Calculating the Probability of Hospitalization as a Function of Time Using the SAS® System's PROC PHREG

Tyler C. Smith, Naval Health Research Center, San Diego, CA

ABSTRACT

Objective To analyze the differences between hospitalizations for two cohorts with incomplete data using Cox's proportional hazard modeling with censoring. This will ultimately allow for the creation of probability of hospitalization curves by cohort.

Methods Using PROC PHREG, two cohorts were compared for a specific hospitalization outcome over the time period of March 10, 1991 through September 30, 1995. An individual contributed to the Cox proportional hazards model until time of event, or until censoring at time of withdrawal or termination of study. The baseline option in PROC PHREG was used to create and output the survival estimates, which are a function of the probability estimates over time.

Results Cox modeling showed no increased risk for hospitalization among subjects in the exposed cohort when compared to the nonexposed cohort.

Conclusions The SAS System's PROC PHREG with baseline option is a powerful tool for dealing with attrition of subjects over a study period and for producing probability of event curves as a function of time.

Introduction

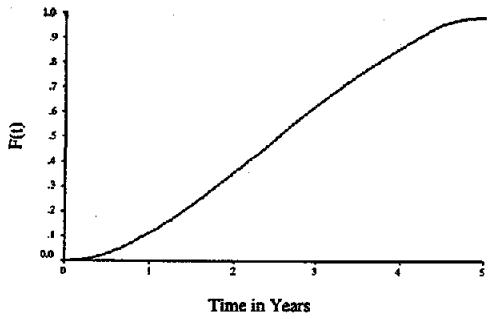
The term survival analysis pertains to a branch of statistics designed for studying the time between entry into a study and a subsequent event, traditionally death. Although time until death was the original outcome for survival analyses, the uses have grown to include outcomes such as time until onset of disease, stockmarket crash, equipment failure, earthquake, etc. These events are best defined as simply a transition from one discrete state to another at an instantaneous moment in time.

Survival analysis may have its origins in mortality tables from centuries ago, but it was not until World War II that a new era of survival analysis emerged. This new era was stimulated by interest in reliability (or time until failure) of military equipment. At the end of the war these statistical ideas, which had worked so well in creating more reliable weaponry, quickly spread through private industry as customers became more demanding of safer, reliable products. As survival analysis adapted to new uses, parametric models gave way to nonparametric models for their appeal in dealing with the ever-growing field of clinical trials in medical research. Survival analysis was well suited to such work, because the studies could start without all of the experimental units and could end before all of the experimental units had experienced an event. This was important in allowing each experimental unit to contribute to the model only for the amount of time it was followed.

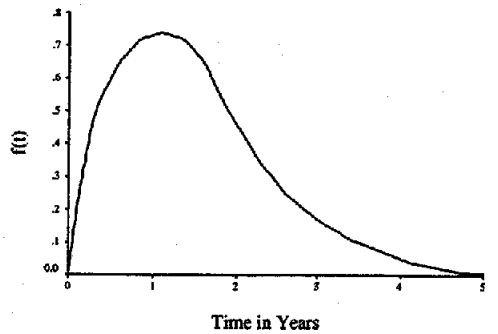
Probability Density Function and Cumulative Distribution Functions

Since time is continuous, and the probability of an event at a single point of a continuous distribution is zero, it is challenging to define the probability of events over time. This relationship is best described by graphing the distributions of event times. The probability density function (pdf) and cumulative distribution functions (cdf) are very useful tools in describing the continuous probability distribution of a random variable such as time in a survival analysis.

The cdf of a random variable T, denoted $F_T(t)$, is defined by $F_T(t) = P_T(T \leq t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function F(t) can be listed as a consequence of the knowledge of probabilities. Because F(t) is a probability $0 \leq F(t) \leq 1$, F(t) is a nondecreasing function of t, and as t approaches ∞, F(t) approaches 1.



The pdf of a random variable T, denoted $f_T(t)$, is defined by $f_T(t) = d F_T(t) / dt$. That is, the pdf is the derivative or slope of the cdf. Every continuous random variable has its own density function, the probability $P(a \leq T \leq b)$ is the area under the curve between times a and b.



