Analysis of Censored Data
IMS Lecture Notes – Monograph Series (1995) Volume 27

Graphical Tools for Censored Survival Data

Debasis Sengupta
Indian Statistical Institute, Calcutta
203, Barrackpore Trunk Road, Calcutta 700 035, India

Abstract

This review article discusses a number of graphical procedures for Survival Analysis. These include descriptive plots such as Event history diagrams, scatterplots and plots of estimated curves. Diagnostic plots for checking model assumptions and comparing two samples are also discussed, with special reference to the Cox regression model. A few specific suggestions have also been made for the modification of some of these plots for better quantitative assessment and suitability to the human visual system. It is hoped that this work will help build a comprehensive package for the analysis of survival data, and stimulate further research to fill the gaps in the current state of the art.

1 Introduction

Graphical methods of examining data have gained considerable popularity over the last two decades. Many analysts feel that simple descriptive plots allow them to comprehend the overall pattern, if any, which is not understood easily from a tabular representation. Sometimes these plots help form conjectures or open up unexpected directions of further investigation. In the context of exploratory data analysis, graphical representation is an essential part of model building. Graphical tools are also quite useful in communicating the findings of an applied statistician to the customer. For example, an estimated curve or confidence band is best presented through a plot. Sometimes it is also possible to supplement model-based prediction by representing the associated uncertainty visually.

Apart from the aspect of presentation, graphical methods can also contribute to a better analysis of the data. For instance, a formal statistical test may be accompanied by a plot to examine how the data does or does not conform to the null hypothesis. When the stakes are high, it is often wiser to also examine the pattern in these plots instead of making a decision based solely on a single binary outcome of the formal test, or a p-value. This strategy is particularly relevant when the statistical test is marginally

significant or insignificant. Besides, graphical methods play a role in understanding the relative importance and the degree of misfit of individual cases to an assumed model. Such diagnostic plots, including the graphical 'tests', may help crystalize fresh ideas for improving the model itself.

There has been a considerable surge in the development of graphical tools for Survival Analysis in the recent years. These differ from the corresponding methods in mainstream statistics in at least one of the following aspects: (a) suitability to models and formulations which are typical of Survival Analysis and (b) ability to handle censored data. The objective of this article is to compile systematically some of these methods for the benefit of the practitioners. It is hoped that a thorough assessment of the available methods would also reveal areas where fresh research is needed.

For simplicity, the graphical methods have been broadly classified here into two groups. Graphical representations of raw or smoothed data, plots of estimators, confidence bands and predictive distributions are called *descriptive plots*. On the other hand, graphical tests are classified as *diagnostic plots*, a set which traditionally includes plots of residuals and other casewise diagnostics vs. time or a covariate and so on. Of course some descriptive plots also furnish diagnostic information, and some diagnostic plots may be used for presentation purposes. However, the overlap of the two classes will be ignored here.

While the availability of numerous diagnostic plots is a positive development, many of them have certain weaknesses which can be overcome by suitable modifications. Specifically, a good diagnostic plot should have the following features:

(a) There should be a reference (e.g., a curve) to remind the user of the ideal shape of the plot.
(b) There should be guidlines to help determine whether the deviation is within statistically permissible limits.
(c) There should be provisions to interpret at least certain forms of deviations from the ideal shape (eg, a strictly monotone trend as opposed to constancy).
(d) There should be adjustments to compensate for certain weaknesses of the human eye.

The last feature is desirable as the potential of the human visual system to extract spatial information can not be fully harnessed unless its inabilities are also taken into account. A number of examples of procedures friendly to the eye are given by Tufte (1983) and Cleveland and McGill (1984).

An attempt has been made in this article to improve some existing diagnostic plots in view of the above considerations. A few new plots have also been proposed.

2 Descriptive plots

2.1 Event history diagrams

Goldman (1992) proposed a plot consisting of a number of horizontal lines. The length of each line represents the duration of survival of an individual. Each line is placed at a height which represents the calender time when the measurement begins for the corresponding individual. A special symbol at the right endpoint of a line indicates an observed death. Withdrawal from the study, death due to unrelated causes and other forms of censoring are indicated by other symbols. If a number of individuals are alive at the time of conclusion of the study, the right endpoints of the corresponding lines lie in a straight line, named the *now-line* of that date by Goldman, stretching from the top left corner to the bottom right. An attractive feature of this 'eventchart' is that one can visualize the events of interest both in calender time and in time measured from the date of diagnosis (or any other chosen starting point). The status of all the individuals present in the study at a given time may be found by examining the events along the corresponding now-line. This helps assess the effects of special events such as an epidemic, change in policies etc. These and other advantages pointed out by Goldman make the plot a useful part of the practitioner's toolkit. Its only weakness is that it becomes rather congested with lines and symbols for large data sets.

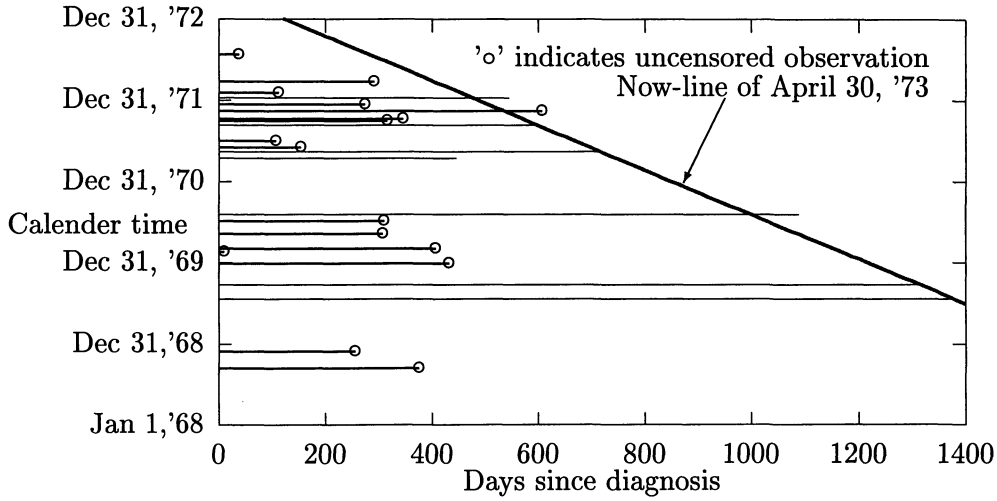


Figure 1: Eventchart for the Oropharynx Carcinoma Data

Figure 1 shows the Eventchart for a section of the Oropharynx Carcinoma Data (Source: Kalbfleisch and Prentice, 1980, pp.225–229). Only the female population in the test (non-standard) treatment group has been considered. The now-line of April 30, 1973 shows that four subjects were censored on this date, while three others remained in the study beyond this date.

2.2 Scatterplots

The task of plotting failure times against the respective covariate values becomes complicated in the presence of censoring. A simple solution is to use different symbols for censored and observed failure times. (Here a censored observation also refers to possible left-censoring). However, the different symbols seem to clutter the field of vision and thereby disrupt the process of assessment. Gentleman and Crowley (1991) suggested plotting censored and observed data with a common symbol but different colors or grey scales. This strategy partially solves the problem, and sometimes brings out information about the censoring pattern.

If the sample size is small, the time-dependence of a covariate may be incorporated into the scatterplot by plotting the covariate value of each individual against time for the duration of the observation, and putting an appropriate symbol representing the epoch of failure or censoring.

Unfortunately these plots do not give any indication of how the plot would have appeared *had all the unobserved failure times been observed.* Zhou (1992) suggested plotting only the uncensored observations with symbol size proportional to the corresponding jump size of the product-limit (PL) estimator. A drawback of this plot is that the covariate information of the censored observations are ignored.

Another possiblity is to replace the censoring time in the scatterplot by an estimate of the unobserved life-length. This may be done through an assumed model which incorporates the data and censoring information, and possibly the covariates. If possible, only those observaions with comparable covariate profile should be used for the extrapolation of a censored lifetime.

