An Introduction to Survival Analysis: Whether, When, and How

Margaret K. Keiley, Ed.D.
Auburn University
keilemk@auburn.edu

SRA Peer Pre-Conference, 2012, Vancouver, BC, Canada

1

Survival Analysis: Types of Research Questions

"Whether" and, if so, "when" a critical events in a life-course occur:

- Sexual initiation
- Marriage
- Birth of first child
- First arrest
- Dropping out of high school

And what predicts those event occurrences

2

Why Common Statistical Methods Fail to Answer these Questions Well

Whether and *when* an event occurs is unknown for some people under investigation

These cases are CENSORED
The outcome is unknown

3

Censoring

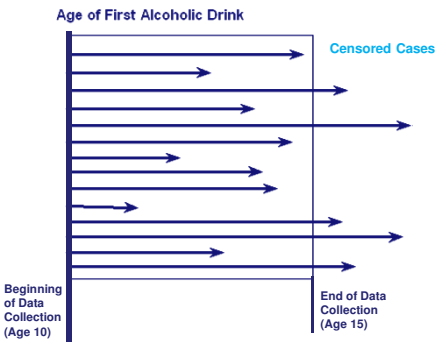
Censored cases:

They are people in the sample who do not experience the target event while they are being observed or, they may do so *after* data collection ends

They are the ones who are *least likely* to experience the event, the participants who "survive" the longest

4

Censored Cases



5

Common Strategies for Dealing with Censoring and Their Problems

Ignore the censored cases completely, treating them as if they were missing

- Reduces statistical power
- Negatively biases estimates of time to event

Imputing unknown event times

- Event time equal to length of data collection
- Underestimates the true length of time-to-event

Dichotomization

- Collapses perfectly good continuous duration data to create a single new dichotomous outcome that may be meaningless!

- Individuals who first experienced depression before or after age 25
- Those with initial depression in early childhood would be pooled with those who did not become depressed until their 20s then compared with those who did not become depressed until after 25!
- Such individuals may differ enormously in the causes of depression

6

Survival Analysis, Event History Analysis, Hazard Modeling:

A data-analytic method that deals evenhandedly with both the non-censored and the censored observations

7

Measuring Time

Define a "beginning of time" in some meaningful and unambiguous way – one that places all individuals in the population *at risk* of experiencing the forthcoming target event

Establish a suitable metric in which the passage of time can be recorded

Irrevocably recognize the target event when it occurs

(Must be mutually exclusive and exhaustive states)

8

Measuring Time: Continuous or Discrete?

Continuous: If the exact time of an event is known (e.g., age of teen mother in days from her own birth to the birth of her first child)

Discrete: Events occur (or can be remembered to have occurred, in the case of retrospective data) across discrete periods, or separable and often large units of time (e.g., age of first sexual intercourse)

I focus on discrete-time event history analysis for this presentation, as this framework allows for a more straight-forward and simple explanation of the method of event history analysis itself

For more information on continuous-time event history analysis see Singer and Willett (2003) or Allison (1995) 9

DESCRIPTIVE ANALYSIS OF DISCRETE-TIME SURVIVAL DATA

Prior to fitting discrete-time survival models, we recommend that you conduct a descriptive analysis of the data

(Keiley, Martin, Canino, Singer, & Willett, 2007; Singer & Willett, 2003)

Sample Hazard Probabilities, Sample Survival Probabilities, Median Life Time

10

Our Research Questions:

When are adolescents and young adults most likely to have sexual intercourse for the first time?

Do attachment style and religiosity predict the timing of first sexual intercourse?

Dissertation by Dr. Canino: Using survival analysis to explore the relationship among attachment theory, religiosity, and sexual initiation

11

Data Collection

Retrospective longitudinal data from 618 university students in a large, mid-western public university who completed a confidential online web survey in 2004 about their:

- **Sexual history**
- **Level of religiosity**
- **Quality of their romantic attachment style**
- **Selected demographic information**

12

Self-Report Measures

Attachment Style

Experiences in Close Relationships (**ECR**; Brennan, Clark, & Shaver, 1998)

- Romantic Attachment
- Categorical: Secure, Avoidant, Preoccupied

Religiosity

Duke Religion Index (**DUREL**; Koenig, Meador, & Parkerson, 1997)

- Scores range from 5-27

Sexual Intercourse Experience

7 questions about sexual experience including age at first sexual intercourse

13

Demographics

SEX: 72% Women
28% Men

AGE: Age range: 16 to 45 (*M*=21, *SD*=2.3)

RACE: 86% Caucasian
7% African-American
3% Hispanic
3% Asian

MARITAL STATUS: 86% Single
9% Engaged
4% Married
1% Divorced

14

Hazard Probability

Answers the question, "When is the target event most likely to occur?"

