Survival Analysis

Survival analysis, sometimes called event history analysis, is used for longitudinal data in which the outcome is a binary event (e.g., heart attack, death, purchase of a car, college graduation) and it is possible for the event to occur for some individuals after the study ends. The data are considered "censored" for cases that do not experience the event by the end of the study.[1] Some of these cases may have experienced the event and some may have not experienced the event had the study continued. Ignoring censoring, however, will lead to potentially biased estimates and statistical tests. Thus, it is inappropriate to use standard ordinary least squares regression or logistic regression to predict the occurrence of the event or the time to occurrence of the event.

There are two common approaches to survival analysis, *discrete time* survival analysis, and *Cox regression* (also referred to as "proportional hazard models"), and I will focus on these approaches here. There are several other approaches (e.g., accelerated failure time models, Weibull, log-normal), but they tend to be used less often in the social sciences. Discrete time survival analysis is designed for instances where there are few time points measured and the exact time that the event may not be known. Cox models are intended for the circumstance in which there are many time periods measured or the exact time to the event is known. Both of these models can be adapted for use for the situation in which there are multiple types of events occurring (e.g., stroke vs. heart attack) or repeated events.

The first step in the analysis is often a descriptive analysis that examines how many events occur in each time period. Tables and figures are usually generated to get an overall sense of when the event is most likely to occur. The two most common approaches to these descriptive tables are the *Kaplan-Meier* (Kaplan & Meier, 1958) and the *life-table* approaches. The Kaplan-Meier makes more sense if there are fewer possible time points to the event or the time to event is more discrete (i.e., fewer intervals known). If there are many time intervals, the Kaplan-Meier tables become unwieldy, so the life-table method, which groups event times into a set of intervals, is typically used.

The focus of the tables is usually an examination of the estimated probability of survival based on the proportion of cases that have not experienced the event at each time point or interval. In symbols, we could say:

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

The $S(t)$ is the survival probability, and $\Pr(T > t)$ is the probability that the event T occurs after some time interval or time point *t*. Thus, if the event has not happened T comes after the time point under consideration at the moment.

The probability of failure (i.e., probability that the event occurs) is simply the converse of the probability of survival, where *F(t)* is the probability failure has occurred up to time *t*:

$$F(t) = 1 - S(t)$$

The hazard probability is the probability that the event has occurred during a certain interval (or at a particular point in time) given that it has not occurred yet. It is, therefore, a type of conditional probability. In symbols, we could say:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

[1] The data are right censored in this case. It is also possible for data to be left censored or interval censored, but I will not discuss these circumstances here.

The numerator of the right side of the equation is a conditional probability that the event T has occurred between some initial time point, t, and some later time point, $t + \Delta t$. Δt represents some increment in time, which can be any amount. The right side of the | symbol represents "given that the event has not occurred before time t." The limit $(\lim)$ is a way of indicating we are computing the probability as the time increment becomes incrementally small. So, the entire right side can be read as the probability that an event occurs in a certain time interval, regardless of its length, given that it has not already occurred to that point.

We can also say the hazard probability is the probability of failure divided by the probability of survival:

$$h(t) = \frac{f(t)}{S(t)}$$

The hazard probability and the survival probability are opposite concepts and they are linked by a log transformation.

$$h(t) = \frac{d}{dt} \log S(t)$$

So, the hazard function is the derivative with respect to *t* of the log transformation of the survival function. In more general terms, as the probability of the event increases (hazard), the probability of survival decreases for any time interval.

Kaplan-Meier Estimation Example

To illustrate hazard and survival estimates, I'm going to use the example from Graham, Willet, and Singer (2012). This example comes from the Wisconsin Longitudinal Study and models the time until divorce for those who marry. The event in this case is divorce, where the hazard represents the probability of divorce during a certain period and survival represents the probability of continued marriage up to and including the period considered. There are 11 periods representing 4-year intervals each, so the time spans 44 years. Data are censored because some couples may divorce after the 44 years.

