_2^2} se_2^2 \right)\right)$$

13

Two Group Inference

.....

- As with any (approximately) normally distributed estimator, CI and P values are computed using

100(1- α)% confidence interval

$$(est) \pm z_{1-\alpha/2} \times (std\ err)$$

Normalized Z statistic for $H_0 : \theta = \theta_0$

$$Z = \frac{(est) - (null\ hyp)}{(std\ err)}$$

14

Example: PSA Data

.....

- Men with prostate cancer
 - Hormonal treatment
 - Followed for signs of progression
- Interested in estimating probability of remaining in remission for three years
 - Testing hypothesis that three year survival differs between bone scan score less than 3 or bone scan score equals 3

15

Example: Preparing Data

.....

- Reading in data (note string variable)

```
. infile ... obstime str8 inrem using psa.txt
```
- Creating indicator of relapse

```
. g relapse = 0
. replace relapse = 1 if inrem=="no"
```
- "Setting" survival variables

```
. stset obstime relapse
```

16

Dichotomizing Bone Scan Score

- Method 1 (must consider missing data)

```
. g bss3= 0
. replace bss3=1 if bss==3
. replace bss3=. if bss==.
```
- Method 2 (recode handles missing data)

```
. g bss3= bss
. recode bss3 1/2=0 3=1
```

17

Stata: KM Listing

```
. sts list, by(bss3) at(12 24 36 48)
Beg.          Surv   Std.
Time Total Fail   Fctn  Error [95% Conf Int]
bss3=0
  12   18   1   0.9444 0.0540   0.6664 0.9920
  24   14   3   0.7778 0.0980   0.5110 0.9102
  36   12   1   0.7130 0.1092   0.4398 0.8699
  48    6   3   0.4801 0.1356   0.2101 0.7082
bss3=1
  12   22  10   0.6667 0.0861   0.4692 0.8047
  24   15   6   0.4667 0.0911   0.2839 0.6304
  36    9   5   0.2963 0.0841   0.1464 0.4630
  48    2   4   0.1058 0.0659   0.0209 0.2713
```

18

Stata: Difference and SE

- Three year survival probabilities
 - Bone scan score < 3: 0.7130 (SE 0.1092)
 - Bone scan score = 3: 0.2963 (SE 0.0841)
- Estimated diff in 3 year survival probability

```
. display 0.7130 - 0.2963
.4167
```
- Standard error of estimated difference

```
. display sqrt( 0.1092^2 + 0.0841^2 )
.13783124
```

19

Stata: 95% CI and P value

- 95% confidence interval: 0.147 to 0.687

```
. display invnorm(.975)
1.959964
. display .4167 - invnorm(.975) * 0.13783124
.14655573
. display .4167 + invnorm(.975) * 0.13783124
.68684427
```
- Two-sided P value : P = 0.0025
– (note use of negative)

```
. display 2 * norm( - .4167/.13783124 )
.00250065
```

20

Interpretation

.....

- The Kaplan-Meier estimate of difference in survival is that men with a bone scan score less than 3 have an absolute improved 3 year survival of 41.7% relative to $bss=3$
- With 95% confidence, such an observation is not consistent with a true absolute improvement less than 14.7% or greater than 68.7%
- Based on the P value of 0.0025, we reject the null hypothesis of no association between bone scan score and 3 year survival prob

21

Comparing Hazard Functions

.....

Logrank Test

22

Scientific Questions

.....

- With time to event data, we are generally interested in probability that an event will occur in a specified time
 - Right censored data presented problems, because the measurement of events was over varying amounts of time
 - Effect modification by time?
 - Confounding by time?
 - Increased precision by accounting for time?

23

General Strategy

.....

- We want to use methods that adjust for the time of observation
 - Kaplan-Meier estimates at a fixed time
 - Logrank and modified Wilcoxon statistics by averaging effects over time

24

Hazard Function

.....