Smoothing of the scatterplots may also be considered. This is particularly needed when the raw scatterplot is too crowded owing to a large sample size. Thus one has to plot a single smoothed failure time for each covariate value. For a given value of the covariate, the smoother may be based on the set of observations (censored or observed) with comparable covariate values. Many common smoothers in the complete samples case, e.g., the sample mean and the sample median, may be written as a functional of the empirical survival function. Gentleman and Crowly (1991) suggested that these smoothers be adapted to the right-censored data case by using the PL estimator instead of the empirical survival function. A further extension to the case of right- *and* left-censoring can be made by using the non-parametric maximum likelihood estimator (NPMLE, see Turnbull, 1974).

2.3 Plots of estimated curves

Plots of the estimates of certain functions are commonly used. These plots are sometimes aided by plotting the loci of pointwise confidence limits or

simultaneous confidence bands. The most popular of these are the PL estimator of the survival function and the Nelson-Aalen (NA) estimator of the cumulative hazard function (CHF). Pointwise confidence limits for the survival function based on the PL estimator may be obtained from Greenwood's formula (see Miller, 1981). A variety of confidence bands may be found in Csorgo and Horvath (1986), Hollander and Peña (1989) and the references therein. Pointwise confidence limits and confidence bands for the CHF based on the NA estimator are available in Andersen and Borgan (1985) and Bie et al. (1987). These confidence limits and bands, as well as the others mentioned later in this article are all asymptotic in nature.

One may also consider plotting an estimator of the hazard rate function. The estimators due to Ramlau-Hansen (1983), Senthilselvan (1987) and Antoniadis (1989), among others, are available for this purpose. While pointwise confidence limits can be constructed through the asymptotic distributions of these estimators given by the respective authors, confidence bands for the hazard rate function appears to be unavailable in the literature.

The mean residual life (MRL) function is sometimes used in actuarial studies and medical statistical literature. Nonparametric estimators of the MRL function are given by Ghorai and Rejtoe (1987) and Park et al. (1993). Guess and Park (1991) give confidence bands.

Estimators of the cumulative hazard, the hazard rate and the MRL functions under the assumption of monotonicity are available in the literature (see, for instance, Robertson et al., 1988). These functions may also be estimated parametrically. However, these plots will not be discussed here.

3 Goodness of fit plots

Often one is interested in checking whether a specified distribution or a *family* of distributions fits the data well. The diagnostic plots for this purpose are based on the principle of comparing a nonparametric estimate of the survival function or the CHF with the corresponding parametric estimate. Let the PL estimator of the survival function and the NA estimator of the CHF be $1-\widehat{F}$ and $\widehat{\Lambda}$, respectively. The corresponding parametric maximum likelihood estimators (MLE) are denoted by $1-\widehat{F}_0$ and $\widehat{\Lambda}_0$, respectively. The common graphical 'tests' of the goodness of fit are as follows.

(a) Overlaid plots of $\widehat{F}$ and $\widehat{F}_0$.

(b) Overlaid plots of $\widehat{\Lambda}$ and $\widehat{\Lambda}_0$.

(c) The plot of $\widehat{F}_0^{-1}(\widehat{F}(t))$ vs. t (Q-Q plot).

(d) The plot of $\widehat{F}(t)$ vs. $\widehat{F}_0(t)$ (P-P plot).

(e) The plot of $\frac{2}{\pi}\sin^{-1}\widehat{F}^{1/2}(t)$ vs. $\frac{2}{\pi}\sin^{-1}\widehat{F}_0^{1/2}(t)$ (stabilized P-P plot).

(f) The plot of $\widehat{\Lambda}_0^{-1}(\widehat{\Lambda}(t))$ vs. t (cumulative hazard plot).

(g) The plot of $\widehat{\Lambda}(t)$ vs. $\widehat{\Lambda}_0(t)$.

In the above it is assumed that $\widehat{F}_0$ and $\widehat{\Lambda}_0$ are continuous and strictly increasing so that the inverse functions are uniquely defined. In plots (c)–(g) the reference is a straight line of unit slope through the origin. Some practitioners prefer plot (c), since it tends to show a linear trend even when the assumed model holds with misspecified location and scale parameters. However, it can sometimes have a misleading visual impact, as discussed below. The stabilized P-P plot (Michael, 1983) is thought to have a faster rate of convergence to the plot corresponding to the 'true' distributions.

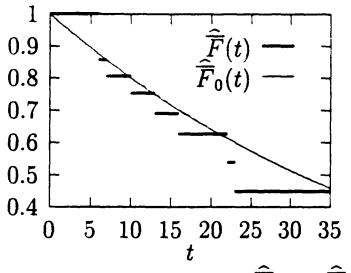
In the special case when the goodness of fit of a *completely specified* distribution is considered, $\widehat{F}_0$ and $\widehat{\Lambda}_0$ are deterministic. In such a case, the confidence bands mentioned in the previous section can be plotted in (a) and (b) for reference. Several confidence bands for the other plots may be found in Csorgo and Horvath (1986), Michael and Schucany (1986) and Hollander and Peña (1989). Most of these bands are suitable for randomly right-censored data. Guilbaud (1988) give an a small-sample Kolmogorov-Smirnov test for left-truncated and right-censored data, which can be converted naturally into a graphical test.

However, these bands are not very useful when the 'null hypothesis' leaves one or more parameters unspecified. For example, a confidence band for $\widehat{F} - \widehat{F}_0$ would be much more appropriate for plot (a). There appears to be a void in the literature in the area of confidence bands for the above plots in the general case of unspecified parameters.

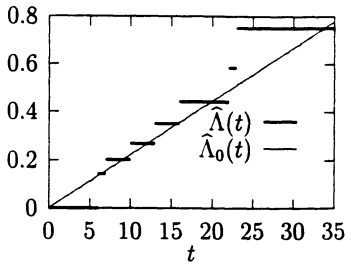
Another weakness of these plots (except, to some extent, plots (d) and (e)) is that the transition points of the plots may not be evenly spread. When the data set is large, most of the points would be concentrated in a narrow zone, while a handful of points would be spread over a wide region. The human eye is likely to give undue importance to the part of the plot with fewer points, leading to possibly biased conclusions. This problem is partially solved by using the P-P plot, since it has a uniform horizontal spacing between the points in the uncensored case. However, this advantage is lost in the censored data case, since the jumps of the PL estimator are not of uniform size. Waller and Turnbull (1992) proposed to rectify this problem by plotting $\widehat{F}_u(\widehat{F}_0^{-1}(\widehat{F}(t)))$ vs. $\widehat{F}_u(t)$, where $\widehat{F}_u$ is the empirical distribution of the uncensored observations. This empirically rescaled plot has the property that the points are evenly spread along the horizontal axis.

The plots (a)–(g) and the empirically rescaled P-P plot are illustrated in Figure 2 with the Leukemia Data of Freirich et al. (1964). The data represent the times of remission (in weeks) of 21 Leukemia patients under the drug 6-MP.

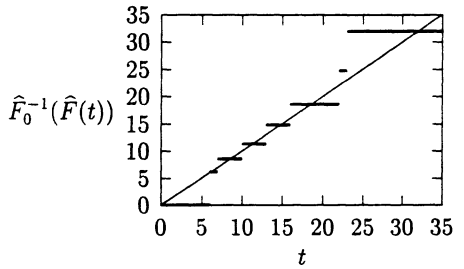
When the sample size is large, specific types of departure from the straight line with unit slope in plots (c)–(g) can suggest alternative models. For example, an S-shape of any of these plots indicate that the actual



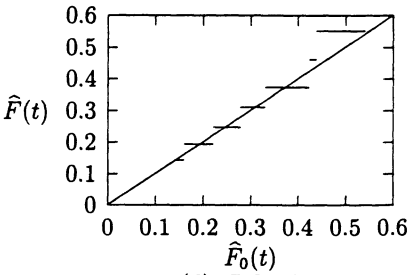
(a): Overlaid plots of $\hat{\bar{F}}$ and $\hat{\bar{F}}_0$



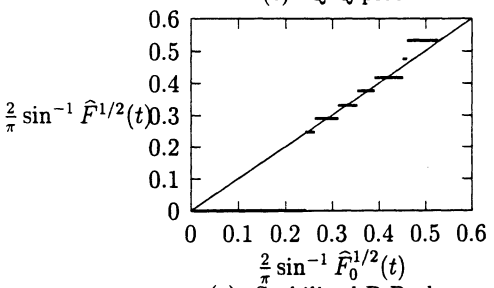
(b): Overlaid plots of $\hat{\Lambda}$ and $\hat{\Lambda}_0$



