ABSTRACT

Survival analysis (also known as failure time analysis) is a way of examining time to a particular event in a set of data. Originally developed to study time until death, survival analysis is being used with increasing frequency by research professionals across many disciplines. For many public health professionals and others, however, survival analysis remains a daunting and unfamiliar series of procedures. As a result, the potentially beneficial and flexible application of survival analytic techniques is often overlooked.

This paper asks the question: could survival analysis be used to examine my data? The goal of this paper is to assist the non-statistician or public health researcher in considering survival analysis as an alternative technique using SAS®. As an example, the paper explores a non-traditional application of survival data—a study of drinking during pregnancy—where the event of interest is not death. Development of this example is accompanied by SAS code illustrating data preparation, preliminary analysis, and model development. A brief overview of survival analytic techniques and terminology for the non-statistician is provided, and is supplemented with numerous resources for further study.

BEYOND PROC LOGISTIC

Most researchers in public health and related disciplines have some experience with longitudinal studies which explore a binary event of interest. An event of interest can be death, relapse, re-arrest, or any of a number of ‘success/failure’ type events.

In investigating this event of interest, a study will often seek to answer research questions such as:

- What is the impact of our new treatment on the event of interest, compared to the standard treatment?
- What factors (demographic, social, economic) appear to determine which subjects are more likely to experience the event?

Standard statistical techniques are typically used to answer such questions with regard to binary outcomes, given certain assumptions regarding the underlying distribution (Collett, 1991). For example, an odds ratio can be calculated to compare the relative number of ‘success’ events in two sets of binary data. To determine which covariates impact the probability of ‘success,’ a linear logistic model can be fit to binomial data. These techniques for statistical modeling of binary data may help answer a number of relevant research questions. However, depending on the study design, certain questions regarding treatment effects or covariate impact may also require an exploration of the time to a designated event.

WHY SURVIVAL ANALYSIS?

In survival analysis, the outcome variable of interest is time from a particular point of origin (recruitment, commencement of treatment) to a particular event (Kleinbaum, 1996). Evaluation of time to event is crucial for many types of follow-up studies. For example, to explore the efficacy of a brief intervention for prevention of alcohol consumption during pregnancy, a researcher may wish to know not only if the intervention performs effectively, but also the median time to occurrence of the event of interest (e.g., time to first drink) or the probability that a subject will drink a certain number of months following intervention. Also, he or she may wish to know if a particular covariate (e.g., demographic, social, medical) impacts time to first drink following intervention.

In developing an analysis plan to address these types of questions, the researcher will have to account for subjects with incomplete data on the event of interest. This analytic problem, called censoring, may occur when a subject withdraws from the study, becomes lost to follow-up, or does not experience the event of interest prior to the conclusion of the study. The researcher may wish to determine if there are particular characteristics of subjects which are related to their failure to experience the event of interest. Although classical theory of linear regression and least-squares estimation does not extend to data with censored observations, survival analysis techniques can accommodate censoring and permit the estimation of treatment effect adjusted for covariates (Marubini and Valsecchi, 1995).

Survival analysis is also able to accommodate two additional features of time to event data: (1) the positively skewed distribution typical of time to events, and (2) covariates whose values vary over the period of observation (time-dependent covariates) (Collett, 1994). For instance, a treatment may impact one group of
subjects immediately following administration, but the effect may disappear after a certain period of time.

Therefore, survival analysis is recommended for:

- estimation and interpretation of (a) the potential, per unit time, of an event occurring, and (b) the probability of an event occurring beyond a specified time 
- to compare study groups based on (a) the potential of an event occurring, and (b) the probability of an event occurring beyond time 
- to assess the relationship of explanatory variables to the probability of an event occurring beyond time 

(Kleinbaum, 1996)

Estimation of the probability of an event occurring beyond a specified time is provided by the survival function, . Estimation of the potential per unit time of an event occurring is referred to as the hazard function, . The hazard function is also referred to as the conditional failure rate, or the instantaneous death rate (Collett, 1994).

In situations where the relative hazard of an event occurring is being compared between groups, a ratio of the hazards of an event can be calculated. Typically, the higher is for a given , the smaller is , and vice versa (Kleinbaum, 1996). The references at the conclusion of this report can provide additional, in-depth information on these topics.

BEYOND THE CLINICAL TRIAL

Traditionally, applications of survival methods have been the domain of clinical research. Much of survival analysis has been developed and applied in relation to cancer clinical trials (Parmar, 1995). Examples of applications within the medical literature include: time to death following a heart transplant; varying survival rates of cancer patients following chemotherapy or radiation; and time to relapse for leukemia patients following remission (Kleinbaum, 1996).

