


PENNSTATE



Lecture 7: Survival Analysis

Christopher S. Hollenbeak, PhD
Jane R. Schubart, PhD

The Outcomes Research Toolbox

Review Homework

Variable	Odds Ratio	95% Confidence		P-value
		Lower	Upper	
Age				
0-39	Reference			
40-49	1.61	0.91	2.85	0.10
50-59	2.22	1.27	3.85	0.01
60+	2.97	1.63	5.39	<0.0001
Sex				
Male	Reference			
Female	0.90	0.62	1.32	0.60
Race				
Nonblack	Reference			
Black	0.73	0.25	2.20	0.58
HLA Mismatches	1.02	0.82	1.26	0.87
Body Mass				
Normal/Underweight	Reference			
Obese	0.76	0.44	1.30	0.31
Red Cell Transfusions (per unit)	1.05	1.00	1.10	0.03
Surgical Site Infection	1.53	1.05	2.23	0.03

Review Homework

- To determine whether age works better as a continuous or categorical variable, fit both models and compute the area under the ROC curve
 - Also check whether the categorical effects look linear or nonlinear
- Same process for HLA as a continuous or categorical variable

Overview

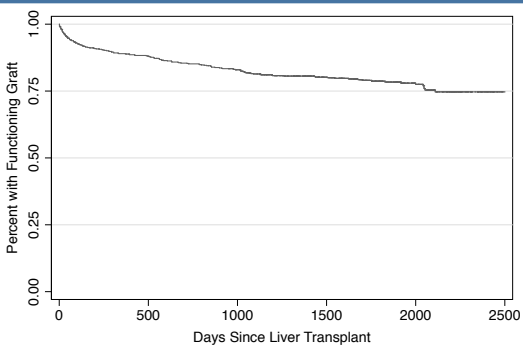
- Risk of events: Adding a time dimension
- Survival curves
 - Kaplan-Meier method
- Risk as hazard
- Survival regression
 - Cox Proportional Hazards model



Survival Analysis

- Logistic regression models the likelihood of an event happening
- It ignores how long it takes, assumes everyone has similar exposure time
- Sometimes, the time dimension is important
- Example: how long does a transplanted liver last before it gives out?





Survival Analysis

- The outcome measure that is appropriate for this question is the **survival time**
- What kind of variable is this? Continuous? Categorical? Binary?
- What is our favorite method to analyze this kind of outcome measure?



Survival Analysis

- Survival time is a continuous variable
- Why can't we just use a linear regression to analyze time to event as our dependent variable?



Survival Analysis

- Three patients
 - **Patient 1** has a liver transplant on January 1, and on February 1 the organ is rejected and s/he gets retransplanted
 - **Patient 2** has a liver transplant on July 1. As of today the organ is functioning fine.
 - **Patient 3** has a liver transplant on February 1, and on March 1 moved to Thailand, never to be heard from again
- What is the survival time for each patient?



Survival Analysis

- The survival time is:
 - Patient 1: 31 days
 - Patient 2: ?? At least 37 days...
 - Patient 3: ?? At least 28 days...
- Because of either lack of follow-up or the end of follow-up these data are “censored”
- We could use linear regression if there was no censoring
 - But if we apply linear regression to censored data we will get biased results



Survival Analysis

- If we compare time to event, we ignore censoring
- If we compare proportions, we ignore time
- Survival analysis allows us to address both issues
 - We study time to event while dealing with censoring



Survival Analysis

- Our outcome measure requires two variables
 1. Time to event
 2. A censoring indicator that shows whether the end of the time to event was an event or a censor
- For example, our data for our three example patients would be
 - 31 days, graft failure=1
 - 37 days, graft failure=0
 - 28 days, graft failure=0



Survival Analysis

- There are both univariate and multivariate approaches for survival analysis
- Univariate: Kaplan-Meier analysis
- Multivariate: Cox Proportional Hazards Regression
 - There are others, but most of the time these are the methods that are used