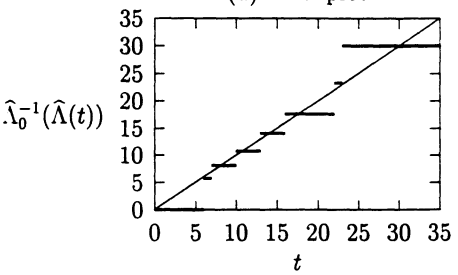
(c): Q-Q plot



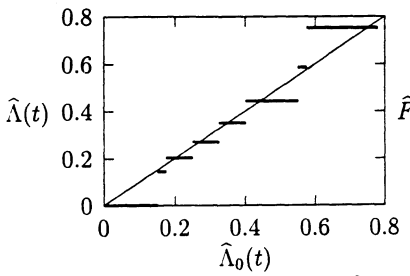
(d): P-P plot



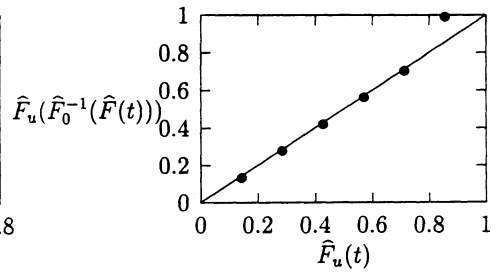
(e): Stabilized P-P plot



(f): Cumulative hazard plot



(g): Plot of $\hat{\Lambda}$ against $\hat{\Lambda}_0$



(h): Empirically re-scaled P-P plot

Figure 2: Goodness of fit plots for the Leukemia Data

distribution may have lighter tails than the assumed one. If the assumed distribution is exponential, as is often the case, a convex, star-shaped or superadditive trend in the plots (c), (f) or (g) suggest that the actual distribution may belong to the increasing failure rate (IFR), increasing failure rate average (IFRA) or new better than used (NBU) classes, respectively. These aging classes of life distributions (see Barlow and Proschan, 1975 for definitions, properties and interrelationships) are useful in Reliability. A plot is called star-shaped provided it intersects any straight line through the origin at most once and from below. A plot is superadditive if it does not intersect the graph of any uniform staircase function, originating from (0,0) with a horizontal line, after touching it at the first vertical jump point.

Another plot that brings out specific types of departure from exponentiality is the empirical total *time on test* (TTT) plot. For a distribution F,

the TTT plot is an empirical version of the scaled TTT transform defined as

$$\phi(u) = \frac{\int_0^{F^{-1}(u)} \overline{F}(t)dt}{\int_0^{\infty} \overline{F}(t)dt}, \quad 0 \leq u \leq 1,$$

where $\overline{F} = 1 - F$. The TTT transform can be plotted within the unit square, and is a non-decreasing funtion passing through the origin and the point (1,1). The transform for any exponential distribution is the diagonal straight line. It can be shown that a distribution belongs to (a) the IFR class if and only if $\phi(\cdot)$ is concave, (b) the decreasing mean residual life (DMRL) class if and only if $(1 - \phi(u))/(1 - u)$ is a decreasing function of u and (c) the new better than used in expectation (NBUE) class if and only if $\phi(u) \geq u$ for all $u \in (0, 1)$. The definition and some properties of the DMRL and NBUE classes may be found in Barlow and Proschan (1975). These classes are also found to be useful in Reliability. Each of the three properties mentioned above are easy to check visually.[1] If the empirical TTT plot is seen to have any of the above properties, a reasonable guess about a more appropriate model can be made. Klefsjö (1982) gives a comprehensive account of these characterizations of the TTT plot. It can be shown that a censored data analogue of the empirical TTT plot is obtained by plotting $\int_0^{t_j} \widehat{\overline{F}}(t)dt / \int_0^{t_n} \widehat{\overline{F}}(t)dt$ against $\widehat{F}(t_j)$, where $t_1 < t_2 < \cdots < t_n$ are the ordered failure times. See Csorgo et al. (1987) for further details.

Miller (1981), D'Agostino (1986) and Kunitz and Pamme (1991) describe other goodness of fit plots for a few special parametric families of life distributions. Parametric methods and models will not be discussed in the remaining part of this article.

[1]Note that $(1 - \phi(u))/(1 - u)$ is a decreasing function of u if and only if the TTT plot, after a 180° rotation, is star-shaped.

4 Comparing samples graphically

4.1 Checking the equality of two samples

The simplest plot to check the equality of two distributions is the *Box-plot.* This consists of two boxes plotted in a coordinate system where the vertical axis represents lifetime, but the horizontal axis has no particular interpretation. Consequently the width of the boxes can be chosen arbitrarily. It is suggested that the width be chosen proportional to the corresponding sample size. The upper and lower edges of each box represents the upper and lower sample quartile, respectively. Another horizontal line corresponding to the sample median is also drawn. These plots can also be used to compare more than two samples. Incorporation of the sample size information through the box width provides an indirect way of indicating the degree of confidence in the estimated quantiles. Censored data can also be handled easily by using quantiles of the NPMLE or other quantile estimators for censored data. However, one of the quartiles may not be observed when there is heavy censoring. In such a case one can either use a model-based quartile estimate or an open-ended box extended up to the largest observed failure time (see Gentleman and Crowley, 1991).

The Box-plot is illustrated with Frierich's Leukemia Data set in Figure 3. The control group and the group under drug 6-MP each have 21 patients. It is quite evident from the plot that the lower quartile of the latter group is higher than the upper quartile of the control group.

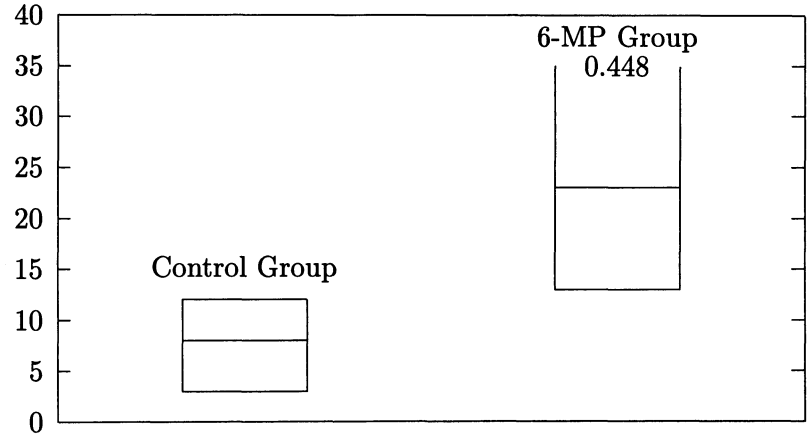


Figure 3: Box-plots for the two groups of leukemia patients

The estimated survival functions or the estimated CHFs for the two populations are often plotted simultaneously. A quantitative guideline for comparison may be provided by plotting poitwise confidence limits for the

curve difference, centered around the average of the two curves. If for $i = 1, 2$ $(\widehat{\overline{F}}_i(\cdot) - l_i(\cdot), \widehat{\overline{F}}_i(\cdot) + r_i(\cdot))$ is a $1 - \alpha/2$ level confidence band of $\overline{F}_i(\cdot)$ (see Section 2.3), then one can use the conservative confidence band

$$(\widehat{\overline{F}}_1(\cdot) - \widehat{\overline{F}}_2(\cdot) - l_1(\cdot) - r_2(\cdot), \widehat{\overline{F}}_1(\cdot) - \widehat{\overline{F}}_2(\cdot) + r_1(\cdot) + l_2(\cdot))$$

for $\overline{F}_1 - \overline{F}_2$. This corresponds to a confidence level of at least $1 - \alpha$. Accordingly one has to plot the functions $r_1(\cdot) + l_2(\cdot))$ and $l_1(\cdot) + r_2(\cdot)$ as conservative upper and lower confidence bands, respectively, in the curve difference plot.