- With censored data, we often compare probability distns using hazard functions
 - Hazard = Instantaneous risk of an event
 - Among subjects at risk of an event, what is the probability of having an event in the next instant
 - Advantage of using hazard with censored data
 - Only need to consider subjects currently at risk
 - Only need to consider whether they have an event right then

25

Hazard Function

.....

- Estimates of the hazard at each time look somewhat like a binomial proportion
 - We do not often estimate the hazard function over time
 - However, we do compare hazard functions
 - Usually we estimate a hazard ratio: relative risk of an event
 - We want to average the estimates of the hazard ratio over all times

26

Stratified Analyses

.....

- Recall that we are often interested in comparing groups within strata
 - Confounding:
 - Comparisons within strata are all similar, but failure to stratify results in a comparison that is misleading due to bias
 - There are nuances here as we go from analyses of means to analyses of nonlinear summary measures (e.g., odds- more later)
 - Interactions:
 - Comparisons within strata result in different estimates

27

Adjusting for Covariates

.....

- We can remove confounding by “adjusting” for the confounder using a stratified test statistic
 - “Adjustment” for a covariate means making comparisons between subjects who have similar levels of that covariate
 - E.g., in FEV data, compare smoking children to nonsmokers of same age, height
 - Average the differences seen in age, height strata

28

Role of Effect Modification

- Adjustment for a covariate does not remove interactions
 - Interactions means that the question has different answers in different strata
- Adjustment for a covariate will merely average the effect across strata
 - Usually weighted by the sample size in each stratum

29

Stratified Estimates

- Obtained by combining estimates from each (independent) stratum
 - Generally, best to average the estimates (sometimes weighted) rather than Z scores
 - SEs for the stratified estimates are obtained using properties of independent random variables
 - Standard errors are the sum of squared standard errors from the independent strata

30

Example

- Effect of hepatomegaly on survival after adjustment for sex?
 - Summarize response by 5 year survival
 - Hepatomegaly effect by sex: For each sex, compute difference in survival across hepatomegaly groups
 - Adjusted measure of effect: Compute the average difference between hepatomegaly effects
 - Usually a weighted average

31

SE for Stratified Estimates

	Males	Females
Hepatomegaly	$\hat{\theta}_{M1} \sim N(\theta_{M1}, se(\hat{\theta}_{M1}))$	$\hat{\theta}_{F1} \sim N(\theta_{F1}, se(\hat{\theta}_{F1}))$
No Hepatomegaly	$\hat{\theta}_{M0} \sim N(\theta_{M0}, se(\hat{\theta}_{M0}))$	$\hat{\theta}_{F0} \sim N(\theta_{F0}, se(\hat{\theta}_{F0}))$
	↓	
Weighted average	$p(\hat{\theta}_{M1} - \hat{\theta}_{M0}) - (1-p)(\hat{\theta}_{F1} - \hat{\theta}_{F0})$	
Approx Distn	$\sim N(\text{mean} = p(\theta_{M1} - \theta_{M0}) - (1-p)(\theta_{F1} - \theta_{F0}),$	
	$se = \sqrt{p^2(se^2(\hat{\theta}_{M1}) + se^2(\hat{\theta}_{M0})) + (1-p)^2(se^2(\hat{\theta}_{F1}) + se^2(\hat{\theta}_{F0}))}$	

Mantel-Haenszel Statistic

.....

- Generally regarded as the best choice of methods for comparing binary data across strata
 - Based on the odds rather than the proportion
 - It is rare that we might expect the difference in proportions to be constant across strata
 - Other methods can be based on the asymptotic distribution of the log odds
 - (More on these methods next quarter)

33

Logrank Test

.....

- The Mantel-Haenszel test is also the basis for a very popular method of comparing censored survival data across populations: The logrank statistic
 - The data are stratified by time of event
 - Often only a single event is observed in each stratum
 - Stratified estimates of the odds ratio are obtained

34

Noninformative Censoring

.....

