TECHNICAL EXPOSITION OF GRAPHICAL TOOLS IN SURVIVAL ANALYSIS

G. G. Enas, C. D. Hardison, and F. W. Rockhold
Eli Lilly and Company

SECTION 1 - INTRODUCTION

Graphical techniques are useful tools in many areas of statistical evaluation and this is especially true in the analysis of time to failure data from censored clinical trials. Hypothetical rating procedures are becoming increasingly more available in the form of computer software. They are also becoming more sophisticated, but there are few descriptive techniques useful in summarizing the data. All of this increases the importance of the role graphics has to play in the statistical analysis of these data.

In this paper we present the outline of a menu-driven system written in SAS which provides the data analyst with tools to characterize a set of failure time data, examine the validity of statistical assumptions, and test hypotheses. One need for this system grew out of a real application in the analysis of a clinical trial. The desire was to provide the statistician with graphics as aid in making inferences and as a method of presenting results to nonstatistical audiences. This presented in a summary of the various options available in the system ranging from basic descriptive graphics to relatively complex diagnostics. We begin by presenting in Section 2 an overview of the statistical methodology utilized by the system. In Section 3 we then how the JAM method is used and give a real-world example in section 4.

SECTION 2 - METHODS

1: Survival Function Estimation

Let \( T_i \) and \( C_i = 1, 2, \ldots, n \) be i.i.d. random variables with distribution functions \( F(t) \) and \( G(t) \) respectively. These functions will describe the survival and censoring distributions respectively. It is only possible to observe the random variables \( (Y_i, S_i)^T = (Y_i, C_i) \), where

\[
Y_i = \min (T_i, C_i)
\]

\[
S_i = \begin{cases} 1 & (T_i \leq C_i) \\ 0 & (T_i > C_i) \end{cases}
\]

\( S_i \) is 1 if \( T_i \leq C_i \) (failure) and 0 if \( T_i > C_i \) (censored)

A random censoring mechanism is assumed to hold. Thus \( T_i \) and \( C_i \) are independent. An important aspect of many controlled clinical trials is the estimation of the survival distribution

\[ F(t) \rightarrow P(T < t) \]

The classic actuarial method of estimation utilizes the standard life table where the time domain is partitioned into mutually exclusive intervals of equal length. The estimation of interval \( \hat{F}(t) \) is the product-limit estimator \( F_{pl}(t) \). The time intervals may be of variable length. \( F_{pl}(t) \) is a consistent estimator and is shown to be the nonparametric maximum likelihood estimate of \( F(t) \).

Let \( Y_i, Y_i', \ldots, Y_i' \) be the distinct survival times. Also let

\[
\begin{align*}
\hat{n}_j & = \text{number alive at time } Y_i' \\
\hat{d}_j & = \text{number of failures at time } Y_i' \\
\hat{\delta}_j & = \begin{cases} 1 & \text{if observations at time } Y_i' \text{ are failures} \\ 0 & \text{if censored} \end{cases}
\end{align*}
\]

The product-limit estimator \( g(t) \) is then given by

\[ F_{pl}(t) = \prod_{j} \left( 1 - \frac{\hat{d}_j}{\hat{n}_j} \right) \]

Assume that any tied, censored intervals are split apart in infinitesimal units and uncensored observations occur just prior to censored observations when censored and uncensored observations are tied.

Confidence intervals may be computed by using the asymptotic convergence of \( \hat{F}(t) \) to a Gaussian process whose variance may be approximated using Greenwood's formula for the variance of the classical actuarial estimator. An \( (1-\alpha)% \) confidence interval is then given by

\[ [F_{pl}(t) - z_{1/2} \sqrt{\text{V}(F_{pl}(t))}, F_{pl}(t) + z_{1/2} \sqrt{\text{V}(F_{pl}(t))}] \]

Confidence intervals may be computed by using the asymptotic convergence of \( \hat{F}(t) \) to a Gaussian process whose variance may be approximated using Greenwood's formula for the variance of the classical actuarial estimator. An \( (1-\alpha)% \) confidence interval is then given by

\[ [F_{pl}(t) - z_{1/2} \sqrt{\text{V}(F_{pl}(t))}, F_{pl}(t) + z_{1/2} \sqrt{\text{V}(F_{pl}(t))}] \]

and \( z_{1/2} \) is the appropriate critical value from the standard normal distribution.