In our example, "At what ages are adolescents and young adults at *greatest risk* of sexual initiation?

The *population hazard probability* is defined as the *conditional* probability that a randomly-selected person will experience the target event in the any given time-period, *given that he or she had not experienced the event in an earlier time-period*

A Conditional Probability

15

Survival Probability:

Answers the question, "How much time passes before people are likely to experience the event, that is have sexual intercourse?"

In our example, "How long do adolescents remain virgins?"

The *population survivor probability* in any discrete time-period is the probability that a randomly selected person will *'survive' beyond* the current time-period without experiencing the target event

In other words, it is the probability that he or she has *not* experienced the event during the current, or any earlier, time period

16

Median Life Time

The population median life time is the length of time that must pass before the population survival probability drops below a value of *one half*

17

Table 1. Life table describing the age at which respondents first had sexual intercourse (*N*=618).

Age in years	Number			Proportion of	
	Respondents at the beginning of the age period who had not had sexual intercourse **(Risk Set)**	Respondents who had sexual intercourse during this age period	Censored at the end of the age period **(Drop Outs)**	Respondents at the beginning of the age period who did have sexual intercourse during this year **(Hazard Function)**	All respondents who still had not had sexual intercourse at the end of the age period **(Survival Function)**
11	618	-	-	-	1.0000
12	618	1	0	0.0016	0.9984
13	617	2	0	0.0032	0.9952
14	615	14	0	0.0228	0.9725
15	601	52	0	0.0865	0.8884
16	549	98	0	0.1785	0.7298
17	451	87	1	0.1929	0.5890
18	363	123	14	0.3388	0.3894
19	226	63	23	0.2788	0.2808
20	140	35	24	0.2500	0.2106
21	81	18	63	0.2222	0.1638

Median Life Time

Hazard= Experienced Event/Risk Set

Survival = (1-Hazard)* Previous Survival

18

Person-Period Data Set

Data for discrete-time survival analysis are best stored in a *person-period format* in which each person contributes *one record* (row) to the dataset for *each discrete time-period* in which he/she is at risk for the event occurrence, in this case first sexual intercourse

I will briefly show this process in the next few slides

See Singer & Willett (2003) or Keiley, Kirkland, Zaremba, & Anders (2011) for details about how to do this

19

Censor Variable

CENSOR Censor (1=censored, 0=Not censored)

If R had sex, she is not censored (Censor = 0)

If R did not have sex, she is censored (Censor = 1)

That is we do not know if she had sex after we lose track of her either because data collection ended or she dropped out

If no one drops out of the study, you can calculate the "life table" by doing:

```
proc freq;
   tables yr_age*censor/list missing;
  run;
```

yr_age = How old respondent was when first had sex

20

Creating the Person Period Data Set from the Person Level Data Set

Person Level Data Set

TCID	Yr_Age at First Sex	Current Age	Censor
11027	.	20 Drop-Out Age =20	1
11028	16	19	0
11030	21	23	0
11031	17	20	0

21

SAS Program Creating Person Period Data Set

```
array p[11] a12-a21;  *Variables in this array will be the age
                       dummy variables;
do period = 12 to min(drpo_age, yr_age, 21);
     if period = yr_age and censor = 0 then Y=1;
          else Y=2;  *Creates event variable, Y;
                  if Y=1 then did not have sex at that age, if Y=2 did not;
do index = 1 to 10;  *Creates the actual values for age
                      dummy variables;
     if index=(period-11) then p[index] = 1;
       else p[index] = 0;
end;
     output;
end;
```