The above plots are quite popular, as they bring out comparative information *along with* descriptive information about the individual samples. However, there are pitfalls. Cleveland and McGill (1984) point out that the human eye often confuses horizontal separation with vertical separations. It is entirely plausible that the vertical separation of two curves is much smaller compared to the horizontal separation, or vice versa. In either of these cases the curves would be visually interpreted as 'close,' sometimes leading to wrong conclusions. Using confidence bands do not help much in these situations, as in the marginal situations, these additional curves only overcrowd the picture in a narrow range along the horizontal or the vertical direction. As a remedial measure, one may consider plotting the curve-difference along with the confidence bands. However, the nature of deviations of this plot from the horizontal axis often can not be interpreted satisfactorily. There are better ways of comparing the two estimated curves graphically, if the diagnostic objective can be delinked from descriptive task of displaying the estimated curves simultaneously.

One such plot is that of $\widehat{\overline{F}}_1$ vs. $\widehat{\overline{F}}_2$, the respective estimators of the survival functions in groups 1 and 2. The PL estimator is used in the case of random right-censoring. This may be called a two-sample P-P plot. In the medical statistical literature, where one group is disease-free and the other is diseased, this plot is known as the *receiver operating characteristic* (ROC). The two-sample Q-Q plot consists of plotting the graph of $\widehat{\overline{F}}_2^{-1} \circ \widehat{\overline{F}}_1$. If the set $\widehat{\overline{F}}_2^{-1}(t)$ is not a singleton at any t, the whole set is plotted in the form of an interval along the vertical direction. The straight line with unit slope through origin serves as a reference for both the plots. The points of the P-P plot are more uniformly spaced compared to those of the Q-Q plot, provided the censoring is not too heavy. On the other hand, the Q-Q plot is more interpretable in terms of the departures from the ideal shape. A convex shape of this plot indicates that F_1 is convex ordered with respect to F_2, while a star or superadditive shape corresponds to the star or superadditive order, respectively. Bootstrap confidence bands for the P-P and Q-Q plots in the censored data case are given by Lu et al. (1994) and Campbell (1994).

Another interesting graphical tool for comparing two samples is the plot

of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$ (Lee and Pirie, 1981), where the Nelson-Aalen estimators are used. The results of Schumacher (1984) can be used to construct confidence bands. This plot clearly brings out 'proportional hazard'-type deviations from equality through a straight line pattern with slope other than unity. It turns out that the plot has much more information to offer through easily recognizable shapes. This is discussed in more detail in the next section.

4.2 Checking the proportionality of hazards

Proportionality of hazards may be questioned in situations where there may be long-term benefits or adverse effects of a particular treatment. In order to detect whether the hazards in two samples are proportional, one may simply plot $\log \widehat{\Lambda}_2$ alongside $\log \widehat{\Lambda}_1$. The separation of these two curves should be approximately uniform if the proportional hazards (PH) model holds. It may be recalled that visually assessing the separation of two empirical curves is not very easy. The task may be even more strenuous as one has to look, not for the existence of a separation but instead, for its *constancy.*

Dabrowska et al. (1992) suggested that the difference $\log \widehat{\Lambda}_1 - \log \widehat{\Lambda}_2$ should be plotted and its closeness to a straight line should be examined. They provided confidence bands for this plot. Dabrowska et al. (1989) furnished confidence bands for (a) the plot of $(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)/\widehat{\Lambda}_2$ and (b) the plot of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$ (Lee and Pirie, 1981) mentioned in the previous section. The first plot is very similar to the log-hazard difference plot, since the theoretical counterparts $\log \Lambda_1 - \log \Lambda_2$ and $(\Lambda_1 - \Lambda_2)/\Lambda_2$ are stricty monotone functions of one another. If the PH model holds, both of these plots should be close to a horizontal straight line, while the plot of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$ should resemble a straight line of arbitrary slope through origin.

It may be noted that a monotone trend of the plot of $\log \widehat{\Lambda}_1 - \log \widehat{\Lambda}_2$ or $(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)/\widehat{\Lambda}_2$ indicates that the ratio of the 'true' cumulative hazards Λ_1/Λ_2 is monotone. This is an important form of departure from the PH model, and can describe the reversal of relative benefits of two treatments in the long run which is not very uncommon in the medical literature. As a special case this includes the 'monotone hazard ratio' relationship. The plot of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$, which coincides with the graph of $\widehat{\Lambda}_1 \circ \widehat{\Lambda}_2^{-1}$, is even more informative. Sengupta and Deshpande (1994) showed that the function $\Lambda_1 \circ \Lambda_2^{-1}$ is

(a) convex if and only if the ratio of hazard rates λ_1/λ_2 is increasing (provided the ratio exists),

(b) star-shaped if and only if Λ_1/Λ_2 is increasing,

(c) superadditive if and only if $q_{1,t}(x) \geq q_{2,t}(x)$ for all $x, t \geq 0$, where $q_{i,t}(x)$ is a quantile of the distribution F_i rescaled suitably to account for aging. Specifically, $q_{i,t}(x)$ satisfies $P[X_i > q_{i,t}(x)] = P[X_i > x + t | X_i > t]$, where X_i has CHF Λ_i, $i = 1, 2$.

These characterizations make the deviation of the plot of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$ from a straight line more interpretable.

The main shortcoming of these plots is that these are prone to large fluctuations near the right end-point, especially for moderate and small sample sizes. This is because of the impact of the smaller risk sets on the NA estimators. This may be rectified following the suggestion of Gill and Schumacher (1987), who proposed to plot $\widehat{\Lambda}_1^K$ vs. $\widehat{\Lambda}_2^K$, where

$$\widehat{\Lambda}_i^K(t) = \int_0^t K(u) d\widehat{\Lambda}_i(u), \qquad i = 1, 2,$$

and $K(\cdot)$ is a weight function that is predictable with respect to the filtration corresponding to the counting processes of observed failures in the two populations. If the weight function is monotone decreasing, the plot is Expected to be more stable near the end. Gill and Schumacher suggested a number of weight functions satisfying the requisite conditions.

It is evident that only the first of the above three interpretations continues to apply to the modified plot. Although Gill and Schumacher proposed their modification only for the plot of $\widehat{\Lambda}_1$ vs. $\widehat{\Lambda}_2$, there is no reason why the same cannot be applied to the plots of $\log \widehat{\Lambda}_1 - \log \widehat{\Lambda}_2$ or $(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)/\widehat{\Lambda}_2$. A monotone hazard ratio of the samples should reflect through a monotone shape of each of these curves. The confidence bands mentioned above should also be applicable with minor modifications.

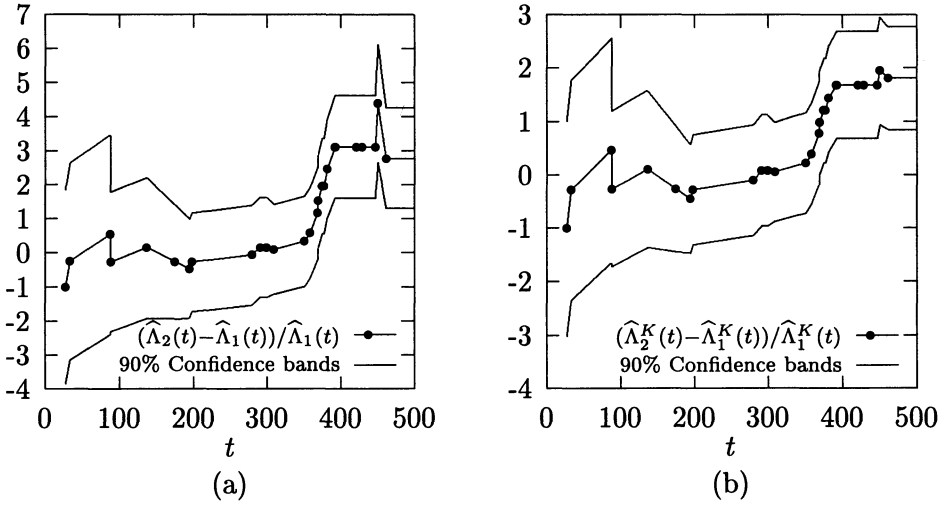


Figure 4: Plot of $(\widehat{\Lambda}_2(t) - \widehat{\Lambda}_1(t))/\widehat{\Lambda}_1(t)$ vs. t and its modification

Figures 4(a) shows the plot of $(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)/\widehat{\Lambda}_2$ against time, along with the 90% simultaneous confidence bands, for the Ovarian Cancer data of Fleming et al. (1980). There are 15 patients with low grade tumor (stage II) in group 1 and 20 patients with high-grade tumor (stage IIA) in group 2.