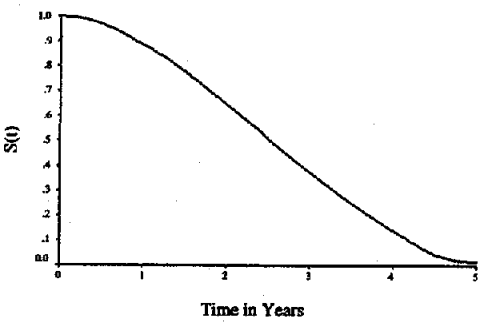
The Survival Function

Let $T \geq 0$ have a pdf f(t) and cdf F(t). Then the survival function takes on the following form:

$$S(t) = P\{T > t\}$$

$$= 1 - F(t)$$

That is, the survival function gives the probability of surviving beyond time t. Because S(t) is a probability it is positive and ranges from 0 to 1. It is defined as $S(0) = 1$ and as t approaches ∞, S(t) approaches 0. The Kaplan-Meier estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times.



The Hazard Function

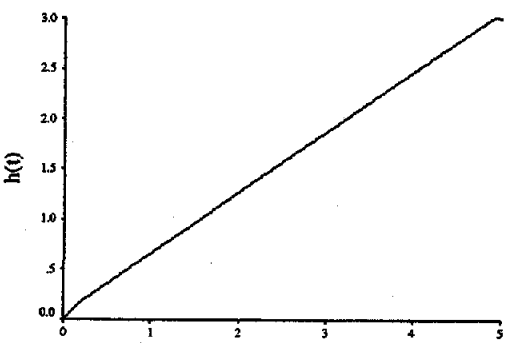
The hazard function h(t) is given by the following:

$$h(t) = P\{t < T < (t + \Delta) \mid T > t\}$$

$$= f(t) / (1 - F(t))$$

$$= f(t) / S(t)$$

The hazard function, describes the concept of the risk of an outcome (e.g. death, failure, hospitalization) in an interval after time t, conditional on the subject having survived to time t. It is the probability that an individual dies somewhere between t and $t + \Delta$, divided by the probability that the individual survived beyond time t. The hazard function seems to be more intuitive to use in survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t.



Censoring

The observation of survival time has two components which must be carefully defined in any survival analysis. There is a beginning point (t=0) and a reason or cause for the observation of time to end. For example, in a complete observation cancer study, observation of survival time may begin on the day a subject is diagnosed with the cancer and end when that subject dies as a result of the cancer. Complete observation data is desired but is not realistic in most studies. There is a good possibility that the patient might recover and live for 30 more years or the patient might die due to another cause. In other words, the study cannot go on indefinitely and unforeseen things happen to subjects. The censoring of study participants deals with situations where there are incomplete observations of time due to assumed random natural factors not related to the study design. This differs from truncation where observations of time are incomplete due to a selection process inherent to the study design.

The most common form of incomplete data is right censoring. This occurs when there is a defined time (t=0) where the observation of time is started for all subjects involved in the study. A right censored subjects time terminates before the outcome of interest is observed. For example, a subject could move out of town, could die of an unexpected reason, or could simply choose not to participate in the study any longer. Right censoring techniques allow subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal), or they have an event. Conversely, an observation is left censored if the event of interest has already occurred when observation of time begins. For the purposes of this study we focused on right censoring.

The Study

The objective of this study was to compare hospitalization experiences of a population stratified by exposure to an environmental factor. The study data consisted of an exposed cohort (cohort A; n=124,487) and an unexposed cohort (cohort B; n=224,804).

Descriptive data available for analysis included unique identifiers, gender, date of birth, race, ethnicity, home of record, marital status, occupation, military pay grade, length of military service, salary, and military service branch.

Hospitalization data were captured from all hospitals for the period of March 10, 1991 through September 30, 1995. These data included date of admission and up to eight diagnoses associated with the visit to the hospital. Additionally, a pre exposure period covariate (coded as yes or no) was used to reflect a hospital admission during the 12 months prior to the start of the exposure period. Diagnoses were coded according to the *International Classification of Diseases, Ninth Revision*, (ICD-9). For these analyses, we ignored the decimal component of the ICD-9 diagnoses and considered diagnoses with the same whole number (up to 3 digits) to be equivalent.