However, non-medical applications of survival analytic techniques have become increasingly popular. Applications in non-traditional settings might include:

- **Criminal justice**
  In a study on violence among teenagers, two groups of teens with a history of fighting are followed for a 1 year period. One group received peer counseling at school. The principal investigator would like to determine if the counseling experience helps reduce the number of teens who are involved in fights, and keeps track of the number of subjects who are suspended, arrested, or reported to juvenile services within the study period. Of the original group of 200 teenagers, 50 subjects were involved in a fight, and 20 subjects were lost during the follow-up period. Survival analysis can be used to determine if a difference in time to first violent encounter following study enrollment differs between the two groups.

- **Social science research**
  In a study investigating counseling needs for two-year college students, researchers are interested in determining social or demographic factors associated with completion of an Associate’s degree. A group of 100 subjects is monitored for 3 years from initial enrollment in college. 20 subjects are lost to follow-up. Information is collected on age, high school grades and attendance, financial status, family/social support, and other factors thought to influence motivation to complete a degree. Survival analysis can be used to determine which factors, if any, are associated with a longer period of school attendance, controlling for additional factors such as financial aid.

- **Public health research**
  Concerned with a rise in rates of prenatal alcohol consumption, researchers are seeking a brief intervention which will effectively deter drinking during pregnancy. A group of 250 women is randomized to receive a brief intervention during a prenatal care visit. 3 subjects are lost to follow-up. Each subject records her subsequent alcohol use. Survival analysis can explore time to first drink following randomization, to see if the intervention impacted the drinking behavior of the subjects.

At first glance, these examples may appear to be quite different, both from one another and in comparison with a ‘traditional’ clinical trial. However, each study design contains design elements making them potentially amenable to the application of survival analysis:

**An Event:** Death, relapse, first drink, re-arrest or any other clearly-defined event can be designated as the endpoint of interest in a study utilizing survival analysis. When the event occurs, the time at which it occurs must be carefully recorded.

**Time to event:** This is our outcome variable of interest in survival analysis (Kleinbaum, 1996). A time origin (e.g., time of randomization, beginning of follow-up) must be unambiguously defined in order for the time to an event to be calculated. Time to an event is often labeled ‘survival time.’

**Censored data:** If the event of interest does not occur prior to the end of study follow-up, or the subject is lost to follow-up, that subject’s survival time is considered censored. In the peer counseling example, those who did
not fight are considered censored data. Likewise, the 20 inmates who were lost to follow-up are considered censored data.

A NON-TRADITIONAL EXAMPLE USING SAS

It is beyond the scope of this paper to consider each of the many aspects of the application and interpretation of survival analysis using SAS. Instead, we will explore one example of a ‘non-traditional’ application of survival analysis.

The study
The following sample data and code are taken from a study which measured the impact of a brief intervention on drinking during pregnancy. The goal of the brief intervention was to reduce drinking during pregnancy. In this study, the researchers wished to evaluate the impact of the brief intervention.

A sample of 250 pregnant women were randomized to receive either an assessment of drinking behavior and a brief intervention, or the assessment alone, in their first trimester. Following delivery, all subjects were asked to provide detailed information on their drinking habits during pregnancy. Using a calendar, patients were asked to record the dates on which they drank.

A question central to this study involved time to the event of interest (drinking): specifically, did those subjects who received the brief intervention demonstrate a difference in time to first drink following assessment when compared with those subjects who received the assessment alone?

Defining the event
Here, the event of interest is the first occurrence of drinking following randomization to either assessment and brief intervention, or assessment alone.

Defining the time to event
Time to event--the number of weeks between randomization and first drink--is our outcome variable of interest.

To create a variable indicating time to event or time to censor, a simple array was used to count the individual weeks (variables PALC8 through PALC42) in each subject’s drinking calendar.

The variable FIRSTDR is set equal to the number of weeks post-randomization when the subject first drank.

A second variable (TIMEVENT) was then created to accommodate the time to censor for those subjects who did not experience an event. If the subject experienced an event (FIRSTDR ≥1), then the variable TIMEVENT was set equal to value of the FIRSTDR week counter. If the subject did not experience an event (was censored), her TIMEVENT was set equal to the number of weeks she remained in the study (her time to censor). In this study, time to event was calculated using the gestational age of each subject (recorded as number of weeks) at randomization and at delivery.

In our dataset, there were a number of instances in which more than one subject relapsed in the same week. Tied data is a concern in survival analysis, and will be addressed briefly in the section describing PROC PHREG.

Creating a variable representing censored data
The variable representing censored data is often denoted by \( \delta \), where \( \delta = 1 \) if an event is observed (uncensored), and 0 if an event is censored. Here, our variable FIRSTDR, time to first drink, is used to create the value of CENS, our censor variable. If a subject did not experience an event (did not drink) prior to the end of follow-up, her CENS variable is set equal to 0. For the purposes of this portion of the study, there were no losses to follow-up.