The computations are done as in Dabrowska et al. (1989) with the choice $q(u) = \{u(1-u)\}^{-1/2}$, suggested by them. The modified plot of Figure 4(b), which makes use of the weight function $K(t) = Y_1(t)Y_2(t)/(Y_1(t) + Y_2(t)$, is smoother for larger values of t. Thus it is easier to observe the increasing trend in the curve.

Since the 'null hypothesis' here is that of proportional hazrds with an unspecified constant of proportionality, there is no fixed reference curve for the plots discussed above. However, a natural reference for the plot of of $\widehat{\Lambda}_1^K$ vs. $\widehat{\Lambda}_2^K$ is the straight line connecting the end-point of this curve to the origin. The slope of this straight line corresponds to an estimator of the constant of proportionality belonging to the class of estimators proposed by Andersen (1983). This class is a rich one which includes as special cases certain estimators as efficient as the Cox estimator. However, it is not proper to examine whether this line is included in the confidence band proposed by Dabrowska et al. (1989), as the line itself is stochastic. One may judge how bad the deviation is, in a more meaningful way, by examining the vertical separation $\widehat{\Lambda}_1^K - \widehat{\theta}_K \widehat{\Lambda}_2^K$, where θ_K is the slope mentioned above. Let τ be a prespecified time lying within the support of the underlying life distribution, but large enough to be larger than most of the observed failure times with a high probability. Suppose further that the weight function K converges in probability to a deterministic function in $D[0, \infty)$. The standard theory of counting processes ensures that under certain conditions and after proper normalization (see Gill and Schumacher, 1987) the limiting joint distribution of $\widehat{\Lambda}_1^K(t)$, $\widehat{\Lambda}_1^K(\tau)$, $\widehat{\Lambda}_2^K(t)$ and $\widehat{\Lambda}_2^K(\tau)$ is multivariate normal for every $t \in (0, \tau)$. Note that $\widehat{\theta}_K = \widehat{\Lambda}_1^K(\tau)/\widehat{\Lambda}_2^K(\tau)$. It follows by an application of the delta-method that, under the null hypothesis of proportional hazards, the normalized difference $(\widehat{\Lambda}_1^K(t) - \widehat{\theta}_K \widehat{\Lambda}_2^K(t))/\widehat{\sigma}_y(t)$ converges in distribution to a standard normal variate for every $t \in (0, \tau)$, where

$$\widehat{\sigma}_y^2(t) = \left(1 - \frac{2\widehat{\Lambda}_2^K(t)}{\widehat{\Lambda}_2^K(\tau)}\right)(V_1(t) + \widehat{\theta}_K V_2(t)) + \left(\frac{\widehat{\Lambda}_2^K(t)}{\widehat{\Lambda}_2^K(\tau)}\right)^2 (V_1(\tau) + \widehat{\theta}_K V_2(\tau)),$$

$$V_i(t) = \int_0^t K(s) Y_i^{-1}(s) d\widehat{\Lambda}_i^K(s), \qquad i = 1, 2,$$

and $Y_i(t)$ is the number at risk at time t for the ith sample. Thus one can plot pointwise confidence limits for the vertical separation of the plot of $\widehat{\Lambda}_1^K$ vs. $\widehat{\Lambda}_2^K$ from the reference line of slope $\widehat{\theta}_K$. The locus of these limits may be plotted, centered around the actual plot. If the straight line falls outside this range, lack of proportionality is indicated. Although a confidence band would be more appropriate in this context, it is not easy to obtain.

It may be argued that the horizontal separation is as important as the vertical one when a comparison of two samples is intended. This and a few related issues will be discussed in Section 6.

5 Diagnostic plots for the Cox regression model

5.1 Checking the proportional hazards assumption

If all the covariates are discrete, one can form a group for every covariate profile and check the proportionality of hazards of pairs of group in the manner discussed in the previous section, provided the sample size per group is not too small. If there is only one continuous covariate in addition to the discrete covariates, it can be stratified. In the case of multiple continuous covariates, the whole space of these covariates can be partitioned for pairwise comparison of groups. However, this is not a practical proposition when the number of such groups is large, as the number of individuals in some groups may be very small.

However, a tractable solution can be reached if the effects of all but a few covariates on the hazard are known to be proportional. For example, there may be several treatment groups within each of which a PH model is valid. Alternatively, there may be one or two continuous covariates such that they do not have have a proportional effect on the hazard, but given a profile of these, the effects of the others are proportional. The space of these possibly problematic covariates can be partitioned into groups. Thus in either case the effective model is of the form

$$\lambda_j(t; z(t)) = \lambda_{j0}(t) \exp(\boldsymbol{\beta}_j{}' \boldsymbol{z}(t)), \qquad j = 1, \ldots, p,$$

where p is the number of strata and $\boldsymbol{z}(t)$ represents that part of the covariate vector which is known to have a proportional effect on the hazard. One can find an estimator of $\boldsymbol{\beta}_j$ (say, $\widehat{\boldsymbol{\beta}}_j$) from the observations of group j. Breslow's (1974) estimator of the integrated baseline hazard $\Lambda_{j0}(t) = \int_0^t \lambda_{j0}(s)ds$ for group j is then

$$\widehat{\Lambda}_{j0}(t) = \int_0^t \left(\sum_{i=1}^{n_j} Y_{ji}(s) \exp[\widehat{\boldsymbol{\beta}}_j' \boldsymbol{z}_{ji}(s)] \right)^{-1} dN_j(s), \qquad j = 1, \ldots, p, \quad (1)$$

where $Y_{ji}(t)$ is an indicator whether individual i of group j is at risk at time t, $\boldsymbol{z}_{ji}(t)$ is the covariate profile of this individual at time t, n_j is the number of individuals in group j and $N_j(t)$ is the counting process for the failures in group j. In the absence of covariates this estimator coincides with the NA estimator. In the spirit of Gill and Schumacher one may also consider the weighted estimators of the form

$$\widehat{\Lambda}_{j0}^K(t) = \int_0^t \left(\sum_{i=1}^{n_j} Y_{ji}(s) \exp[\widehat{\boldsymbol{\beta}}_j' \boldsymbol{z}_{ji}(s)] \right)^{-1} K(s) dN_j(s), \quad j = 1, \ldots, p. \quad (2)$$

Once these estimators are found, they may be checked for proportionality graphically in the manner indicated in the previous section. The method

proposed by Andersen (1982) is a special case of this. Confidence bands for some of these plots may be constructed using the results of Dabrowska et al. (1992) with minor modifications for the weight function.

Arjas (1988) suggested a plot of $H_j(k)$ vs. k, where

$$H_j(k) = \int_0^{T_{jk}} \sum_{i=1}^{n_j} \left(\sum_{i=1}^{n_j} Y_{ji}(s) \exp[\widehat{\boldsymbol{\beta}}_j' \boldsymbol{z}_{ji}(s)] \right)^{-1} Y_{ji}(s) \exp[\widehat{\boldsymbol{\beta}}_j' \boldsymbol{z}_{ji}(s)] dN_{ji}(s),$$

for $j = 1, \ldots, p$. In the above, N_{ji} is the counting process for the ith individual in group j and the indices of the individuals are according to the order of their failures within the group. Arjas argues that the plots for the different groups should approximately overlap along a straight line if the PH assumption holds and diverge in the middle part if the pairwise hazard ratio is monotone. Some guidelines for calibrating the plots are also given by Arjas.

Therneau et al. (1990) suggested plotting $U(\widehat{\boldsymbol{\beta}}, t)$ vs. t, where $U(\boldsymbol{\beta}, t)$ is the score process for the covariate in question or the partial derivative (with respect to this covariate) of the logarithm of the Cox likelihood evaluated at time t. They provided asymptotic confidence bands for this process under the assumptions that the PH model holds and the covariate in question is independent of the other covariates. The latter assumption is rather strong. Instead, one may use simulated samples from the asymptotic null distribution of the process in the manner indicated by Lin et al. (1993). However, systematic deviations of this plot from the zero-line do not have obvious interpretations.