13

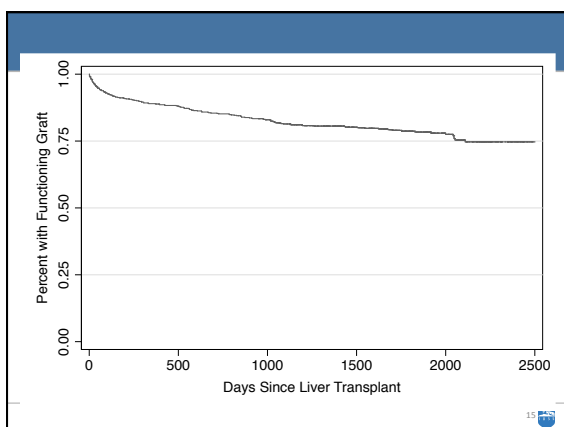
Kaplan-Meier Analysis

- Kaplan-Meier analysis produces “survival curves”
- Survival curves are estimates of the survivor function
- The survivor function is

$$S(t) = 1 - \Pr(T > t)$$

- The probability of surviving beyond some time period t

14



15

Kaplan-Meier Analysis

- The Kaplan-Meier method computes the survival probability as a compound probability
 - The probability of being alive at time 2 is the probability of surviving time 1 times the probability of surviving time 2
- At $t = 0$, everyone is alive
- Then for each time period after, the probability of surviving is a function of patients available in the current period
 - The denominator changes at each new time period
- Censoring is handled by dropping them from the denominator

16

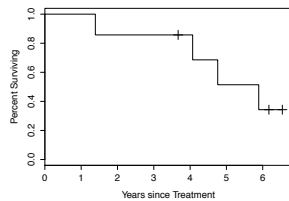
Kaplan-Meier Example

Time Period	Patients At Risk	Patients Censored	Patients Died	Patients Survived	Kaplan-Meier Survival Probability
Year 1	100	3	5	95	$(95/100)=0.95$
Year 2	92	3	10	82	$(95/100) \times (82/92)=0.8467$
Year 3	79	3	15	64	$(95/100) \times (82/92) \times (64/79)=0.70$
Year 4	61	3	20	41	$(95/100) \times (82/92) \times (64/79) \times (41/61)=0.4611$
Year 5	38	3	25	13	$(95/100) \times (82/92) \times (64/79) \times (41/61) \times (13/38)=0.1577$

17

Reading a Kaplan-Meier Curve

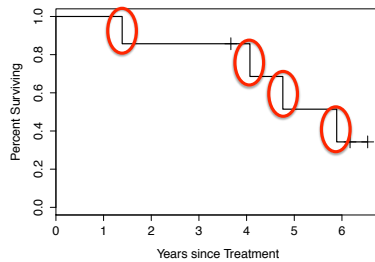
- How many events occurred during this study?
- How many patients were censored?



18

Reading a Kaplan-Meier Curve

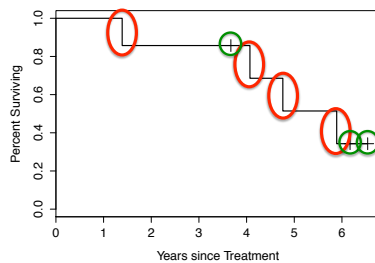
Events occur where curve bends



Reading a Kaplan-Meier Curve

Events occur where curve bends

Censoring is indicated with symbols



Reading a Kaplan-Meier Curve

- Considerations for large data sets
- With a large number of observations
 - There may be too many events to count on a curve
 - There may be too many censoring events to plot along the curve
- Frequently the censoring symbols are omitted

Stata Code

- `stset` command is used to tell Stata the format of your survival data
- Example: `stset gs_days, failure(gfail==1)`
 - Only have to “tell” Stata once, after which all survival analysis commands (the `st` commands) will use this information
- Stata needs to know the **time at risk** (e.g., time from diagnosis to death or censoring) AND the **failure indicator** (e.g. whether or not the patient died)

22

Stata Code

- After you run the `st set` command, other commands are available to:
 - Plot the Kaplan-Meier curve
 - Perform a statistical test to compare Kaplan-Meier curves
 - List the points that are graphed
- Code to plot Kaplan-Meier curves
 - `sts graph, by(strata)`
- Code to compare Kaplan-Meier curves
 - `sts test strata`
- Code to list survival percents
 - `sts list`