22

TCID	a12	a13	a14	a15	a16	a17	a18	a19	a20	a21	CENSOR	Y	
11027	1	0	0	0	0	0	0	0	0	0	1	2	11027: Never experienced the event (only 20 years old)
11027	0	1	0	0	0	0	0	0	0	0	1	2	
11027	0	0	1	0	0	0	0	0	0	0	1	2	
11027	0	0	0	1	0	0	0	0	0	0	1	2	
11027	0	0	0	0	1	0	0	0	0	0	1	2	
11027	0	0	0	0	0	1	0	0	0	0	1	2	
11027	0	0	0	0	0	0	1	0	0	0	1	2	
11027	0	0	0	0	0	0	0	1	0	0	1	2	
11027	0	0	0	0	0	0	0	0	1	0	1	2	
11028	1	0	0	0	0	0	0	0	0	0	0	2	11028: experienced event at age 16
11028	0	1	0	0	0	0	0	0	0	0	0	2	
11028	0	0	1	0	0	0	0	0	0	0	0	2	
11028	0	0	0	1	0	0	0	0	0	0	0	2	
11028	0	0	0	0	1	0	0	0	0	0	0	1	
11030	1	0	0	0	0	0	0	0	0	0	0	2	11030: Experienced the event at age 21
11030	0	1	0	0	0	0	0	0	0	0	0	2	
11030	0	0	1	0	0	0	0	0	0	0	0	2	
11030	0	0	0	1	0	0	0	0	0	0	0	2	
11030	0	0	0	0	1	0	0	0	0	0	0	2	
11030	0	0	0	0	0	1	0	0	0	0	0	2	
11030	0	0	0	0	0	0	1	0	0	0	0	2	
11030	0	0	0	0	0	0	0	1	0	0	0	2	
11030	0	0	0	0	0	0	0	0	1	0	0	2	
11030	0	0	0	0	0	0	0	0	0	1	0	1	
11031	1	0	0	0	0	0	0	0	0	0	0	2	11031: Experienced the event at age 17
11031	0	1	0	0	0	0	0	0	0	0	0	2	
11031	0	0	1	0	0	0	0	0	0	0	0	2	
11031	0	0	0	1	0	0	0	0	0	0	0	2	
11031	0	0	0	0	1	0	0	0	0	0	0	2	
11031	0	0	0	0	0	1	0	0	0	0	0	1	

23

MODELING EVENT OCCURRENCE

Assumptions for Survival Analysis

1) For every value of a predictor, an hypothesized logit hazard function does exist
2) The linear additivity assumption: Each of the hypothesized logit hazard functions in #1 have an identical shape (violations are allowed and they are interesting – moderation)
3) The proportionality assumption: The hypothesized logit hazard functions in #1 are all equidistant from each other (violations are allowed and they are interesting – time varying covariates)

For information about these assumptions, please see Singer & Willett (2003) or Keiley et al. (2011)

24

Baseline Fitted Logit-Hazard Model: Time as Predictor

A *logistic* model is fit to the person-period data set with "age" as the only predictor of event occurrence

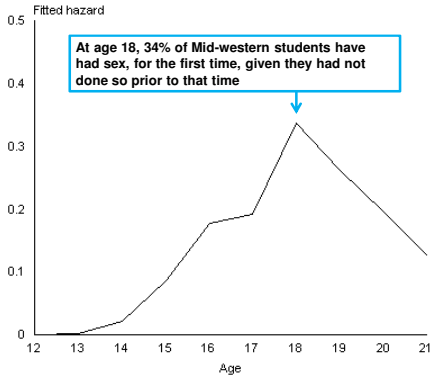
$$logit\ h_i(t_j) = \left[\alpha_{12}A12_j + \alpha_{13}A13_j + \cdots \alpha_{21}A21_j\right]$$

The logistic regression analyses is fit with the "no-intercept" option selected, as there is no stand-alone intercept included in the model specification

In fact, in our model specification the *α*-parameters function as a set of 10 intercepts, one per discrete-time period

25

Fitted Baseline Logit Hazard Function

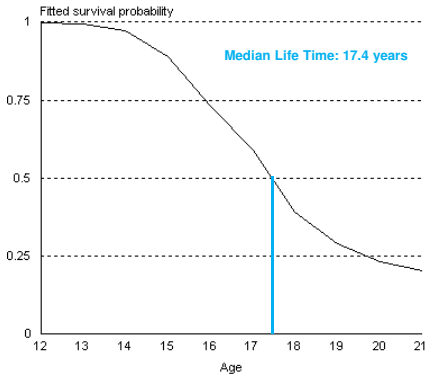


The parameter estimates of the time period dummy predictors, *A12* through *A21*, provide the fitted shape of the baseline logit-hazard profile for the average mid-Western adolescent