With a focus on exposure or lack of exposure, the study outcomes were defined to be "any cause" hospitalization and hospitalization with a diagnosis in each of 15 broad ICD-9 diagnostic categories. The aggregation of many diagnoses into large ICD-9 categories might mask population risk differences due to individual diagnoses. We therefore chose to also examine specific ICD-9 diagnoses, suggested by an expert panel, as possible manifestations of the exposure. For each subject, hospitalizations (if any) were scanned in chronological order and diagnostic fields were scanned in numerical order for the ICD-9 codes of interest. Only the first hospitalization meeting the outcome criteria was counted for each subject.

Analysis

Our study had a set start date when every subject's follow-up time began. Follow-up time was calculated from March 10, 1991 until hospitalization, withdrawal from study, or September 30, 1995, whichever came first. Subjects were allowed to leave the study and assumed a random early departure distribution. Delayed entry, and events occurring before the start date of the study were not a concern in this study, therefore only right censoring was needed to allow for the early departure of subjects.

Cox's proportional hazards modeling was chosen to explain the effect of covariates on survival times for the following reasons:

1) The Cox model has a simple interpretation as a "relative risk" type ratio. For example, when we have a two level covariate with a value of 0 or 1, the

hazard ratio becomes e^{β}. If the value of the coefficient is $\beta = \ln(3)$ then it is simply saying that the subjects labeled with a 1 are three times more likely to have an event than the subjects labeled with a 0.

2) The distribution of survival times can be described either by specifying the density function of a parametric distribution or by specifying the hazard function. Cox semiparametric modeling allows for no assumptions to be made about the parametric distribution of the survival times (making the method considerably more robust) instead making the assumption about the proportional hazards over time. The proportional hazards assumption refers to the fact that the hazard functions are multiplicatively related. That is, their ratio is assumed constant over survival time. Specifying the hazard function instead of specifying the density function was desirable for our study so that we could directly address the aging process over time. Through the analysis we get parameter estimates that will allow us to compare the survival experience of the two exposure cohorts of study.

3) With the use of the partial likelihood function, the Cox model has the flexibility to introduce time-dependent explanatory variables and handle censoring of the survival times. This was important to our study in that any temporal biases due to differences in hospitalization practices for different strata of the significant covariates over the years of study would need to be handled correctly. This ensured that any differences in hospitalization experiences between the exposed and non-exposed would not be coming from these temporal differences.

4) With the SAS option BASELINE, a SAS data set containing survival function estimates can be created and output. These estimates correspond to the means of the explanatory variables for each stratum.

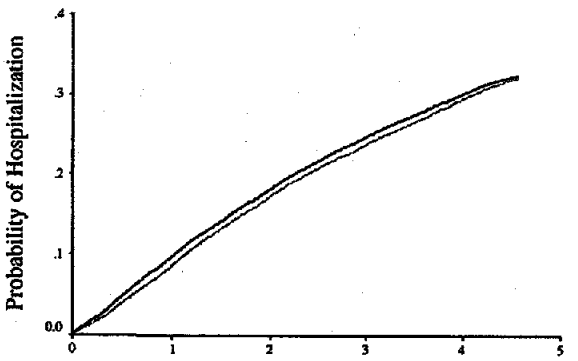
Using PROC FREQ, an initial univariate analysis of the demographic variables crossed with hospitalization experience was carried out to determine possible significant explanatory variables to be included in the model runs. All variables with a chi-square value of .15 or less were considered possibly significant and were therefore entered into the model analysis.

Using PROC PHREG, a saturated Cox model was run after creating dummy variables, necessary for the output of hazard ratios, for the categorical explanatory variables. A manual backward stepwise analysis was employed to create a model with statistically significant effects of explanatory variables on survival times.

After the final model of significant explanatory variables was created, the proportional hazards assumption was validated by adding a time-dependent covariate representing the interaction of the original covariate (the covariate most likely to change over time) and time. These interaction terms were insignificant, therefore it was concluded that the proportional hazards assumption held. This was necessary to ensure that there was no adverse effect from time-dependent covariates creating different rates for different subjects thus making the ratios of their hazards non constant.