23

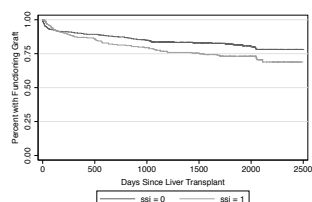
Kaplan-Meier Example

- We want to know whether surgical site infection increases the likelihood of losing the organ after liver transplantation
- We have the following variables:
 - `gs_days` tells how long the organ lasted
 - `gfail` is an indicator for whether the liver failed
 - `ssi` is our surgical site infection variable
- Step 1: `stset` the data
 - `stset gs_days, failure(gfail)`

24

Kaplan-Meier Example

- Step 2: Plot the Kaplan-Meier curves, stratified by SSI
 - `sts graph, by(ssi) title("") xtitle("Days Since Liver Transplant") ytitle("Percent with Functioning Graft")`



25

Kaplan-Meier Example

- Step 3: Perform log rank test to compare the curves

`sts test ssi`

Log-rank test for equality of survivor functions

ssi	Events observed	Events expected
0	86	100.49
1	73	58.51
Total	159	159.00

chi2(1) = 5.69
Pr>chi2 = 0.0171

26

Kaplan-Meier Example

- Step 4: What is the percent of organs still functioning one year after transplant?

`sts list, by(ssi)`

Seg. Time	Total	Net Fail	Lost	Survivor Function	Std. Error	[95% Conf. Int.]
ssi=0						
1	485	6	0	0.9876	0.0050	0.9727 0.9944
2	479	0	3	0.9876	0.0050	0.9727 0.9944
.						
334	423	1	0	0.9009	0.0137	0.8704 0.9246
371	422	1	0	0.8988	0.0139	0.8680 0.9228
ssi=1						
6	292	0	1	1.0000	.	. .
9	291	1	0	0.9966	0.0034	0.9759 0.9995
.						
330	235	1	0	0.8700	0.0202	0.8243 0.9045
376	234	0	1	0.8700	0.0202	0.8243 0.9045

27

R Code

- R follows a similar pattern
 - Create a survival object
 - Apply functions to the survival object
- Step 1: Install the “survival” package
 - `install.packages("survival")`
- Step 2: Load the “survival” library
 - `library(survival)`
- Step 3: Create a survival object
 - `sv1 <- Surv(time, failure) ~ strata`



R Code

- Step 4: Perform log rank test using `survdif()`
 - `survdif(sv1)`

```
Call:
survdif(formula = sv1)

      N Observed Expected (O-E)^2/E (O-E)^2/V
dat1$ssi=0 485      86   100.5      2.09    5.69
dat1$ssi=1 292      73    58.5      3.59    5.69

Chisq= 5.7 on 1 degrees of freedom, p= 0.0171
```



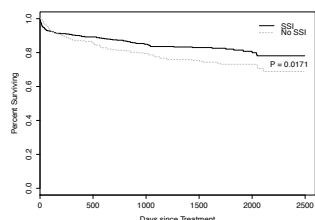
R Code

- Step 4: Produce Kaplan-Meier plots by plotting a `survfit`
 - `plot(survfit(sv1))`
 - Use options for aesthetics
 - `plot(survfit(sv1), xlab="Days since Treatment", ylab="Percent Surviving", lty=c(1,2), col=c("black","grey75"), lwd=2, cex=2, mark.time=FALSE)`
 - `lty` = line type (1 = solid, 2=dash, 3=dots, etc.)
 - `lwd` = line width (scaling factor)
 - `mark.time` = turn censoring markers on or off



R Code

```
sv1 <- Surv(datl$gs_days, datl$gfail) ~ datl$ssi
plot(survfit(sv1), xlab="Days since Treatment",
     ylab="Percent Surviving", lty=c(1,2),
     col=c("black","grey75"), lwd=2, cex=2, mark.time=FALSE)
```



R Code

- To add a figure legend use the `legend()` function
 - `legend(2000, 1, c("SSI", "No SSI"),`
`lty=c(1,2), col=c("black", "grey75"),`
`bty="n")`
- To add a p-value use the `text()` function
 - `text(2350, .73, "P = 0.0171")`