The hazard increases from age 11 until age 18 and then decreases, but not back to the original level

26

Fitted Baseline Survival Function



By the median life time, half of the population of mid-western college students will have had sex and 50% will still be surviving

However, by age 21, there are still approximately 23% still surviving

27

Adding Other Predictors: Attachment Style

$$logit\ h_i(t_j) = \left[\alpha_{12}A12 + \alpha_{13}A13_j + \cdots \alpha_{21}A21_j\right]$$

$$+\left[\beta_1 AVOID_i + \beta_2 PREOCC_i\right]$$

28

Model Fit Statistics: Models 1 & 2

Deviance Statistic (-2 log likelihood)
Baseline Model 1: Deviance = 2514, df=11
Model 2: Attachment, Deviance = 2488, df=13

Akaike Information Criterion (AIC)
(Corrects for sample size and the # of parameters)
Baseline Model 1: AIC= 2936
Model 2: Attachment, AIC = 2514

Bayesian Information Criterion (BIC)
(Corrects for sample size)

29

Comparison of Models 1 & 2:

Model 1: No Predictors (Deviance = 2514)
Model 2: Attachment style (Deviance = 2488)

Δ Deviance =26, *df*=2, *p*<.001

Attachment Style is a significant predictor of sexual initiation

30

Adding Another Predictor: Religiosity

$$logit\ h_i(t_j) = [\alpha_{12}A12 + \alpha_{13}A13_j + \cdots \alpha_{22}A22_j]$$
$$+[\beta_1 AVOID_i + \beta_2 PREOCC_i]$$
$$+[\beta_3 RELIGIOSITY_i]$$

31

Model Fit Statistics: Model 1-3

Deviance Statistic (-2 log likelihood)

Model 1: no Predictors: Deviance = 2514, df=11
Model 2: Attachment: Deviance = 2488, df=13
Model 3: Attachment, Religiosity: Deviance = 2451, df 14

Akaike Information Criterion (AIC)

Model 1: No Predictors: AIC= 2936
Model 2: Attachment: AIC = 2514
Model 3: Attachment, Religiosity: AIC = 2479

Decreasing Fit Statistics: Better fit as we add predictors

32

Comparison of Models

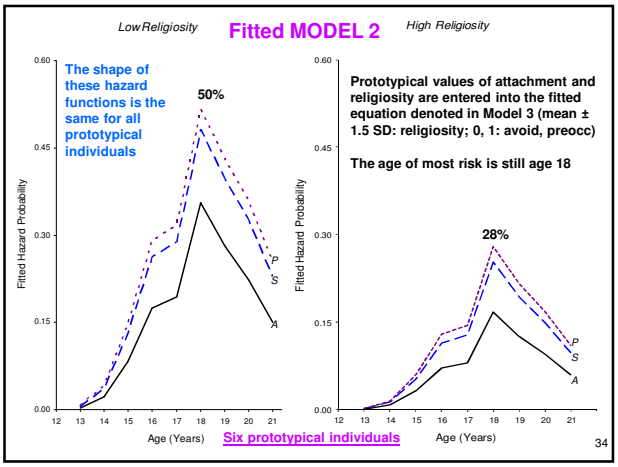
Model 2: Attachment Style (Deviance = 2488, df=13)
Model 3: Attachment, Religiosity (Deviance = 2451, df=14)

Δ Deviance =37, *df*=1, *p*<.001

Religiosity is a significant predictor of sexual initiation, even controlling for Attachment Style

33

Fitted MODEL 2



Prototypical values of attachment and religiosity are entered into the fitted equation denoted in Model 3 (mean ± 1.5 SD: religiosity; 0, 1: avoid, preocc)