Below is SAS code for requesting Kaplan-Meier estimates of hazard and survival probabilities at each period (`METHOD=KM`). `METHOD=LT` would produce life-table estimates, which are computed slightly differently. The `INTERVALS` statement requests estimates for each of the 11 periods. The `TIME PERIOD*Y(0);` command specifies the variables used as the time variable, which is period here, and the variable used as the indicator of whether or not the event occurred, which is Y. The value in parentheses, (0), indicates the value of the outcome that corresponds to censoring (i.e., the value representing a non-occurrence of the event). (Values appear in Table 11.1 in the Graham et al., 2012).

```
proc lifetest data=person method=km plots=(s,h)
 intervals = 1 2 3 4 5 6 7 8 9 10 11;
  time period*y(0);
  run;
```

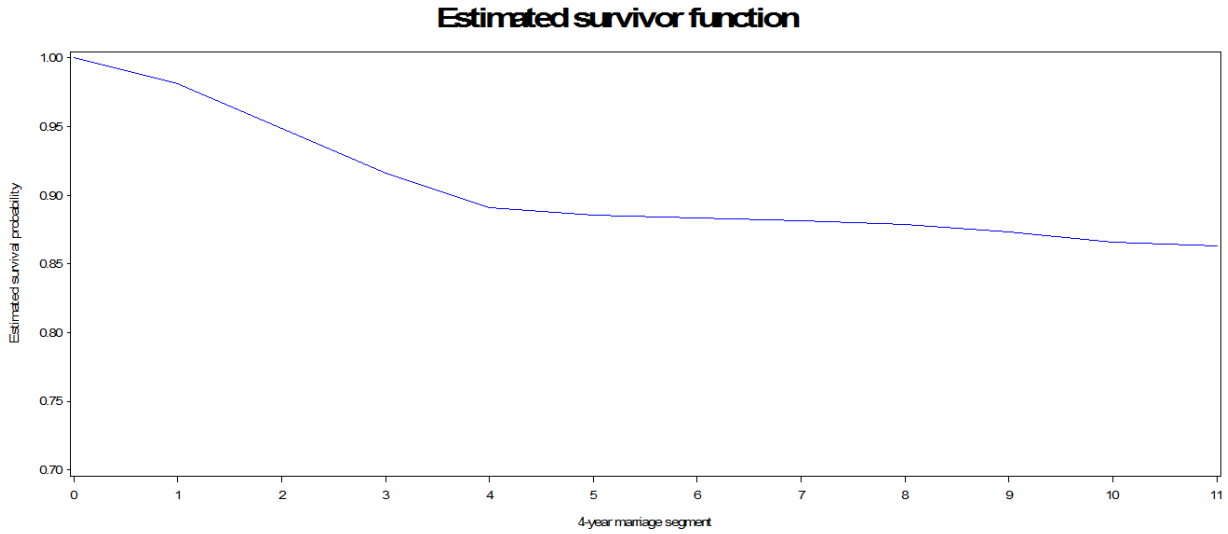
```
The LIFETEST Procedure
               Product-Limit Survival Estimates
                                    Survival
                                    Standard    Number   Number
  period     Survival    Failure     Error       Failed     Left
  0.0000       1.0000           0         0            0     7860
  0.0000       0.9809      0.0191   0.00154          150     7710
  1.0000       0.9487      0.0513   0.00249          403     7451
  2.0000       0.9157      0.0843   0.00314          661     7167
  3.0000       0.8912      0.1088   0.00352          851     6903
  4.0000       0.8855      0.1145   0.00361          894     6656
  5.0000       0.8831      0.1169   0.00364          911     6422
  6.0000       0.8815      0.1185   0.00367          923     6302
  7.0000       0.8786      0.1214   0.00371          943     6073
  8.0000       0.8730      0.1270   0.00380          979     5655
  9.0000       0.8655      0.1345   0.00394         1022     4973
 10.0000       0.8630      0.1370   0.00399         1034     4141
 11.0000*           .           .         .         1034        0
```