The median survival time may be estimated from the survival table. Generally a confidence interval (low median survival is most informative) is most informative. The method of G. G. Enas, C. D. Hardison and F. W. Rockhold recommends calculating

\[ \hat{s}(Y_i) = \frac{N_i}{N_i} \]

\[ \hat{s}(Y_i) = \frac{N_i}{N_i} \]

\[ \hat{s}(Y_i) = \frac{N_i}{N_i} \]

An approximate \( (1-\alpha)% \) confidence interval for the median survival is then given by the set \( Y_i(j) \) for which

\[ \hat{s}(Y_i(j)) \leq \hat{s}(Y_i(j)) \]

where \( Y_i(j) \) is appropriately chosen from the standard normal distribution. The system presented here computes \( \hat{s}(Y_i) \), and allows one to visually inspect the augmented life table to determine the appropriate confidence interval.

The hazard function \( \hat{h}(t) \) and survival function \( \hat{s}(t) \) are related by

\[ \hat{s}(t) = \exp(-\hat{h}(t)) \]

where

\[ \hat{h}(t) = \int_0^t \hat{h}(u) \, du \]

\( \hat{h}(t) \) is known as the cumulative hazard function.

It may be shown that if the Nelson estimator of \( \hat{h}(t) \) is given by

\[ \hat{h}_n(t) = \sum_{j=1}^{n} \frac{d_j}{n_j} \left( Y_i(j)^{d_j-1} \right) \]

is substituted into (4), then in some but not all situations,

\[ \text{E}(\hat{h}_n(t)) = \frac{\hat{F}_{pl}(t) - F(t)}{\hat{F}_{pl}(t)^2} \]

546


II. Graphical Tools for Evaluating the Effect of Covariates in Cox's Model

Cox's proportional hazards regression model allows the capability of assessing the effect of an explanatory vector \( g \) on the underlying hazard function at any given time. The individual's unique hazard is specified as

\[
H_i(t | g) = A_i(t) \exp(g' \beta) 
\]

where \( g \) is a \( k \)-dimensional vector of regression coefficients and \( A_i(t) \) is an arbitrary function characterizing the underlying hazard when \( g = 0 \). The underlying cause-specific hazard is then given by

\[
H_i(t) = \int_0^t A_i(s) \exp(g' \beta) \, ds 
\]

The partial likelihood function for this model depends only on the ranked survival times. Denote this set of ranked times as \( t = (t_1, t_2, ..., t_n) \)

Let \( \epsilon_j = \exp(g' \beta_i) / \exp(g' \beta_j) \)

\[
\epsilon_j = \frac{\exp(g' \beta_i)}{\exp(g' \beta_j)} 
\]

It may be shown that the \( \epsilon_j \)'s approximate a unit exponential distribution when the model \( (8) \) is correct. However, the \( \epsilon_j \)'s are explicitly dependent on the times to failure and their ranks are not invariants to monotone transformations of the time scale.

A similar set of residuals which are consistent with the rank-invariance property of Cox's model partially likelihood have been defined.

Let \( z_j^* \exp(2g_i^* \beta_i^*) \)

\[
z_j = \exp(2g_i^* \beta_i^*) 
\]

It may be that the \( z_j \)'s approximate a unit exponential distribution when the model \( (12) \) is correct. However, the \( z_j \)'s are explicitly dependent on the times to failure and their ranks are not invariants to monotone transformations of the time scale.

A similar set of residuals which are consistent with the rank-invariance property of Cox's model partially likelihood have been defined.