5.2 Checking the effect of a covariate

The problem of checking the significance of a covariate in the presence of other covariates is similar to the two-sample problem. One can obtain an estimator of the form (1) or (2) for the cumulative baseline hazard, *including* and *excluding* the variable in question and compare them using the methods of Section 4.1. Since the two estimators are obtained from the same sample, the confidence limits mentioned in Section 4.1 are not applicable in this case. If the covariate is discrete or can be discretized with a reasonable number of observations per cell, then one can judge the significance of the covariate through pairwise graphical comparison between groups, with the appropriate confidence limits. One may also use Arjas' plots for the strata (see Section 5.1), which would resemble divergent straight lines when the covariate has a significant and proportional effect on the hazard. Finally, a systematic pattern in the the plot of any one of the residuals described in Section 5.3 against a covariate excluded from the current model would suggest that the covariate contains relevant information.

Once a covariate is found to be significant, one can also check its functional form which may sometimes be nonlinear. One possibility is to work with the strata corresponding to a discretized covariate. Consider the ratio $\hat{f}(j) = \hat{\Lambda}_{j0}^K(\tau)/\sum_{k=1}^p \hat{\Lambda}_{k0}^K(\tau)$ for $k = 1, \ldots, p$, where the notation is as in (2). These should be consistent estimators of $f(j)/\sum_{k=1}^p f(k)$, $k = 2, \ldots, p$, where $f(j)$ is the functional value for a covariate level representative of stratum j (say, $z(j)$). Consequently a plot of $\hat{f}(j)$ vs. $z(j)$ would reflect the approximate shape of the functional form. Note that it is not necessary to estimate the scale factor, which would be absorbed in the estimated coefficient anyway. Confidence limits of the plot can also be obtained using the standard counting process theory. Another approach for checking the functional form is based on *martingale residuals*

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp[\hat{\boldsymbol{\beta}}' \boldsymbol{z}_i(s)] d\hat{\Lambda}_0(s),$$

where the subscript i refers to the individual and $\hat{\Lambda}_0(t)$ is Breslow's estimator of the cumulative baseline hazard of the entire sample. Therneau et al. (1990) suggested omitting the covariate in question and plotting a smoothed version of the resulting martingale residual $\widehat{M}_i(\tau)$ against the omitted covariate. The rationale seems to be the following approximation of the right hand side:

$$\begin{aligned} \text{RHS} &\approx N_i(t) - \int_0^t Y_i(s) \exp[\boldsymbol{\beta}' \boldsymbol{z}_i(s)] d\Lambda_0(s) \\ &= M_i(t) + \int_0^t Y_i(s) (\exp[f(x_i)] - 1) \exp[\boldsymbol{\beta}' \boldsymbol{z}_i(s)] d\Lambda_0(s), \end{aligned}$$

where x_i is the value of the dropped covariate for the ith observation and f is its functional form. Therneau et al. (1990) further justified this through some asymptotic arguments, assuming that the dropped covariate is independent of the others. Lin et al. (1993) suggested another way of smoothing this plot and provided the asymptotic distribution of the modified plot for calibration.

Sometimes the effect of a covariate on the life distribution of an individual is also of interest. Suppose the profile of most of the cavariates ($\boldsymbol{z}$) for an individual is known, and the task is to determine the effect of a controllable covariate x. A simple comparison would be to discretize x (unless it is naturally discrete) and plot the PL-estimator or the NA estimator for each of these values. If x is a binary variable, one may plot the difference of the two log-transformed NA estimators along with calibration (see Dabrowska et al., 1992). However, a more informative plot would be the plot of an estimated time-to-failure against the covariate value, as suggested by Heller and Simonoff (1992). For this purpose one may use the Cox estimator defined implicitly by $(\widehat{\widehat{F}}_0)(t)^{\exp(\hat{\beta}' z)} = 0.5$.

5.3 Checking the overall fit and identifying discordant cases

The graphical methods described above are designed to check one violation of the model assumptions at a time. In reality, more than one assumptions can go wrong at the same time. There is no guarantee that the above methods would continue to provide the intended information in such a situation. Therefore it is necessary to check the overall fit through general diagnostic plots which are the graphical analogues of omnibus tests. Unfortunately, these plots usually are not suggestive of corrective measures. However, sometimes the source of discordance can be pinpointed through the identifiction of a few influenial or outlying observations.

Note that the cumulative hazard of the ith individual at its moment of failure (if it is uncensored) is $\exp[\boldsymbol{\beta}' \boldsymbol{z}_i(T_i)]\Lambda_0(T_i)$, if the PH model holds. The distribution of this is easily seen to be unit exponential. Therefore these random variables may be thought of as censored samples from this distribution, provided the PH model is valid. Replacing $\boldsymbol{\beta}$ and Λ_0 by their estimators, one can device a graphical check for the goodness of fit of the unit exponential distribution (see Section 3), which will indicate the goodness of overall fit of the PH regression model. This test was proposed by Kay (1977). Lagakos (1981) suggested a diagnostic based on a permutation of the 'observed' rank vector (which accounts for covariates) which should be plotted against each covariate. Any pattern in this plot would indicate a lack of fit.

Another option is to plot *residuals* against their indices. The diagnostic of Kay may be called a *generalized residual* in the sense of Cox and Snell (1968), although it does not have the 'observed – expected' interpretation. Several other residuals have been proposed. A *failure time* residual can be constructed as

$$r_i = T_i\delta_i - \sum_{k=1}^{n} T_k \delta_k \widehat{p}_i(T_k),$$

where δ_i is the censoring indicator for individual i and $\widehat{p}_i(t)$ is the estimated relative risk, that is,

$$\widehat{p}_i(t) = Y_i(t)\exp[\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i(t)] \left(\sum_{l=1}^{n} Y_l(t)\exp[\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l(t)] \right)^{-1}.$$

The residual brings out the difference between the observed failure time of an individual with a weighted average of the other observed failure times, where the estimated relative risk of each individual at its time of failure serves as the weight. Similarly, a *failure count* residual would be

$$c_i = \delta_i - \sum_{k=1}^{n} \delta_k \widehat{p}_i(T_k).$$

It is easily seen that this coincides with the martingale residual $\widehat{M}_i(\tau)$ of Section 5.2. Noting that the martingale residuals have a skewed range, $(-\infty, 1]$, Therneau et al. (1990) proposed *deviance* residuals

$$d_i = \text{sgn}(\widehat{M}_i)[-2\{\widehat{M}_i + \delta_i \log(\delta_i - \widehat{M}_i)\}]^{1/2}.$$

Two other residuals, called the *score residual* and *partial residual* will be described later. Barlow and Prentice (1988) pointed out that f_i and c_i belong to a general class of residuals of the form $f_i\delta_i - \sum_{k=1}^{n} f_k\delta_k\widehat{p}_i(T_k)$, and provide scaling procedures for them. Plots of the standardized residuals against the ranks of the failure time are expected to indicate the overall goodness of fit of the assumed model. The ideal scenario is that of no pattern in the plot. Index plots of these residuals may also be used. Sometimes one or two residuals stand out of the rest, indicating that the model may be inadequate for these.

Often a few observations influence the fit in a significant way, and in the process get themselves fitted very well. Since the residuals described above are likely to be small for these observations, special tools are needed to identify them. It is computationally rather prohibitive to consider the actual amounts of change in the estimated regression coefficients after dropping one observation at a time, although ideally this would bring out the influence of the individual cases on these estimates. There are two alternative approaches. Starting from the current set of estimates, one may take the first step of the Newton-Raphson iterations towards the estimate corresponding to the data set with one observation deleted. The resulting change in the estimator can be taken as a diagnostic of the influence of the deleted observation. Storer and Crowley (1985) observed that the effect of dropping an observation on the regression coefficients is the same as adding a new binary covariate which distinguishes one observation from the rest. Motivated by this equivalence, they proposed a diagnostic which corresponds to the first step of iterations in the reformulated problem. This diagnostic is easier to compute than the one-step diagnostic in the usual set-up. The second approach is to consider only an infinitesemal step towards the deletion of the ith observation. This gives rise to the empirical influence functions for the Cox regression parameters, derived by Cain and Lange (1984). The vector of influence functions when the ith case is dropped is $\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{v}_i$, where $\mathcal{I}(\boldsymbol{\beta})$ is the observed information matrix and

$$\boldsymbol{v}_i = \int_0^\tau \left[\boldsymbol{z}_i(t) - \sum_{l=1}^{n} \boldsymbol{z}_l(t)\widehat{p}_l(t)\right] \{dN_i(t) - \widehat{p}_i(t)dN(t)\},$$

where $N(\cdot)$ is the counting process for the entire population. The quantity in the squared brackets can be interpreted as the difference between the ith covariate and the mean covariate value of the risk set at time t. The

quantity v_i is seen to be a special case of the class of residuals considered by Barlow and Prentice (1988, see above). It may be called a *score residual* in the following sense: if the above integral is evaluated from $t-$ to t, one gets the the contribution of the ith individual to the score vector at time t. Summing this over all the failures at distinct failure times produces the respective contributions of these time points to the score vector. A special case of the latter (for constant covariates) is the *partial residual* of Schoenfeld (1982). The diagnostics proposed by Reid and Crépeau (1985) are identical to the influence functions of Cain and Lange.

