

Jason & Glenwick, eds.

Methodological Approaches to Community-Based Research

Chapter 9/Web Appendix guide

Web Appendix Guide for Chapter 9/ Survival Analysis in Prevention and Intervention Programs

Survival Analysis Statistics

The survival function, $S(t)$, and the hazard function, $h(t)$, are the primary statistics of interest when conducting a survival analysis. Details on the calculation of each are outlined below.

Survival Function

Formulas for the survival function vary depending on the specific type of survival analysis method being used. The Kaplan-Meier method illustrates one such calculation of the survival function. Event times are arranged in increasing order beginning at $t_1 < t_2 < \dots < t_k$, with k representing the total number of observed target events. To estimate $S(t)$ to a particular time j , the following formula is used:

$$S(t_j) = S(t_{j-1}) \times \Pr[T > t_j \mid T \geq t_j]$$

in which the probability of surviving past the previous failure time t_{j-1} is multiplied by the conditional probability of surviving past the specified time t_j , given survival to at least the specified time t_j .

Hazard Function

The hazard function, $h(t)$, for discrete time survival models such as the Life Table method is calculated using the following formula:

$$h(t_{ij}) = \Pr[T_i = t_j \mid T_i \geq t_j]$$

This reflects the probability that an individual i will experience the target event in time period j , given that he or she did not experience the event during a prior interval.

For survival analysis in which time is measured continuously, the hazard function is expressed with a limit that changes as time from one instant to the next approaches 0 using the following formula:

$$h(t_{ij}) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[t_j \leq T_i < t_j + \Delta t \mid T_i \geq t_j]}{\Delta t}, t \geq 0$$

and ranges from 0 to infinity to represent the rate of hazard rather than the probability.

Data Set-Up

Online Table 9.1 depicts a typical data set-up to conduct survival analysis with either discrete or continuous time, as long as no time-varying indicators are used. The table includes data for a random sample of 20 participants for whom data was generated in the hypothetical example described in the chapter. In addition to **Case ID**, the data file includes elements related to event duration and selected covariates. Duration-related data elements include the following: **start day** (i.e., the day the participant entered the study from 1-180), **end day** (i.e., the day the participant recidivated from 1-365), and target event occurrence or **event** (i.e., whether the

participant was charged with a new offense during the observation period). From these data elements, we can calculate study duration (**time**) by subtracting the start day from the end day for those who experienced the target event. For censored individuals who did not experience the target event during the study period, duration is calculated by subtracting the start day from the last day of the observation window (i.e., 365).

Model covariates hypothesized to influence rates of recidivism relevant to our evaluation study include the following: treatment condition (**Tx**), the primary focus of our evaluation, and two potential covariates: gender (**Sex**; 75% male) and a hypothetical juvenile justice risk assessment measure (**Risk Score**) ranging from 0–5 with higher scores reflecting increased risk for recidivism. In the simulated dataset, assignment to condition resulted in 95 individuals being assigned to the comparison group and 105 individuals being assigned to the treatment group.

Life Table Analysis

Table 9.2 in the book's chapter summarizes results of the life table analysis of overall recidivism rates based upon the hypothetical case example. Column 1 is the start time for each two-week interval. Columns 2 and 3 indicate the number of participants remaining in the study at the beginning of the interval and the number of participants who are censored from the study during the interval, respectively.¹ Column 4 represents the number of participants exposed to risk during the interval. The life table method assumes that censoring occurs uniformly throughout the interval, so the total number of at-risk individuals in an interval is obtained by subtracting 1/2 of the censored cases (Column 3) from the number of participants entering the interval (Column 2) (Gruber, 1999). Column 5 represents the number of terminal events (i.e., recidivism occurrences) during each interval. The proportion terminating for each interval (Column 6) is obtained by dividing Column 5 by Column 4. Column 7 represents the proportion surviving through the interval and is calculated by subtracting the value in Column 6 from 1. The cumulative proportion surviving (Column 8) is calculated by multiplying the proportion surviving for all times up to the period of interest. Thus, the cumulative proportion surviving through the interval starting in week 4 can be found by multiplying the proportion surviving for weeks 1 through 3 (i.e., $0.885 \times 0.859 \times 0.888 = 0.675$) or by multiplying the cumulative survival rate for the previous time interval and the current proportion surviving (i.e., $0.760 \times 0.888 = 0.675$). Finally, Column 9 calculates the hazard rate for participants during each interval to reflect the probability of an individual participant experiencing the target event (i.e., recidivating) during a particular interval, given that he or she did not exit the study prior to the start of the interval.

Cox Proportional Hazard Survival Analysis

A key assumption of the Cox proportional hazard model is the proportionality of the hazard function across levels of a covariate, and there are many strategies for evaluating this assumption (Allison, 1995; Bradburn, Clark, Love, & Altman, 2003; Kleinbaum & Klein, 2005). One common approach is to plot the cumulative hazard function for different levels of a covariate-sometimes referred to as a log(-log(survival)) plot-and inspect the results to see whether the lines are parallel (i.e., do not cross or intersect after the origin). Because of the subjective nature of this approach, a number of statistical procedures for assessing the proportionality assumption have been developed. Among the easiest approaches is the Schoenfeld residuals test (Grambsch & Therneau, 1994; Schoenfeld, 1982), which involves

¹ Since the study in the hypothetical example involved a six-month follow-up for all participants, no censoring occurs until week 26.

estimating a Cox regression model with covariates and saving the partial residuals for covariates in the model. These residuals are calculated only for individuals experiencing the event. A new variable must then be created that includes the rank order of survival time for individuals who experienced the event. This variable is then correlated with the partial residuals of covariates; a statistically significant correlation suggests a violation of the proportionality assumption.