Let \( z_j = \exp(g_i \beta) \)

\[
z_j = \exp(g_i \beta) 
\]

The partial likelihood function for this model depends only on the ranked survival times. Denote this set of ranked times as \( t = (t_1, t_2, ..., t_n) \)

Let \( z_j = \exp(g_i \beta) / \exp(g_j \beta) \) and \( V_j = \min(t_j, t_{j+1}) - \sum_{i=j}^{n} \exp(-g_i \beta) \)

The \( z_j \)'s do not have unit exponential distribution as the \( z_j \)'s but approximate the conditional expectation of the \( z_j \)'s given the ordered survival times \( t_j \) and are independent of the \( t_j \)’s.

Let \( w_i \) be the rank of \( g_i \) among \( g_1, g_2, ..., g_k \). Let \( X \) be a possible explanatory variable not yet included in the model. If \( X \) is included in the model, then the fitted variables \( W \) will not be able to express the effect of \( X \) on the hazard.

\[
H_i(t | g) = \exp(g' \beta + w \cdot X) 
\]

The hazard ratio for a given level of \( X \) is then given by

\[
H_i(t | g) / H_i(t | g - w \cdot X) = \exp(w \cdot X) 
\]

Thus, the \( g_j \)'s allow one to examine in a sequential manner

1. The proportionality assumption of the variables already included in the model.
2. Checks on the inclusion (exclusion) of important (unimportant) explanatory variables.
3. Indication of the need for transformations of variables already included in the model.

SECTION 3 - USE OF THE SYSTEM

The survival analysis macro facility is invoked upon execution of a 700 CLINT or a CMS EXEC depending upon where the user's data resides. Initially, Screen 1 appears. Option 3 on the main menu (Screen 1) allows the user to get into interactive SAS from the macro facility. Thus one can do all necessary data manipulations before accessing the survival analysis macros. The user may choose to utilize either the survival estimation techniques of Kaplan and Meier or the Cox model building techniques previously discussed. The user is always given the option of not continuing and returning to the main menu from any screen. If one selects the Kaplan-Meier analysis, Screen 2 appears. The user must first enter the name of the SAS data set which contains the data to be analyzed. The data set must contain at least two variables: one for time and another for event, possibly a transformation of \( \epsilon \) to be included for the lack of fit tests present.
In order to invoke the proportional hazards options, the user must be aware of the SAS PROC PHGLI, which fits the Cox model to the data, and provides parameter estimates. This step in the process is currently being incorporated into the macro, although some familiarity with the procedure is helpful in any case. The rank correlation estimates come directly from the macro using PROC CORR.

If the proportional hazards analysis is selected, a few additional questions need to be answered as shown in Screen 3. The names of all the variables included in the Cox model need to be specified as well as an inclusion of whether the variable in question is to be excluded from the model at this point. Both the proportional hazards assumption may be tested or the inclusion or lack of fit of a certain variable. The help screen associated with Screen 3 is shown in Screen 3A.

The system is very easy to use. Minor variations in each job submitted are easily facilitated since all the previous answers to the questions remain until changed. Hence, multiple runs are readily performed. The user may choose to rerun the job either in batch mode or interactively.

SECTION 4: AN EXAMPLE

An illustration of these tools, consider the data from the Veterans Administration [67] lung cancer trial presented in reference [67]. In this trial, 132 males with advanced lung cancer were randomized into one of two chemotherapy treatment groups (standard, 2-trial). Of concern here is the effect of performance status (PERFORM) and tumor type (CELL TP) on survival (TIME). PERFORM is a measure of global fitness on a scale of 1-100. Large cell, small cell, squamous cell, and adenocarcinomas are included. The difference between the two chemotherapies are already shown to be insignificant.

Survival times range from 1 to 999 days with 9 observations being right-censored (CESECR=9). A subset of the SAS dataset "LUNGADJ" containing these three variables is shown in Figure 1. This dataset is a member of the OS dataset "L.RADJ".