R Code

- Step 5: Get a detailed list of survival percentages by time using `summary(survfit(sv1))`

```
> summary(survfit(sv1))
Call: survfit(formula = sv1)
```

```
datl$ssi=0
time n.risk n.event survival std.err lower 95% CI upper 95% CI
1 485 6 0.988 0.00502 0.978 0.998
3 476 1 0.986 0.00542 0.975 0.996
5 471 1 0.983 0.00580 0.972 0.995
6 468 1 0.981 0.00616 0.969 0.993
```

```
datl$ssi=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
9 291 1 0.997 0.00343 0.990 1.000
20 287 1 0.993 0.00487 0.984 1.000
23 286 1 0.990 0.00596 0.978 1.000
26 284 1 0.986 0.00688 0.973 1.000
30 282 1 0.983 0.00770 0.968 0.998
```

Summary

- Analysis of time-to-event data requires special methods
- Not all subjects will have experienced the event by the end of the study; others may be lost to follow-up
- This is called censoring
- Survival curves account for this censoring as well as the total time of exposure
- Kaplan-Meier analysis is the most common method for estimating survival curves
 - It allows simple (one variable) stratification and comparisons



Yet Another Measure of Risk

- When risk of an event involves **time** we need a new measure of risk
- Hazard is the “instantaneous” risk of an event
 - The risk of having an event at time point t given that the event has not yet occurred
- Example:
 - Among all liver transplant patients, 5% of transplants fail per year. This implies that grafts fail at a certain rate per month, or per week, or per day. The hazard is the probability of failure as the time point shrinks to 0



Hazard of and Event

- We can compute the **average hazard rate** as
 - Total number of failures divided by observed survival time (units are therefore $1/t$ or $1/\text{pt-yrs}$)
- Example: In our liver transplant data set we have 159 graft failures and 1,061,029 patient days
What is the average hazard rate?
 - $159/1,061,029 = 0.0001499$ failures per patient day
 - $159/(1,061,029/30) = 0.004496$ failures per month
 - $159/(1,061,029/365.25) = 0.0547$ failures per year



Cox Proportional Hazards Model

- The Cox Proportional Hazards Model is the most commonly used multivariate survival method
- Models the hazard rate of an event as a function of covariates
- Separates the “baseline” hazard rate from covariates



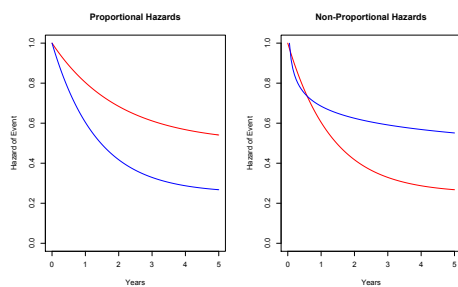
Cox Proportional Hazards Model

$$h(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

- $h_0(t)$ is called the **baseline hazard**
 - It is the hazard assuming all covariates equal zero
 - The hazard for the reference patient
- Covariates impact the hazard rate by scaling the baseline hazard by a constant
- This means that the model assumes that the effect of a covariate is proportional across all units of time: **proportional hazards assumption**
 - e.g. being male implies you have x times the hazard at ANY point in time: time 0, time 10, or time 100000000000



Proportional Hazards Assumption



Proportional Hazards Assumption

- Always compute Kaplan-Meier curves first
- See whether any lines cross for your covariates
- If they cross, proportional hazards assumption is violated and you can't do Cox Regression
- If they don't cross, you have a green light to do Cox Regression



Cox Proportional Hazards Model

- Take logs of both sides, and voila! Linear survival model

$$\ln(h(t)) = \ln(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- Coefficients are the effect of the covariate on the log hazard rate
- Exponentiate the coefficient and you get the **hazard ratio**



Interpreting Hazard Ratios

- Hazard Ratio tells you how covariate changes the baseline hazard proportionally to the reference group
- Recall hazard is risk of an event at any point in time
 - $HR > 1$ means increased hazard
 - $HR = 1$ mean equal hazard
 - $HR < 1$ means reduced hazard



Time Out

- I know this looks scary and the equations are intimidating!
- But Cox regression is really just like logistic regression, only applied to time-to-event data
- We start with an unintuitive measure of risk (hazard versus odds), take logs, and get a linear model
- When we exponentiate the coefficients we get a hazard ratio (versus an odds ratio)
- Interpretation is very similar



Stata Code

- Must start with an `st set` statement
- Then you specify the model with `stcox`
`stcox covar1 covar2 ... covark`
- For example, to estimate the effect of covariates on graft survival in liver transplant patients

```
stset gs_days, failure(gfail)
stcox ssi age4049 age5059 age60 ///
female black ab0 ab1 ab2 ab3
```