In SAS, the KM method does not generate the hazard probabilities (the life-table method does), but that could be computed for any interval by dividing the difference in failure by the survival. For example, for the 3rd period, there were 851 – 661 = 190 new divorces, and the increment in failure is .1088 - .0843 = .0245. This gives a hazard for this period of:

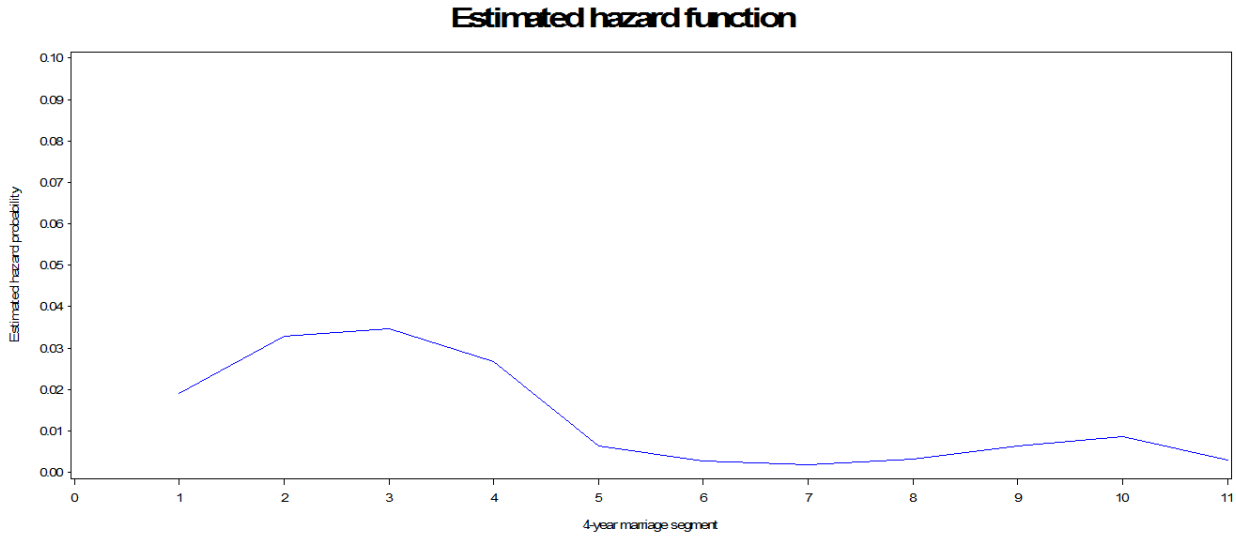
$$h(t) = \frac{f(t)}{S(t)} = \frac{.0245}{.8912} = .0275,$$

which is within rounding error of the value in Table 11.1 in the Graham et al. chapter for this period.

The PLOTS= request produced the following two plots. The survivor function plot is a plot of the survival probabilities, or the probability of staying married over the 44 years. The plot shows a decline in survival over time, but there is a steeper decline in the first 4 periods (16 years).



The hazard function plot suggests that there is a considerably higher risk of divorce between the second and fourth periods (about 8 to 16 years after getting married).



Discrete Time Survival Analysis

Discrete time survival analysis (Cox, 1972) is intended for analysis of the probability of an event occurring when the time variable is discretely measured. In other words, there are few periods to time measured and the exact time of the event is not known. For example, a study might survey respondents annually over 6 years and ask them about a particular event at each interview. In this case we often do not know the exact time of the event, but we know the event occurred between Years 2 and 3 or Years 4 and 5 etc. So, we have only 5 possible time periods within which the event may have occurred.

Discrete time survival models can be tested using traditional logistic regression (but really based on a multinomial distribution), but several special steps must be taken to obtain the correct estimates that can be interpreted as survival estimates. First, the data must be reconfigured into a person-period format (a.k.a., long format). Second, dummy or indicator variables must be constructed to designate whether the event occurred during that period for each person.