- Most often the same subjects are used in several different strata
 - Noninformative censoring argues that the estimates are independent across strata asymptotically

35

Tests Equality of Hazards

.....

- Equal hazard functions implies equal distributions
 - The P value for this test is interpretable as a test that the survival distributions are similar for the two groups
 - This test is more powerful when the true alternative is “proportional hazards”
 - Proportional hazards = constant risk ratio over time
 - Proportional hazards regression will provide estimates of the risk ratio

36

Logrank Test: Stata Commands

- The logrank test can be obtained from Stata using the “sts test” command (after defining survival variables using “stset”
 - “sts test *groupvar*, logrank”
 - *groupvar* indicates the groups to be compared
 - logrank test is default
 - P value based on a chi square statistic
 - Hence a two-sided P value
 - (Obtaining a one-sided P value is deferred until we discuss proportional hazards regression next quarter)

37

Example: PSA Survival by bss

```
. sts test bss3
Log-rank test for equality of survivor functions
```

	Events	Events
bss3	observed	expected
0	9	17.18
1	25	16.82
Total	34	34.00

```
chi2(1) =      8.30
Pr>chi2 =     0.0040
```

38

Example: Interpretation

- Based on the two-sided P value of 0.004, we reject the null hypothesis of equal relapse free survival probabilities between the bone scan score groups
 - (Because the expected events are less than observed in the bss=3 group, we can presume that the higher bss is associated with worse relapse free survival)

39

Hazard Ratio Estimates

- Logrank test does not give estimates
 - However, it is closely related to “proportional hazards regression” (“Cox regression”)
 - Provides estimates of the (average) hazard ratio over time
- Hazard ratio
 - Groups with higher hazards have higher event rates
 - Hazard ratio greater than 1 = Worse “survival”

40

Proportional Hazard Regression

.....

- HR estimates approximately normal in large samples
- Stata commands
 - “`stcox groupvar, robust`”
 - “robust” eliminates need for proportional hazards
 - Gives hazard ratio, 95% CI
 - CI is computed on log hazard ratio scale
 - P values
 - “Wald test” (based on approximately normal estimate)
 - “Likelihood ratio test”
 - (“Score test” would be the logrank test)

41

Example: PSA Survival by bss

.....

```
. stcox bss3
No. of subjects = 48      Number of obs =      48
No. of failures = 34      Time at risk =     1408
                               LR chi2(1) =      8.35
Log likelihood = -106.9   Prob > chi2 = 0.0038
```

Robust						
t	HazRat	StdErr	z	P> z	[95% CI]	
bss3	2.96	1.11	2.89	0.004	1.42	6.16

42

Example: Interpretation

.....

- We estimate that at any given time the risk of relapse in men with bss=3 tends to be 2.96 times that of men with lower bss
- 95% CI suggests these results typical if true risk of relapse with bss=3 is 1.42 to 6.16 times that in men with lower bss
- Based on P value of 0.004 we would reject null hypothesis of no association between relapse and bss

43

Comparing Hazard Functions

.....

Wilcoxon Form of Logrank Test

44

Modification of Wilcoxon Test

.....

- Recall that the Wilcoxon test compares distributions based on $\Pr(Y > X)$
 - We need to define what we mean by $Y > X$ in presence of censoring
 - $Y > X$ if
 - uncensored $Y >$ uncensored X
 - censored $Y >$ uncensored X
 - Regard as unknown (and omit from analysis)
 - censored $Y <$ uncensored X
 - Y and X both censored

45

Wilcoxon Test Distribution

.....

- The modified Wilcoxon statistic can be shown to be asymptotically normally distributed
 - The standard errors for the modified Wilcoxon test under the null hypothesis can be computed from permutation distributions
 - Hence, a test of equality of the entire distribution

46

Other Interpretations

.....

- The modified Wilcoxon statistic can also be viewed as a weighted logrank statistic
- A weighted average of difference in hazards
- Places greater weight on differences in the survival curve that appear “early”
- Other ways to weight logrank statistics also exist
 - Logrank test is best if hazard ratio is constant over time

47

Stata Commands

.....

