ANALYSIS OF SURVIVAL DATA: A COMPARISON OF THREE MAJOR STATISTICAL PACKAGES (SAS, SPSS, BMDP)

Corey J. Pelz and John P. Klein, Medical College of Wisconsin
Corey J. Pelz, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226

KEY WORDS: Survival analysis, SAS, SPSS, BMDP

Survival analysis techniques have become standard tools for the statistician in medical research. The application of survival models to data is valid when the endpoint of interest is the "time to the occurrence of a particular event." Survival models may be applied to a variety of fields such as biology, medicine, engineering, and economics. With modern computing technology, the analysis of "time-to-event" data has become inexpensive in terms of time. There are several statistical packages on the market today that can be used to do survival analyses. The most commonly used packages are SAS, SPSS, and BMDP. These three packages are compared based upon their capabilities, accuracy, and user-friendliness as applied to survival analysis. Example data sets are used to demonstrate standard and nonstandard conditions that occur when modelling survival data in each of the packages. Several survival analysis applications are presented to determine the agreement among the three packages. Both the univariate and multivariate survival analysis procedures are presented for each package.

1. INTRODUCTION

The application of survival models to data is valid when the endpoint of interest is the "time to the occurrence of a particular event." Survival models may be applied to a variety of fields such as biology, medicine, engineering, and economics. An example of an application in engineering is to model the time it takes for a ball-bearing to wear. The focus will be on applications in biology and medicine where the event of interest may be time to death or time to a particular event such as relapse of a disease. The standard statistical techniques for data analysis are usually not applicable to survival data. First of all survival data are typically not symmetric. A histogram of survival times will indicate that they tend to be positively skewed. As a result it is not reasonable to assume data of this type to be normally distributed. Another feature of survival data that makes it difficult to use standard techniques is that survival times are frequently "censored." The survival time of an individual is said to be censored when the endpoint of interest has not been observed.

Right censoring, which is the most common form, occurs when the exact survival time is not known. All that is known is that the exact survival time exceeds the recorded value. This type of situation can occur if the subjects do not experience the event of interest when the study terminates or they are lost to follow-up. Such data cannot be analyzed by ignoring the censored observations because in general those who tend to live longer are more likely to be censored.

Another feature of survival data is the potential for truncation. For left truncation only subjects that experience a certain intermediate event are made known to the investigator. For example, if the focus of the study is to look at relapse of leukemia prior to death, left truncation occurs because only those who experience the intermediate event (relapse) are observed.

There are both parametric and nonparametric techniques available to model survival data. The parametric methods of estimation assume that the probability density function of the time to a particular event follows a specific distribution, such as the exponential distribution, while the nonparametric methods do not. The three major statistical packages (SAS, SPSS, and BMDP) are compared for both parametric and nonparametric survival analysis methods. Recommendations are given as to when each package is superior under both standard and nonstandard conditions. Several datasets are analyzed by each of the packages so that direct comparisons can be made. These include: (1) ovarian cancer data, Edmunson *et al.* (1979); (2) the Stanford heart transplant data, Crowley and Hu (1977); (3) larynx cancer data, Kardaun (1983); (4) breast feeding data, National Labor Survey of Youth (NLSY); and (5) melanoma data, Lee (1992).

2. COMPARISON OF THE PACKAGES

There are a number of similarities between SAS(version 6.09), SPSS (version 5.0), and BMDP (version 1990) in terms of computational methods for survival analysis. For the most part, the three packages agree with one another with respect to parameter estimation and calculation of available statistical tests. Table 1 lists the procedures that are found in each of the three statistical packages that perform the major survival analysis techniques: Kaplan-Meier method (Kaplan and Meier, 1958), life table methods (Gehan, 1969), Cox proportional hazards models (Cox, 1972), and the accelerated failure time model (Andersen, Borgan, Gill, Keiding, 1993). The life table method is not considered in this discussion since it is no longer commonly used in medical applications. Each of the packages can handle right censored data easily. The major differences among the packages are summarized in Table 2.

2.1 Kaplan-Meier Estimates and Tests

The Kaplan-Meier estimates of the survival function are available in all three packages along with standard errors of the survival function calculated by Greenwood's formula (Greenwood, 1926). The three packages provide the results of the log-rank (Collett, 1994) and the Wilcoxon tests (Gehan, 1969) for comparing the survival of two or more groups. The Tarone-Ware test (Tarone and Ware, 1977) is available in SPSS and BMDP but not in SAS. The Peto-Prentice (Peto and Peto, 1972) test is available only in BMDP. SPSS has the ability to calculate all pairwise comparisons among the groups by issuing a

linear combination of another covariate. Under the first condition, the estimate of the regression parameter is ± infinity. The second condition leads to a singular matrix that is not invertible and therefore the regression parameters cannot be estimated. The results of how each package handles the nonstandard conditions differs among the packages. Both SPSS and BMDP give warning messages that one of the nonstandard conditions is present. SAS provides results without providing any information that a nonstandard condition is present. For example, when condition 1 is present, SAS provides the results of the estimated regression parameters after 15 iterations. These are not the correct estimates because the parameter estimate for the covariate with the condition present is diverging. This can be seen by using the "/itprint" option, but no warning message is given. BMDP provides references that provide information on how to remedy the nonstandard conditions.

2.6 The Accelerated Failure Time Model

Only SAS and BMDP allow the use of the accelerated failure time model. There is an agreement in the results. SAS allows the use of the generalized gamma distribution. It can be useful is choosing which underlying distribution to use to model the data. One must be careful in interpreting the results of both packages, because the estimates that are provided are for the transformed logarithm of survival time. You will want to transform the estimates back to their original units. This can be done using the delta method.

3. CONCLUSION

SAS, SPSS, and BMDP are all very good packages for analyzing survival analysis applications.

Gehan, E.A., (1969). Estimating survival functions for the life table. *Journal of Chronic Diseases*, **21**, 629-44.

Greenwood, M. (1926). The errors of sampling of the survivorship tables. *Reports on Public Health and Statistical Subjects*, number 33, Appendix 1, HMSO, London.

Lee, E.T. (1992). *Statistical Models and Methods for Lifetime Data*, Wilely, New York.

Kalbfleisch, J.D., and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kaplan E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-81.

Kardaun, O. (1983). Statistical survival analysis of male larynx-cancer patients -- A case study. *Statistica Neerlandica*, **37**, 103-25.

Klein, J.P. and Zhang, M.J. (1996). Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia, *Lifetime Data: Models in Reliability and Survival Analysis*, N.P. Jewel, 175-85.

Klein, J.P. and Moeschberger M.L. (1996) Survival Analysis. Springer-Verlag, New York (in press).

Peto, R. And Peto, J. (1972) Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society*, A, **135**, 185-207.

SAS/STAT User's Guide, Version 6, Volume 1 (1990), SAS Institute, Inc., Cary, NC.

SAS/STAT User's Guide, Version 6, Volume 2 (1990), SAS Institute, Inc., Cary, NC.

SPSS for Unix, Advanced Statistics, Release 5 (1993), SPSS, Inc.

SPSS Statistical Algorithms, 2nd Edition (1993), SPSS, Inc.

Tarone, R.E. and Ware, J, (1977). On distribution-free tests for equality of survival distribution. *Biometrika*, **64**, 156-60.

Table 1: Listing of the procedures by survival analysis topic and statistical package.

Survival Analysis Topic	SAS	BMDP	SPSS