Potential solutions to a violation of the proportionality assumption include conducting a stratified Cox regression model, which involves running separate Cox regression models for different levels of the covariate that violates the assumption. Alternatively, the problematic covariate may be included as a time-dependent covariate (which involves the creation of an interaction term to represent the interaction between time and the identified covariate). This procedure is beyond the scope of this chapter, but does receive attention in some of the advanced texts recommended in Chapter 9.

Advanced Survival Analytic Methods

Apart from these basic applications of survival analytic methods, there are a number of advanced applications that increase the flexibility and versatility of the method. These methods are beyond the scope of this chapter and are addressed in a number of accessible texts including those mentioned at the beginning of Chapter 9 (Allison, 1995; Kleinbaum & Klein, 2005; Singer & Willett, 2003).

Parametric Survival Models

Parametric methods such as the Parametric Cox proportional hazard models and accelerated failure time (AFT) model are similar to semi-parametric methods but differ in how hazard is modeled (Bradburn, Clark, & Altman, 2003; Tabachnick & Fidell, 2007). In each, a specific hazard distribution for event occurrence with known distributional qualities is tested. A number of hazard distributions exist (e.g., Exponential, Weibull, Gompertz) and depict different forms of event occurrence such as a constant, increasing, or decreasing rate of risk over time, as well as a combination of increasing and decreasing hazard over time. For the Parametric Cox model, the hazard distribution of the data is compared to a selected known distribution which allows for more precise and efficient parameter estimates, though the assumption of proportional hazard is still required (Bradburn, Clark, & Altman, 2003). AFT models, on the other hand, emphasize the effect of covariates to compress or elongate the survival curve on the time axis relative to a constant, and proportional hazards are not assumed (Bradburn, Clark, & Altman, 2003; Tabachnick & Fidell, 2007). Parametric models are not implemented in SPSS, but may be implemented in other programs (e.g., SAS and SYSTAT).

Time-Dependent and Time-Varying Covariates

There may be times when the inclusion of time-dependent covariates is necessary to properly conduct a survival model, such as when the proportional hazard assumption is violated. Violation of the proportional hazard assumption indicates that the influence of the covariate changes over time. To address this violation, the researcher must create a new covariate to represent the interaction between the covariate and time that is included in the Cox regression model.

In addition to covariates for which the magnitude or direction of the effect changes over time, researchers may also encounter covariates for which the value of the variable itself changes over time. This type of variable is referred to as a time-varying covariate, and its use within survival analysis allows the researcher to study the effect of non-static risk variables on event

occurrence. In order to model such data appropriately a special procedure, referred to as a counting process, must be used (Allison, 1995; Singer & Willett, 2003). The counting process requires that data be structured into a special person-period format in which individuals have separate records with specified start and end times to represent time periods with distinct values for the time-varying variable. The analysis program uses the start and end dates for each person-period record to determine risk group membership and covariate values for each time interval. Not all statistical programs can utilize this method, however, so it is important to select the appropriate statistical program when planning a study that will include time-varying covariates.

Competing Risk Models

In the models discussed throughout Chapter 9, the event of interest was presented as a dichotomous outcome (e.g., recidivism either occurs or it does not). For some outcomes, however, a number of mutually exclusive events may be possible, and the competing risk model may be more appropriate (Hachen, 1988). Singer and Willett (2003) provide an overview of the method, which involves estimating separate survival models for each outcome and treating individuals who experience an alternate (i.e., competing) event, as well as those who do not experience any event, as censored in the model for a specific target event. Competing risk models can be readily estimated in any statistical program that can run survival analysis. However, there are a number of other assumptions that also require attention, such as the tenability of non-informative censoring when competing risks are being modeled, which are addressed further in Singer and Willett (2003).

Repeatable Events

Another common phenomenon in event-related research studies is the potential for studying repeatable events. Most of the events described in Chapter 9 have focused on single-event occurrence, or were conceptualized in such a way that they could not repeat (e.g., time to first event). Some research questions may focus on events that might repeat a number of times during the observation period. This introduces the potential for dependence among multiple observations for the same participant (i.e., the same individual will have multiple records representing time to each successive event). A number of procedures for estimating these types of models exist including the WLW method proposed by Wei, Lin, and Weissfeld (1989) and the fixed effect partial likelihood (FEPL) method. Each is described in greater detail in Allison (1995).

Latent Variable Survival Models

A relatively recent advance in the field has been the development of latent variable approaches to both discrete and continuous survival methods (Masyn, 2004; Muthén & Masyn, 2005). This increased flexibility expands the utility of the method to capitalize on the strengths of general latent variable modeling applications (e.g., structural equation modeling) and more advanced applications such as latent variable mixture modeling to address unobserved heterogeneity among study participants. This application has received relatively less attention in the research and program evaluation literature and requires the use of specialized statistical programs (e.g., Mplus; Muthén & Muthén, 1998-2010) as such models are not readily implemented in most statistical programs.

REFERENCES

- Allison, P. D. (1995). *Survival analysis using the SAS System: A practical guide*. Cary, NC: SAS Institute.
- Bradburn, M. J., Clark, T. G., & Altman, D. G. (2003). Survival analysis part II: Multivariate data analysis--An introduction to concepts and methods. *British Journal of Cancer*, 89, 431–436.
- Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515–526.
- Hachen, D. S., Jr. (1988). The competing risks model: A method for analyzing processes with multiple types of events. *Sociological Methods & Research*, 17, 21–54.
- Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis: A self-learning text* (2nd ed.). New York: Springer.
- Masyn, K. E. (2004). *Discrete-time survival mixture analysis for single and recurrent events using latent variables* (Doctoral dissertation). Los Angeles: University of California, Los Angeles.
- Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, 30, 27–58.
- Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards model. *Biometrika*, 69, 51–55.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.
- Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065–1073.