Discrete Time Survival Analysis Example

An initial analysis can be conducted without predictors to obtain hazard estimates at each period. To do this, a logistic model is tested using the person-period (long) format and entering in the list of dummy variables. In the Wisconsin Longitudinal Study there were 11 periods, so 11 dummy variables (D1-D11) are constructed. The odds ratios for each dummy estimate the hazard at each time point. The values closely approximate those obtained from the Kaplan-Meier procedure.

```
proc logistic data=persper out=estimate;
  title "Fitted initial hazard model with time indicators";
  model y(event='1') = D1-D11/noint;
  run;
```

```
The LOGISTIC Procedure
             Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
D1            1     -3.9396      0.0824     2283.6827        <.0001
D2            1     -3.3827      0.0639     2799.9442        <.0001
D3            1     -3.3243      0.0634     2752.0517        <.0001
D4            1     -3.5927      0.0735     2386.7139        <.0001
D5            1     -5.0421      0.1530     1086.1510        <.0001
D6            1     -5.9343      0.2429      597.0848        <.0001
D7            1     -6.2637      0.2889      469.9227        <.0001
D8            1     -5.7159      0.2240      651.2799        <.0001
D9            1     -5.0568      0.1672      914.7323        <.0001
D10           1     -4.7506      0.1532      962.1048        <.0001
D11           1     -5.8438      0.2891      408.6141        <.0001

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
D1           0.019       0.017       0.023
D2           0.034       0.030       0.038
D3           0.036       0.032       0.041
D4           0.028       0.024       0.032
D5           0.006       0.005       0.009
D6           0.003       0.002       0.004
D7           0.002       0.001       0.003
D8           0.003       0.002       0.005
D9           0.006       0.005       0.009
D10          0.009       0.006       0.012
D11          0.003       0.002       0.005
```

One or more predictor variables can be included in the analysis to predict the probability of experiencing the event. In the WLS data, we can add an education variable (DEGREE) to see whether those who are more educated are more or less likely to get divorced.

```
proc logistic data=persper out=estimate;
  title "Table 3, Model B";
  model y(event='1') = D1-D11 degree / noint;
  run;
```

```
Analysis of Maximum Likelihood Estimates

                              Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
D1            1     -3.8764      0.0831     2176.1222        <.0001
D2            1     -3.3189      0.0648     2623.3324        <.0001
D3            1     -3.2603      0.0643     2574.8627        <.0001
D4            1     -3.5285      0.0743     2255.1477        <.0001
D5            1     -4.9761      0.1534     1052.6065        <.0001
D6            1     -5.8671      0.2431      582.4403        <.0001
D7            1     -6.1969      0.2892      459.2988        <.0001
D8            1     -5.6499      0.2242      634.8448        <.0001
D9            1     -4.9919      0.1675      887.7818        <.0001
D10           1     -4.6873      0.1535      932.2916        <.0001
D11           1     -5.7907      0.2892      400.8479        <.0001
degree        1     -0.5164      0.1057       23.8795        <.0001

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
D1           0.021       0.018       0.024
D2           0.036       0.032       0.041
D3           0.038       0.034       0.044
D4           0.029       0.025       0.034
D5           0.007       0.005       0.009
D6           0.003       0.002       0.005
D7           0.002       0.001       0.004
D8           0.004       0.002       0.005
D9           0.007       0.005       0.009
D10          0.009       0.007       0.012
D11          0.003       0.002       0.005
degree       0.597       0.485       0.734
```

The results indicate that having a degree is negatively related to (B = -.5164, p < .0001) divorce. Those with higher education are less likely to get divorced. The hazard ratio is interpreted like an odds ratio, and its value here (HR = .597) represents approximately 68% reduction in the likelihood of getting divorced (1/.597 = 1.68).

SPSS

The discrete survival model can be tested in SPSS using the LOGISTIC procedure, but the /noorigin subcommand is needed to test a model without the intercept.

