S02 - 2008
Survival Methods for Correlated Time-to-Event Data

James Bena, Cleveland Clinic, Cleveland, OH
Shannon McIntyre, Cleveland Clinic, Cleveland, OH

ABSTRACT
The use of product-limit (Kaplan-Meier) estimation and Cox proportional hazards modeling is common when analyzing time-to-event data, especially in the presence of censoring. Presenting results from both methods provides the magnitude of loss within the levels of a given variable, and a relative measure of failure risk between the levels. However, since both above methods assume independence of the observations, correlated measurements require adjustment to avoid underestimation of the variance, and overestimation of the statistical significance.

In this paper, we focus on the case of clustered results, where a single observed unit may have several unique observations. The variances of Kaplan-Meier estimates from PROC LIFETEST are adjusted for the clustering using Taylor-series approximation. The standard errors of estimated hazard ratios from Cox proportional hazards models fit using PROC TPHREG are altered using the sandwich estimator, effectively fitting a marginal model.

A SAS macro is described that performs both of these adjusted analyses, and then creates a table displaying the Kaplan-Meier survival estimates at specified time points and hazard ratios from the marginal Cox proportional hazards model.

INTRODUCTION
Often biomedical studies look at the time duration until some event occurs. When all subjects experience the event, the time can be evaluated like any other continuous measure. However, if not all patients experience the event, methods that account for censoring must be used. Typical methods include product-limit (Kaplan-Meier) estimation and Cox proportional hazards modeling. An added complexity occurs when not all of the observations are independent. Examples include repeated failures within the same person as well as different types of events within the same subject. Often subjects may have multiple interventions of the same type at the same time point. While each observation is unique, and has its own failure or censor time, certain subject characteristics may be systemic and affect all observations from the subject. The goal of this paper is to explain the adjustments that are required to the traditional methods when these clustered data occur. A macro that performs these adjustments and outputs tabular summaries is described.

CALCULATING SURVIVAL ESTIMATES
Typically, when each subject has just one observation, SAS/STAT possess the necessary tools to calculate the nonparametric estimates based on the Kaplan-Meier method. Unfortunately, when multiple observations per subject exist, standard errors of the survival estimates underestimate the true amount of variability that exists. Using a method that employs Taylor series linear approximations, and relying on the macro language and data step processing, the necessary adjustments can be made.

USING PROC LIFETEST
Kaplan-Meier estimates of survival are easily calculated using PROC LIFETEST. While there are a variety of options available in the procedure, the call used here is quite simple. Beyond defining the variable that contains the time to an event or censoring, and another that contains an indicator value of whether the time point is an event or not within the TIME statement, the only additional command that is needed is a STRATA statement that contains the variable that is being tested for association. Below is an example of the procedure evaluating differences in a time variable based on gender. Note that the number in parentheses indicates the censoring values of the variable that indicates whether an event has occurred. The ODS statements are used to store the survival estimates in a separate dataset. This is vital to allow for adjustment of the standard errors.

```
proc lifetest data=InData;
  ods output Estimates=ParmEst;
  time TimeVal*PatCens(1);
  strata Gender;
run;
```

ADJUSTING VARIANCE ESTIMATES
Once the Kaplan-Meier estimates are obtained, adjustments of the standard errors are necessary to correct for the correlation between observations from the same subject. Williams (1995) described a method based upon Taylor series approximations of the survival estimates for each observation. He then applied the between-cluster variance estimator that is used often in multi-stage surveys.

Using the outputted dataset of estimates and standard errors from the procedure call above, macro variables containing unique time points, survival estimates, numbers of failures, and numbers at risk at each time point are created. Then linear approximations of the survival function are created for each observation and time point, and then summed within subject and

time. The calculation of the linear survival function for each observation is created using data step processing. Summaries by subject are calculated using PROC MEANS. Finally, the variance estimate between subjects is calculated using a between-cluster variance estimator. Once the adjusted standard errors for each time point are derived, confidence intervals for the estimates can be created.

CALCULATING HAZARD RATIOS

Compared to the number of steps required to calculate adjusted standard errors from the survival function, the adjustments made to the Cox proportional hazards model are minor and easily implemented. The adjustments performed calculate a sandwich estimator of the standard error and the model fit is referred to as a marginal model. This model was described by Lin (1994).