The data were then stratified by exposure (cohort A and cohort B) and the models were run with the exposure flag covariate withdrawn from the model. This allowed for inspection of interaction between exposure status and covariates. These separate models also allowed for the computation of survival function estimates using the BASELINE function in PROC PHREG. The survival curves (step functions) were now available to graph for both cohort A and cohort B.

A simple calculation of 1-S produced the cumulative distribution function. That is, we now had the probability estimates of hospitalization over time. These could then be graphed and visually scanned for differences in hospitalization experiences of the two cohorts.



Results

Based upon univariate comparisons for hospitalizations occurring during the study, the following covariates were selected for further modeling: gender, age group, marital status,

race/ethnicity, occupational category, military pay grade, salary, service type, pre exposure period hospitalization, and exposure. Home of record and length of service covariates were not important to the model. Salary was dropped from analyses due to colinearity with age group. The exposed subjects had similar risks of any cause hospitalization during the March 10, 1991 to September 30, 1995 compared to nonexposed subjects. The corresponding cumulative probability plots (above) were nearly parallel for the 54 months of follow-up. However, the Cox model did reveal some better predictors of post exposure hospitalization, which included female gender (RR=2.63), pre exposure period hospitalization (RR=1.65), enlisted pay grade (RR=1.5), and US Reserve service type (RR=1.33).

The Cox modeling for each of the 15 diagnostic categories and the specific diagnoses associated with the possible manifestations of the exposure over the 54 month follow-up period similarly showed no increased risk for hospitalization among subjects in the exposure cohort.

Conclusions

Using SAS's PROC PHREG, Cox's proportional hazards modeling was performed to compare the hospitalization experience of subjects exposed to an environmental factor with subjects not exposed to this environmental factor. Our analyses included broad outcomes (any cause hospitalization and diagnoses from 15 large diagnostic categories) as well as specific diagnoses suggested by expert panels most likely to reflect new or chronic manifestations of exposure to the environmental factor. None of these models suggested an increase risk among the exposed.

The SAS System's PROC PHREG with baseline option is a powerful tool for dealing with the early departure of subjects during the study period through censoring. It is also useful for producing data sets, including survival function estimates, which can be used in a simple equation to produce estimates of probability of events. When graphed, these show cumulative probability of event curves as a function of time.

References

Hosmer JR. DW, Lemeshow S. *Applied Survival Analysis; Regression Modeling of Time to Event Data.* New York: John Wiley & Sons; 1999

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 1,* Cary, NC: SAS Institute Inc., 1989. 943 pp.

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 2,* Cary, NC: SAS Institute Inc., 1989. 846 pp.

SAS Institute Inc. SAS/STAT® *Software: Changes and Enhancements through Release 6.11.* Cary, NC: SAS Institute Inc., 1996. 1104 pp.

Allison, Paul D., *Survival Analysis Using the SAS® system: A Practical Guide,* Cary, NC: SAS Institute Inc., 1995. 292 pp.

Acknowledgments

Thank you to CAPT Greg Gray at the Naval Health Research Center, San Diego, for his support of new and different concepts and ideas, and his ongoing encouragement of the pursuit of truth and knowledge.

This research was supported by the Department of Defense, Health Affairs, under work unit no. NMRDC.WR.00098(6423).

SAS software is a registered trademark of SAS Institute Inc. in the USA and other countries.

About The Author

Tyler has used SAS for 7 years including work as a student in MATH and STATISTICS at CSUC, as a graduate student at the University of Kentucky Department of Statistics, and currently as a Statistician with NHRC, San Diego. His responsibilities include mathematical modeling, analysis, management, and documentation of large hospitalization/demographic data sets. This is his 4th WUSS conference having been chosen for best contributed paper for the Data Analysis and Statistics Section last year.

Tyler C. Smith, M.S.
Statistician, Henry Jackson Foundation
Emerging Illness Division
Naval Health Research Center, San Diego
(619) 553-7593 (phone)
SMITH@nhrc.navy.mil