R

In R, the glm function can be used with a + 0 term and intercept=FALSE. (I gave … instead of listing all of the dummies here).

```
glm(y~d1 + d2 + … + d11 + 0, family=binomial(link="logit"), intercept=FALSE)
```

Cox Regression Survival Analysis

Cox regression (Cox, 1972) is used for the circumstance in which there is information about the time until the event occurs in addition to information about whether or not the event occurred.

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t \le \mathcal{T} < t + \Delta t \mid \mathcal{T} \ge t)}{\Delta t}$$

The hazard function, $h(t)$, is a representation of the risk or rate of the events occurrence within a certain interval. Δt is defined in terms of an increment in continuous time that can be considered in ever decreasing quantities to a lower limit of 0, which the $\lim_{\Delta t \to 0}$ notation on the right of the equation refers to. As with the discrete hazard, the rate is conditional on the event having not previously occurred—that $\mathcal{T} \ge t$.

The hazard, as defined above, is then modeled in a transformed linear regression equation:

$$h_i = \lambda_0(t)\exp(b_1 x_{i1} + ...b_k x_{ik})$$

The $\lambda_0(t)$ is the intercept or baseline hazard and exp is the exponent function (constant *e* raised to some power). The model can then be "linearized" with log transformation in into the following form:

$$\log h_i(t) = b_0(t) + b_1 x_{i1} + ... + b_{ik} x_{ik}$$

And this equation looks very much like the standard regression equation on the right-hand side.

The Cox regression model is convenient for several reasons. There is no underlying probability distribution of survival assumed by the model (it is considered "semiparametric"). Time values do not have to be exact, and Cox can produce good estimates of hazard probability even with relatively few or inexact time points (Allison, 1984; Thompson, 1977). So, it remains a very flexible approach.

Cox regression then proceeds very much like linear or logistic regression with information about overall model fit and whether predictors are significant. Like logistic regression, the slope coefficients are not particularly meaningful because of the log transformation of the hazard. So, results can also be expressed in terms of *hazard ratios*. A hazard ratio represents the increased risk of the event occurring at any given time for each unit increase in the predictor. Like odds ratios in logistic regression, a hazard ratio of 1.0 represents even odds or not increased or decreased risk. Hazard ratios over 1.0 indicate increased risk, and hazard ratios below 1.0 indicate reduced risk.

The analysis requires a variable for censoring (event or no event by the end of the study), a numeric variable for the time until the event occurs, and the analysis uses standard data format of one person per record (wide format). The individuals that do not experience the event are given a time score equal to the last possible time value.

Cox Regression Example

Cox models can be tested in SAS, SPSS, or R, and other general statistical software. I tend to use SAS because it has more options for survival analysis. Two special variables are needed for the analysis. The first variable is simply whether or not the event occurred, often referred to as the "censor" variable. The second variable is the time variable, representing time until the event occurs. For cases in which the event has not occurred by the end of the study, their values are set to the maximum time possible for the study. If months are the time metric and the study ended after 24 months, individuals who did not participate in the study are given a value of 24.

Using the WLS data, we can estimate the risk of divorce for those with higher vs. lower education level. Here, the Cox regression treats the variable LENGTH (same as our PERIOD variable before) as a continuous time variable. The TIES=EXACT option specifies one way of handling tied time-to-event values, which is preferable if available (Allison, 2010). TIES=EFRON and TIES=EXACT tend to produce better estimates , in general, than the default (TIES=BRESLOW).

```
proc phreg;
  model length*censor(1)=degree / ties=exact;
  baseline out=output1 covariates=values1 survival=surviv1 / nomean;
  run;
```

```
Model Fit Statistics
                 Without           With
Criterion     Covariates     Covariates

-2 LOG L        9932.752       9905.413
AIC             9932.752       9907.413
SBC             9932.752       9912.354

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        27.3392        1         <.0001
Score                   24.3562        1         <.0001
Wald                    23.8321        1         <.0001

               Analysis of Maximum Likelihood Estimates

                     Parameter      Standard
Parameter    DF       Estimate         Error    Chi-Square    Pr > ChiSq

degree        1      -0.51140       0.10476       23.8321        <.0001

Parameter       Ratio    Label
degree          0.600    received college degree (1=yes, 0=no)
```

Results give a similar hazard ratio to that obtained with the discrete time survival model, indicating a significant effect of degree such that those with a college degree are about 67% (1/.60 = 1.67) less like to get divorced (HR = .60, $p < .0001$).