- The Wilcoxon test for censored data can be obtained from Stata using the “sts test” command (after defining survival variables using “stset”
 - `sts test groupvar, wilcoxon`
 - groupvar indicates the groups to be compared
 - P value based on chi square statistic
 - Hence a two-sided P value

48

Comparing Quantiles

.....

Parametric Models for Censored Data

49

Parametric Models forCensored Data.....

- There are times that inference for censored data is based on parametric models
 - Accelerated failure time models
 - Assume a constant ratio between groups for all quantiles of survivor distribution
 - E.g., dogs live 7 years for each year of human life

50

Parametric Models forCensored Data.....

- Commonly used parametric models
 - Exponential:
 - Constant hazard independent of past
 - Weibull:
 - Theoretical derivation: First failure in a series of components (weakest link in a chain)
 - Log hazard is linear
 - Exponential is special case
 - Only accelerated failure time model that is also proportional hazards

51

Parametric Models forCensored Data.....

- Commonly used parametric models (cont.)
 - Gamma:
 - Theoretical derivation: Final failure in parallel components
 - Exponential is special case
 - Lognormal
 - Many other generalizations

52

Caveats

.....

Choice of Summary Measures

53

Parametric Models

-
- All of the parametric models will be sensitive to violation of the distributional assumptions
 - Because these models assume constant ratio of all quantiles, we do not have robustness to other distributions in any particular model (including lognormal)
 - (We will discuss these models with regression next quarter)

54

Semiparametric Models

-
- We do know how to use the proportional hazards model, even when the hazard ratio is not constant
 - However, you need to be careful– it may not estimate anything you care about

55

Hypothetical Example: Setting

-
- Consider survival with a particular treatment used in renal dialysis patients
 - Extract data from registry of dialysis patients
 - To ensure quality, only use data after 1995
 - Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
 - Prevalent cases in 1995: Data from 1995 - 2002
 - » Incident in 1994: Information about 2nd – 9th year
 - » Incident in 1993: Information about 3rd – 10th year
 - » ...
 - » Incident in 1988: Information about 8th – 15th year

56

Hypothetical Example: Analysis

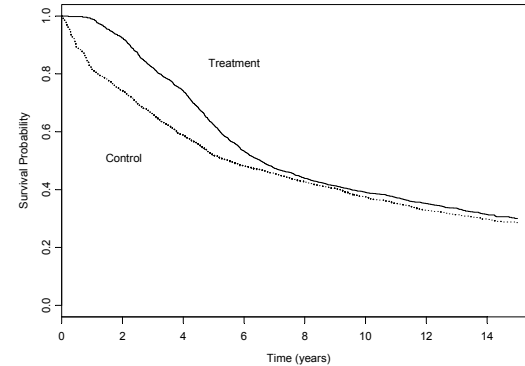
- Choice of summary measure
 - Survival at fixed point in time
 - Median, other quantiles
 - Mean (or restricted mean)
 - Hazard ratio (or weighted average of hazard ratio over time)
- Choice of methods
 - Parametric, semiparametric, nonparametric

57

Hypothetical Example: KM

.....

Kaplan-Meier Curves for Simulated Data (n=5623)



58

Who Wants To Be A Millionaire?.....

Proportional hazards analysis estimates a

Treatment : Control hazard ratio of

- A: 2.07 (logrank P = .0018)
- B: 1.13 (logrank P = .0018)
- C: 0.87 (logrank P = .0018)
- D: 0.48 (logrank P = .0018)

- Lifelines:
 - 50-50? Ask the audience? Call a friend?

59

Who Wants To Be A Millionaire?.....

Proportional hazards analysis estimates a

Treatment : Control hazard ratio of

- B: 1.13 (logrank P = .0018)
- C: 0.87 (logrank P = .0018)

- Lifelines:
 - 50-50? Ask the audience? Call a friend?

60