Stata Results

```
No. of subjects =      777          Number of obs   =      777
No. of failures =      159
Time at risk    =    1061029
Log likelihood   =   -1003.7557      LR chi2(10)    =     26.36
                                      Prob > chi2    =     0.0033
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
ssi	1.505696	.2426318	2.54	0.011	1.097925 2.064915
age4049	1.423765	.3460387	1.45	0.146	.8842152 2.292547
age5059	1.730902	.4066374	2.34	0.020	1.092198 2.743112
age60	2.246839	.5611624	3.24	0.001	1.377143 3.665768
female	.9552656	.1552651	-0.28	0.778	.6946615 1.313636
black	.763481	.3520942	-0.59	0.558	.3092075 1.885153
ab0	.9113616	.6554753	-0.13	0.897	.222579 3.731619
ab1	.2120457	.2136852	-1.54	0.124	.0294203 1.528314
ab2	.952278	.2372452	-0.20	0.844	.5843866 1.551777
ab3	1.273965	.2203778	1.40	0.162	.9076359 1.788147



R Code

- Use the `coxph()` function to create a proportional hazards object
- Then summarize the object
 - `cox1 <- coxph(Surv(dat1$gs_days, dat1$gfail) ~ data1$age4049 + data1$age5059 + data1$age60 + data1$female + data1$black + data1$abmm + data1$ssi)`
 - `summary(cox1)`



R Results

```
> summary(cox1)
n= 777, number of events= 159

              coef exp(coef) se(coef)      z Pr(>|z|)
data1$age4049  0.38022  1.46260  0.24249  1.568 0.116892
data1$age5059  0.58384  1.79309  0.23453  2.490 0.012779 *
data1$age60    0.82178  2.27453  0.24905  3.300 0.000968 ***
data1$female   -0.05649  0.94508  0.16266 -0.347 0.728377
data1$black    -0.26939  0.76384  0.46115 -0.584 0.559099
data1$abmm     0.05595  1.05755  0.09012  0.621 0.534682
data1$ssi      0.40825  1.50418  0.16071  2.540 0.011077 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
data1$age4049  1.4626  0.6837  0.9093  2.353
data1$age5059  1.7931  0.5577  1.1323  2.839
data1$age60    2.2745  0.4397  1.3960  3.706
data1$female   0.9451  1.0581  0.6871  1.300
data1$black    0.7638  1.3092  0.3094  1.886
data1$abmm     1.0576  0.9456  0.8863  1.262
data1$ssi      1.5042  0.6648  1.0977  2.061

Concordance= 0.593 (se = 0.024 )
Requeren 0.025 (max possible= 0.927 )
Likelihood ratio test= 19.43 on 7 df,  p=0.006939
Wald test              = 18.92 on 7 df,  p=0.008425
Score (logrank) test = 19.35 on 7 df,  p=0.007162
```



Variable	Hazard Ratio	95% Confidence		P-value
		Lower	Upper	
Age	REFERENCE			
0-39				
40-49	1.42	0.88	2.29	0.146
50-59	1.73	1.09	2.74	0.020
60+	2.25	1.38	3.67	0.001
Sex	REFERENCE			
Male				
Female	0.96	0.69	1.31	0.778
Race	REFERENCE			
Nonblack				
Black	0.76	0.31	1.89	0.558
HLA Mismatches				
0	0.91	0.22	3.73	0.897
1	0.21	0.03	1.53	0.124
2	0.95	0.58	1.55	0.844
3	1.27	0.91	1.79	0.162
4	REFERENCE			
Surgical Site Infection	1.51	1.10	2.06	0.011



The Narrative

Age had a significant association with graft survival. Patients age 50-59 had a 73% greater hazard of losing their graft ($p=0.02$) and patients age 60+ had 2.25 times greater hazard of losing their graft ($p=0.001$). Patients with a surgical site infection were 51% more likely to lose their graft than patients without a surgical site infection ($p=0.011$).



Homework

- Using the Liver Transplant data:
- Create Kaplan-Meier survival curve using **patient** survival and explore how survival is influenced by age, sex, and SSI
 - Test whether curves are different using log rank test
- Fit a Cox proportional hazards model and control for covariates you used in your logistic regression model of mortality