The index plot of any of these diagnostics is expected to reveal the influence of individual cases on the estimated coefficient of a given covariate.

Most of the diagnostic plots mentioned in this section are applicable to the Generalized Cox model (Andersen and Gill, 1982) for recurrent events. This model is useful in analyzing competing risks data, multiple failures data and transition time data in a Markov chain. Thus the relevance of the methods discussed here transcend the domain of Survival Analysis.

Diagnostic plots for other regression models such as the linear model, the accelerated failure time model and the proportional odds-ratio model are also available in the literature. These plots will not be discussed here.

6 Improving readability of some diagnostic plots

It has been mentioned before that some form of reference is needed to aid the visual assessment of a diagnostic plot. This may be achieved by the overlaid plot of a few samples of the 'ideal' plot through simulations. Several examples of this technique may be found in Fleming and Harrington (1991). A simpler and more traditional way of providing the reference is through pointwise confidence limits or confidence bands. The user is expected to judge how far the ideal plot (usually a straight line) is from these bands *in the vertical direction.* This is alright if the line is horizontal, but the same cannot be said about diagonal 'ideal' lines. As pointed out earlier, the human eye tends to confuse horizontal separation with vertical one. If one were to draw reference curves based on the *horizontal separation* of the estimated curve and the ideal line, say in the one-sample P-P plot, a different set of reference curves would emerge. Specifically, the *right* reference curve would be the graph of $\widehat{F}(t)$ vs. $\widehat{F}_0(t) + c(t)[\widehat{F}(t) - \widehat{F}_0(t)]$, while the *lower* reference curve would be the graph of $\widehat{F}(t) - c(t)[\widehat{F}(t) - \widehat{F}_0(t)]$ vs. $\widehat{F}_0(t)$, where $c(t)$ is a common coefficient representing the confidence level. The vertical reference curves are more misleading when one is supposed to interpret the type of departure from the straight line (say, a convex or star-shape).

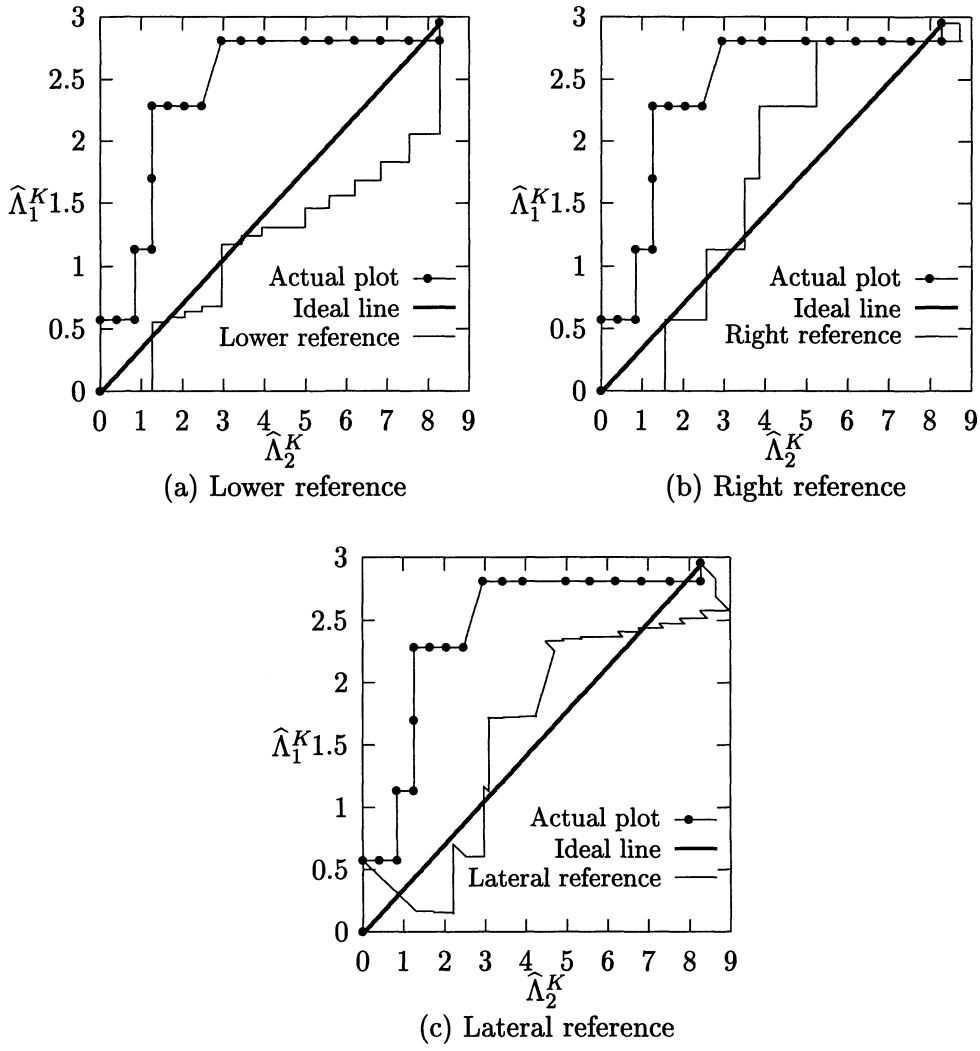
It is proposed that in all the diagnostic plots where the reference is a straight line, the reference curves be drawn on the basis of the *lateral*

separation of the ideal and estimated curves. This principle endorses the usual pointwise confidence limits and confidence bands when the reference line is horizntal. In the other cases, the distribution of the *normal* distance of the estimated curve from the ideal one has to be considered, and the resulting reference points for each individual should be plotted along the same direction. An example of this is the set of bootstrap confidence bands for the two-sample P-P plot considered by Campbell (1994).

The above principle is now illustrated for the plot due to Gill and Schumacher (1987) to test the proportionality of hazards, using the Ovarian Cancer data set of Fleming et al. (1980). It is suggested through the analytical tests of Gill and Schumacher (1987) and Deshpande and Sengupta (1995) that the failure rate of the patients at stage IIA (called Group 1 here) has an increasing ratio with that of the stage II patients (Group II). The plot of $\widehat{\Lambda}_1^K$ vs $\widehat{\Lambda}_2^K$ given by Gill and Schumacher, who used the straight line with slope $\widehat{\theta}_K$ as the 'ideal' line, shows a concave trend. Therefore a single reference curve is considered here instead of a pair. Figure 5(a) shows the plot of $\widehat{\Lambda}_1^K$ vs $\widehat{\Lambda}_2^K$ along with the lower reference curve for the vertical separation, corresponding to the one-sided 95% pointwise confidence limits calculated from the limiting distribution given in Section 4.2. The weight function used here is $K(t) = Y_1(t)Y_2(t)(Y_1(t) + Y_2(t))^{-1}$, where Y_i is the number at risk from Group i at time t, $i = 1, 2$. Figure 5(b) shows the right reference curve, obtained in the same manner by interchanging the populations. It appears that the 'ideal line' by and large lies above the lower reference curve, but strays beyond the right reference curve after a certain time. This is confusing, as each reference curve is as appropriate as the other.