SPSS

Specify the above model in SPSS with length as the time to event variable and censor(0) specifying the event (opposite of SAS).

```
Coxreg length with degree
  /status censor(0).
```

R

The survival package is a convenient way to specify Cox regression in R. Create the censor variable such that 1 = censored and 2 = event.

```
library(survival)
coxmod <- coxph(Surv(length, censor) ~ degree, data = d)
```

Comments on Survival Analysis

The results from the discrete time survival model and the Cox regression were quite similar, so why chose one over the other? With many time points (and, thus, fewer tied time values), the two approaches will give very similar results. With finer grained time intervals, the discrete survival estimates converge with the proportional hazard (Cox regression) results (Thompson, 1977). Categorizing otherwise continuous information (e.g., exact days to the event are grouped into yearly intervals) is not advisable, however, because there is a loss of precision proportionate to the coarseness of the categorization. Interestingly, implementation of the analyses can produce results from the alternative method depending on the specifications. In SAS, for example, with PROC LOGISTIC, specifying the complementary log-log link, gives proportional hazards estimates (Cox regression) and specifying TIES=DISCRETE under PROC PHREG (Cox regression procedure) gives discrete time survival results. The discrete method assumes the event for ties occur at the same time, but the Cox regression approach assumes that there is a true underlying ordering of the time of the event that is unknown.

Both discrete and Cox regression produce hazard ratios and both can incorporate time varying and time invariant predictors. For either method, interactions can also be incorporated to investigate whether the covariate has a differential effect at different periods. So, with many time points there are few important differences and most researchers probably choose Cox regression, because it is somewhat more convenient (e.g.., dummy variables do not need to be computed). Cox regression does assume that the levels of the predictor have a similar (parallel) hazard over the time periods. Thus if the hazard is plotted for two levels of a predictor variable (e.g., no degree vs. degree) the shape of the hazard lines should be approximately parallel. This assumption can be addressed by incorporating interactions, however. With very few time points, it might be preferable to use discrete time survival, because Cox will likely become less accurate as the number of time intervals decreases (and, thus, number of ties increases).

One can also examine competing risks when there is more than one type of event. For example, one may want to compare heart attack with congestive heart failure. Methods also exist for repeated events (e.g., multiple heart attacks).

The data circumstance that we have been considering with survival analysis is that of right censoring, where we do not know whether the outcome occurs after the study ends. Left censoring, in which some cases may have already experienced the event prior to the start of the study, is also a potential biasing issue. Although left censoring is less often an issue, there are special analytic precautions that should be taken. Interval censoring can also occur, a type of right and left censoring together. Typically this occurs in a panel study with regular intervals when the event occurs between the intervals but the precise timing of the event is unknown. Most often researchers will analyze these data with the discrete time survival approach. When there are irregular intervals more complex methods may be needed (e.g., J-S. Kim, 2003).

Further readings

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data.* Newbury Park, NJ: Sage.
Allison, P.D. (2010). *Survival Analysis Using SAS: A Practical Guide, 2nd edition*: SAS Institute, Cary, NC.
Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, 34*, 187–220.
Graham, S. E., Willett, J. B., & Singer, J. D. (2012). Using discrete-time survival analysis to study event occurrence. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging health, and social sciences*. New York: Routledge.
Hosmer, D., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis (2nd Edition.).* Hoboken, NJ: Wiley Series
Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.
Thompson Jr, W. A. (1977). On the treatment of grouped observations in life studies. *Biometics, 33*, 463–470.