The macro system is invoked and Screen 1 appears. We initially want a pictorial display of the estimated survival functions of each tumor cell type (CELL TP) with the associated life tables. Hence, option "T" is selected by Screen 2 appears. The necessary information previously discussed is furnished as shown in Screen 2B. Upon completion of the job run, this analysis produces the survival curves shown in Figure 2 as well as the life tables in Figure 3 from which the plot is derived. The plot suggests differences between the survival functions. Important information concerning these distributions is contained in the 5% confidence intervals for median survival. Figure 1 gives the point estimates and associated confidence intervals computed from information given in the life tables. The variable Z SQUARE is the statistic C of Section 7. Linear interpolation allows computation of the confidence limits.

PROC PHGLI is used to assess the significance of tumor cell type. Three dummy indicator variables are chosen to represent the five levels of the tumor cell type variable as follows:

Original Dummy Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>CELL TP = squamous</th>
<th>CELL TP = small</th>
<th>CELL TP = large</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELL TP</td>
<td>small</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CELL TP</td>
<td>large</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

An explicit assumption of Cox's model is that the hazard functions for each cell type be proportional to each other. For example,

\[ \Lambda (t) = \exp \{ \beta \text{CELL TP} \} \]

where \( \Lambda (t) = \text{hazard function for patients with small cell} \)

\[ \Lambda (t) = \text{hazard function for patients with all other cell types} \]

To test this assumption option 2 is selected in Screen 2 and Screen 2A then appears and the required information is input (See Screen 3B). Both CELLS and CELLS are tested individually using the methods described in Section 2. As seen in Figures 4-5, it is apparent that the cell types appear fairly evenly distributed between both levels of CELLS as well as CELLS.

Once the proportionality assumptions for CELLS and CELLS have been verified, we may examine the list of the variables included thus far in the model. First, CELLS is included in the model alone. The Spearman correlation coefficient \( C = 0.08 \) (p>0.30) and does not show any evidence for association with the cell type. Now with CELLS included in the model, \( C \) is computed between CELLS and the \( r_i \) to check for the inclusion of CELLS.

In this case, CELLS would be included on Screen 3B as a variable already in the model. Here \( C = 0.25 \) (p>0.001) and inclusion of CELLS is indicated in the model. The Spearman correlation coefficients for CELLS and CELLS with the \( r_i \) from the model including both variables were calculated. In both instances the coefficients are very small, giving evidence that CELLS and CELLS do not exhibit an association with the \( r_i \). Hence, no transformation of these data is necessary.

Performance status is known to be a powerful prognosticator of survival. To test this variable as such, its scale was dichotomized as follows:

\[ \text{PERF} = \begin{cases} 0 & \text{PERF} \leq 60 \text{ (non-ambulatory)} \\ 1 & \text{otherwise (ambulatory)} \end{cases} \]

By including PERF alone in the Cox model a significance level of p<0.001 was obtained. Figure 5 shows the results of a check on the proportional hazard assumption. The smaller and middle ranks \( r_i \) are transformed by the non-ambulatory patients' while the larger ranks are dominated by the group of patients continued to be. Thus, this is a case where a deviation from proportionals, hazards may be indicated.

Thus, CELLS and CELLS were retained in the model (p<0.001). It is evident that both adenocarcinoma and small cell cancers in this study have poorer prognosis than squamous and large cell carcinomas. The average increase in the hazard rate for aden cell cancer is over twice the hazard for the remaining types. Likewise the average hazard for small cell cancer is almost three times the hazard for the other three tumor types combined.

This example exhibits only a few of the options available. Other types of plots and statistics may be utilized from this system though not shown here. For example, survival curves adjusted for important explanatory variables may be produced assuming Cox's model. Transformations of the survival curves (e.g. the natural log) are also available. Hazard estimators such as \( \hat{h}(t) \) in equation 6 may also be plotted. The ease of use of the system affords quite convenient to facilitating interactive survival data analysis.
REFERENCES