In order to derive a *lateral* reference curve, note that the separation of the two curves in the direction normal to the straight line is $(\widehat{\Lambda}_1^K(\tau)\widehat{\Lambda}_2^K(t) - \widehat{\Lambda}_1^K(t)\widehat{\Lambda}_2^K(\tau))[m_y^2(\widehat{\Lambda}_1^K(\tau))^2 + m_x^2(\widehat{\Lambda}_2^K(\tau))^2]^{-1/2}$, where m_x and m_y are the scales in the horizontal and vertical directions, respectively. If these scales are expressed in units per centimeter, the above separation is in centimeters. The marginal distribution of this separation (at a given time t) can be found as in Section 4.2. The resulting reference curve is shown in Figure 5(c). The ideal line is away from this curve near the middle range. Since the jumps in the lateral direction have a horizontal and a vertical component, both the coordinates of the lateral curve change at the jump-points of $\widehat{\Lambda}_1^K$ and $\widehat{\Lambda}_2^K$. In fact, consecutive vertical jumps of the lower reference curve and the consecutive horizontal jumps of the right reference curve make them difficult to interpret. The lateral reference curve has no such problem.

In order to judge whether the overall plot is concave, a lateral confidence *band* may have been more useful. The pointwise confidence limits have only been used for the purpose of illustration.



(a) Lower reference

(b) Right reference

(c) Lateral reference

Figure 5: Plot of $\widehat{\Lambda}_1^K$ vs $\widehat{\Lambda}_2^K$ with three reference curves

References

1. Andersen, P.K. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics* **38**, 67–77.
2. Andersen, P.K. (1983). Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics* **10**, 77–85.
3. Andersen, P.K. and Borgan, Ø. (1985). Counting process models for life history data: a review. *Scandinavian Journal of Statistics* **12**, 97–158.

4. Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.
5. Antoniadis, A. (1989). A penalty method for nonparametric estimation of the intensity function of a counting process. *Annals of the Institute of Statistical Mathematics* **41**, 781–807.
6. Arjas, E. (1988) A graphical method for assessing goodness-of-fit in Cox's proportional hazard model. *Journal of the American Statistical Association* **83**, 204–212.
7. Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Mathematical Models*, Holt, Reinhart and Winston: New York.
8. Barlow, W.E. and Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika* **75**, 65–74.
9. Bie, O., Borgan, Ø. and Listol, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics* **14**, 221–233.
10. Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.
11. Cain, K.C. and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* **40**, 493–499.
12. Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* **13**, 499–508.
13. Cleveland, W.S. and McGill, R. (1984). Graphical perception: theory, experimentation and application to the development of graphical methods.
14. Cox, D.R. and Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* **30**, 248–275. *Journal of the American Statistical Association* **79**, 531–554.
15. Csorgo, S. and Horvath, L. (1986). Confidence bands from censored samples. *Canadian Journal of Statistics* **14**, 131–144.
16. Csorgo, M., Csorgo, S. and Horvath, L. (1987). Estimation of total time on test transforms and Lorenz curves under random censorship. *Mathematische Operationforschung und Statistik, Series Statistics* **18**, 77–97.
17. Dabrowska, D., Doksum, K. and Song, J.-K. (1989). Graphical comparison of cumulative hazards for two populations. *Biometrika* **76**, 763–773.

18. Dabrowska, D., Doksum, K., Feduska, N.J., Husing, R. and Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-parametric hazard model. *Statistics in Medicine* **11**, 1465–1476.
19. D'Agostino, R.B. (1986). Graphical analysis, in *Goodness-of-fit Techniques*, eds. R.B. D'Agostino and M.A. Stephens, Marcel Dekker: New York.
20. Deshpande, J.V. and Sengupta, D. (1995). Testing the hypothesis of proportional hazards in two populations. To be published in *Biometrika*.
21. Fleming, T.R., O'Fallon, J.R., O'Brien, P.C., and Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily censored data. *Biometrics.* **36**, 607–625.
22. Fleming, T.R. and Harrington, D.P. (1991). *Counting Process and Survival Analysis*, John Wiley: New York.
23. Freirich, E.J. et al. (1963). The effect of 6-Mercaptpurine on the duration of steriod-induced remissions in acute leukemia. *Blood* **21**, 699–716.
24. Gentlemen, R. and Crowley, J. (1991). Graphical methods for censored data. *Journal of the American Statistical Association* **86**, 678–683.
25. Ghorai, J.K. and Rejtoe, L. (1987). Estimation of mean residual life with censored data under the proportional hazards model. *Communications in Statistics: Theory & Methods* **16**, 2097–2114.
26. Gill, R.D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.
27. Goldman, A.I. (1992). Eventcharts: Visualizing survival and other timed-events data. *The American Statistician* **46**, 13–18.
28. Guess, F. and Park, D.H. (1991). Nonparametric confidence bounds, using censored data, on the mean residual life. *IEEE Transactions on Reliability* **40**, 78–80.
29. Guilbaud, O. (1988). Exact Kolmogorov-type tests for left-truncated and/or right censored data. *Journal of the American Statistical Association* **83**, 213–221.
30. Heller, G. and Simonoff, J.S. (1992). Prediction in censored survival data: a comparison of the proportional hazards and linear regression models. *Biometrics* **48**, 101–115.
31. Hollander, M. and Peña, E. (1989). Families of confidence bounds for the survival function under the general random censorship model and the Koziol-Green model. *Canadian Journal of Statistics* **17**, 59–74.
32. Kalbfleisch, J.D. and Prentice, R.L. (1980), *The statistical analysis of failure time data*, New York: Wiley.
33. Kay, R. (1977). Proportional hazards regression models and the analysis of censored survival data. *Applied Statistics* **26**, 227–237.

34. Klefsjö, B. (1982). On ageing properties and total time on test transforms. *Scandinavian Journal of Statistics* **9**, 37–41.

35. Kunitz, H. and Pamme, H. (1991). Graphical tools for life time data analysis. *Statistische Hefte* **32**, 85–113.

36. Lagakos, S.W. (1981). The graphical evaluation of explanatory variables in proportional hazards regression models. *Biometrika* **68**, 93–98.

37. Lee, L. and Pirie, W.R. (1981). A graphical method for comparing trends in series of events. *Communications in Statistics, Theory and Methods* **10**, 827–848.

38. Lin, D.Y., Wei, L.J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

39. Lu H.H.S., Wells, M.T. and Tiwari, R.C. (1994). Inference for shift functions in the two-sample problem with right-censored data: with applications. *Journal of the American Statistical Association* **48**, 1017–1028.

40. Michael, J.R. (1983). The stabilized probability plot. *Biometrika* **70**, 11–17.

41. Michael, J.R. and Schucany, W.R. (1986). Analysis of data from censored samples, in *Goodness-of-fit Techniques*, eds. R.B. D'Agostino and M.A. Stephens, Marcel Dekker: New York.

42. Miller, R.G. (1981). *Survival Analysis*, John Wiley: New York.

43. Park, B.G., Sohn, J.K. and Lee, S.B. (1993). Nonparametric estimation of the mean residual life function. *Journal of the Korean Statistical Association* **22**, 147–157.

44. Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics* **11**, 453–466.

45. Reid, N. and Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72**, 1–9.

46. Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*, John Wiley: New York.

47. Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.

48. Schumacher, M. (1984). Two-sample tests of Cramér-von Mises and Kolmogorov-Smirnov type for randomly censored data. *International Statistical Review* **52**(3), 263–281.

49. Sengupta, D. and Deshpande, J.V. (1994). Some results on the relative aging of two life distributions. *Journal of Applied Probability* **31**.

50. Senthilselvan, A. (1987). Penalized likelihood estimation of hazard and intensity function. *Journal of the Royal Statistical Society, Series B* **49**, 170–174.
51. Storer, B.E. and Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods. *Journal of the American Statistical Association* **80**, 139–147.
52. Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.
53. Tufte, E.R. (1983). *The Visual Display of Quantitative Information*, Graphics Press: Stamford, CT.
54. Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* **86**, 169–173.
55. Waller, L.A. and Turnbull, B.W. (1992). Probability plotting with censored data. *The American Statistician* **46**, 5–12.
56. Zhou, M. (1992). *A scatterplot for censored data.* Technical Report # 309, Department of Statistics, University of Kentucky (also presented at the 1992 Spring Joint Meeting of ASA, IMS and ENAR at Cincinnati, Ohio).