A number of differences exist between the SAS survival analysis PROCs, notably in terms of distribution and modeling assumptions. Since a complete discussion of these differences is beyond the scope of this report, the reader is encouraged to refer to the References section to select appropriate further reading on each PROC and its assumptions prior to analysis.

To consider possible differences in time to first drink based on a treatment effect or covariate impact, we
selected two SAS PROCs based on the distribution of our data and study needs: (1) PROC LIFETEST, to consider the difference in time to first drink between our two groups based on treatment/control status; and (2) PROC PHREG, to explore the impact of covariates on time to first drink.

**PROC LIFETEST**

PROC LIFETEST can be used to produce survival curves using the Kaplan-Meier method of estimating the survivor function, \( S(t) \) (please see any of the references for a complete description of Kaplan-Meier estimation). Here, PROC LIFETEST is used to graphically display the survivor functions for both treatment and control group, to illuminate possible differences between the groups. The data which we will be examining here is limited to those subjects initially abstinent prior to entering the study.

In the following syntax, created using SAS v6.11 under Windows 3.1, our variable TIMEVENT once again refers to the time to event or censor, and the variable CENS represents the censoring status for each subject, with the adjacent number (0) denoting a censored observation. Use of the option STRATA produces a separate graph for our two groups, treatment (BI=1) and control (BI=0).

```sas
PROC LIFETEST DATA=PH1. MERGE23 PLOTS=(S) GRAPHICS;
TIME TIMEVENT*CENS(0);
STRATA BI;
RUN;
```

**Figure 1. Survival curves for Brief Intervention (Tx) and Assessment only (Co) groups**

The graph produced by PROC LIFETEST reveals a higher survival curve (longer time to first drink) for those subjects in the Brief Intervention treatment group (solid line).

**PROC PHREG**

PROC PHREG utilizes Cox proportional hazards regression in a modeling approach to the analysis of survival data. A model is developed which considers the relationship of the hazard function to selected covariates, including the variable denoting treatment condition (Collett, 1994). Briefly, a proportional hazards model is based upon the assumption that the hazard functions for two groups remain constant across time. SAS provides a number of options for testing this assumption. Following preliminary bivariate analyses, we are able to identify two dichotomous variables--completion of a college education (variable COLLEGE) and treatment condition (variable BI)--which appear to have prognostic importance for the proportional hazards model.

We are also able to explore possible interaction between BI and COLLEGE, as well as to explore possible violations of the proportional hazards assumption, by including interaction terms for each test in the model.

The following code illustrates the MODEL statement used in conjunction with PROC PHREG. The two covariates of interest, COLLEGE and BI, are located on the right side of the equation. The TIES option grants the user the option of selecting the preferred method for handling ties; the options are EXACT and DISCRETE, each of which can be used to provide ordering for tied event times (Allison [1995] discusses the merits of the methods).

```sas
PROC PHREG DATA=PH1.MERGE24;
MODEL TIMEVENT*CENS(0)=COLLEGE BI /TIES=EXACT;
RUN;
```

The results of the Cox regression model for this model (not reproduced here) will not appear unfamiliar for those comfortable with output generated by PROC LOGISTIC. As was noted in the PROC LIFETEST output, the treatment variable BI continues to have an effect on time to first drink using PROC PHREG. The hazard ratio for the variable BI is less than 1, indicating that the hazard of an event (drinking) at a time \( t \) is smaller for an individual receiving the Brief Intervention, relative to an individual receiving an Assessment alone.

**Model checking**

There are numerous options available in conjunction with PROC PHREG, which may be employed depending on
study and/or model-checking needs. Some of these options include:

• Use of an OUTPUT statement in conjunction with PROC PHREG provides the user with the option of generating residual statistics for detection of outliers.

• the ability to include a time dependent covariate in the model to test for violations of the proportional hazards assumption

• the DFBETA statistic, which can used to access the impact of individual observations

In addition, PROC GPLOT or PLOT can be used in conjunction with PHREG, to provide graphs of residuals and predicted values, as required.

CONCLUSION

This report does not begin to penetrate the extensive mathematical background necessary for an in-depth understanding of survival analysis. Despite the somewhat daunting aspect of the underlying mathematical theory, survival analysis can help researchers answer questions which may affect the overall interpretation of study results. The increasing importance of survival analysis techniques outside the medical literature should lead public health researchers and others to investigate their use in non-traditional applications.

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute, Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES


The code and data examples used in this report were taken from a study supported by R01 AA 9670 from the National Institute on Alcohol Abuse and Alcoholism (Dr. Chang).

Key words: statistics, multivariate analysis, survival analysis, proportional hazards model, public health

Margaret Ann Goetz
Data Manager, Alcohol Abuse in Pregnancy Study
Division of Psychiatry/Research
221 Longwood Avenue, Boston, MA 02115
phone 617-278-0530
e-mail MAGOETZ@BICS.BWH.HARVARD.EDU