The age of most risk is still age 18

Six prototypical individuals

34

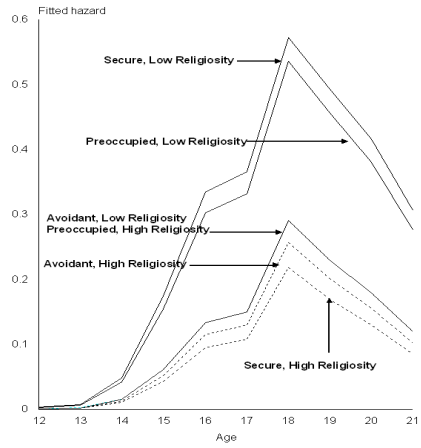
Model with Attachment Style, Religiosity, and Their Interaction: Testing the Additivity Assumption

$$logit\ h_i(t_j) = [\alpha_{12}A12 + \alpha_{13}A13_j + \cdots \alpha_{22}A22_j]$$
$$+[\beta_1 AVOID_i + \beta_2 PREOCC_i]$$
$$+[\beta_3 RELIGIOSITY_i]$$
$$+[\beta_4 RELIGIOSITY * AVOID_i + \beta_5 RELIGIOSITY * PREOCC_i]$$

Do each of the logit hazard functions have an identical shape or does the effect of attachment on risk of sexual initiation vary across levels of religiosity? (moderation)

35

Fitted MODEL 3



Age 18 is the age at which individuals are most at risk for initiating sexual intercourse for the first time, given they had not prior to that

Nearly 60% of secure and 55% of preoccupied individuals with low religiosity have sex for the first time at 18

29% of avoidant individuals with low religiosity & 29% of preoccupied with high religiosity initiated sex at 18

Secure individuals with high religiosity are least likely to initiate sex at any of these ages than the other highly religious groups

36

Extensions of Survival Analysis

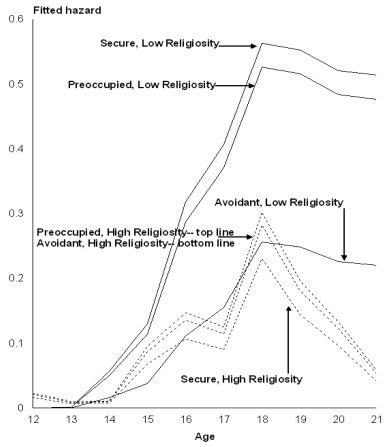
Including time-varying predictors

Alternate specifications for time

Interactions with time -- A predictor's effect may vary over time. It may have a different impact in different time periods

We tested the Proportionality Assumption by adding the interaction of Religiosity with time – It was significant

37



Fitted MODEL 4

Individuals who are *not highly* religious continue to be at higher, more steady, risk for sexual initiation after age 18 than do *highly* religious individuals whose risk declines after that age

Highly religious individuals appear to have a dip in the risk for sexual initiation at age 17, while those who are not highly religious continue their upward trend of high risk at that age

If predictors do have a different effect on the hazard function at different time periods, then relaxing the proportionality assumption is appropriate

38

Take-Home Message:

If you are interested in "Whether" and, if so, "when" critical events in a life-course occur and what predicts the occurrence of those events then conduct a survival analysis!

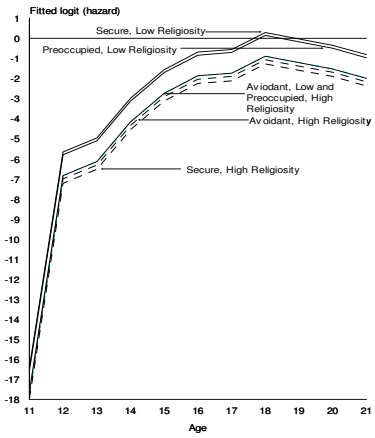
39

Interpretive Note of Caution #1

Apparent differences (gap sizes between the fitted hazard functions for the different prototypical individuals) do not represent a time-varying effect of the predictors, but the hazard scale on which it has been plotted

Plotted on the Logit Hazard Scale on which it was estimated, no varying gap is seen

40



Fitted MODEL 2

41

Interpretive Note of Caution #2

These plots represent fitted functions, *not* sample estimates from an analysis of a sub-group of the sample

Prototypical plots like those presented here can be constructed in any analysis, regardless of the specific methodology used

42

Questions

What is the "whether" and "when" question that you will try to answer with your data?

What is the "beginning of time?"

What is the metric in which the passage of time is recorded?

How is the target event defined?

Who drops out of the study and how do you denote these respondents?

43