USING PROC TPHREG

While many of the same calculations could be performed using PROC PHREG, the experimental procedure TPHREG was used because of its added functionality by allowing class variables. The syntax used to run PROC TPHREG uses all of the same variables used in the PROC LIFETEST call above. However, the syntax used is more similar to other modeling procedures, such as PROC REG, PROC GLM, or PROC GENMOD. Usually, the PROC TPHREG call requires only the procedure and model statements. In order to calculate the sandwich estimator for the variance of the hazard ratio, the phrase `covsandwich(aggregate)` is added to the end of the procedure statement, as is an ID statement, which identifies observations from the same subject. An example call of the procedure is shown below, again using Gender as the class variable of interest. The ODS statement is used to capture the hazard ratio, confidence limits, and p-values from the Cox proportional hazards model.

```
proc tphreg data=InData covsandwich(aggregate);
  ods output ParameterEstimates=Estimate GlobalTests=Overall;
  model TimeVal*PatCens(1)=Gender;
  id PatID;
run;
```

MACRO CREATION AND OUTPUT

Since there are often several variables of interest as potential predictors of risk to investigators, and the same set of actions is required of each variable, a macro that performs each of the steps for several variables, and then combines the results was created. The macro provides a limited number of options for the end user, including an ability to define time points where estimates of survival are printed, as well as the ability to identify an output destination for the table of results. A marginal model option can be used to specify whether correlated data exist. If marginal is set to true, then the above calculations are performed. Otherwise, the procedures are run assuming independence across all observations.

MACRO CALL

The call to the macro is very straight forward and acts mainly as a way to identify the variables of interest for the procedures run within the macro. The ability to identify a destination, or multiple destinations, for the final table is also included. As a convenience for the user, the result dataset from the previous run of the macro within same session can be saved. This allows the user to append the results from the current call to those of a previous call. Note that unless the time points used are the same as the previous call, extra columns with new time points will be created. The variables to be tested for association with the outcome are entered as one string. Each is evaluated to ensure that it exists in the dataset provided, and then a large loop is created that performs the required analyses for each of the variables one at a time, using the syntax described by Carpenter (2004). Below is a sample call of the macro. In this example, a marginal model is requested, using a time variable named TimeVal, and a censoring variable named PatCens with censoring value of 1. The subject id is captured in the variable PatID.

```
%OrgKMRes(ds=InData, Idvar=PatID, Marginal=T, TimeVar=TimeVal, EventVar=PatCens,
CensVal=1, vlist=Gender DM, OutSet=Temp, DeleteOutputDS=T, Times=30 180 365,
ListingFile=T, HTMLFile=T, HTML_Out="&BASEDIR/Temp.html");
```

OUTPUT

In the macro call, the user can specify as many time points as they wish, provided that they fall within the range of the data. The first row of the resulting table that is created by the macro show estimates of survival overall. Subsequent sets of rows show one row for each level. For each of the sets of rows, one column contains the variable name, a second provides the variable levels, and a third shows the number of observations within each level. Then one column of survival estimates and 95% confidence limits for each time point listed in the macro call are created. Following the requested survival estimates, the hazard ratio and confidence limits are listed. For each factor, two sets of p-values are created. The first column of p-values shows the significance of each variable level relative to the first formatted level of that variable, while the last column contains a single p-value that measures the overall significance of that variable.

CONCLUSIONS

The calculations described above allow the user to see the results for several variables in a single call. A table of results is created that can be shared with researchers and is appropriate for inclusion in published manuscripts. Other diagnostic tests

should be performed to assess whether the assumptions for each of the methods used are met. However, the macro is a convenient preliminary tool for understanding the relationship between a set of variables with the outcome of interest.

REFERENCES

Carpenter A (2004), ***Carpenter's Complete Guide to the SAS® Macro Language***, Cary, NC: SAS Institute Inc.

Lin DY (1994), "Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach," ***Statistics in Medicine,*** 13, 2233-2247.

Williams RL (1995), "Product-Limit Survival Functions with Correlated Survival Times," ***Lifetime Data Analysis*** 1, 171-186.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Matthew Karafa for his guidance on macro programming, output creation, error checking, and review of an earlier draft of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

James Bena
Department of Quantitative Health Sciences
Cleveland Clinic
9500 Euclid Avenue/ JJN3-01
Cleveland, OH 44195
benaj@ccf.org