Survival Analysis

2. Non-Parametric Estimation

Germán Rodríguez

Princeton University

February 12, 2018

Overview

We now consider the analysis of survival data without making assumptions about the form of the distribution. We will review

1. The Kaplan-Meier estimator of the survival curve and the Nelson-Aalen estimator of the cumulative hazard.
2. The Mantel-Haenszel test and other non-parametric tests for comparing two or more survival distributions.
3. Cox's proportional hazards model and the partial likelihood, including time-varying covariates and time-dependent or non-proportional effects,

Later we will discuss flexible semi-parametric models that represent a compromise between fully parametric and non-parametric alternatives.

Kaplan-Meier

If there is no censoring, the obvious estimate of the survival function is the empirical survival function or proportion alive at t

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i > t).$$

Kaplan and Meier (1958) extended the estimator to right-censored and left-truncated data by focusing on conditional survival

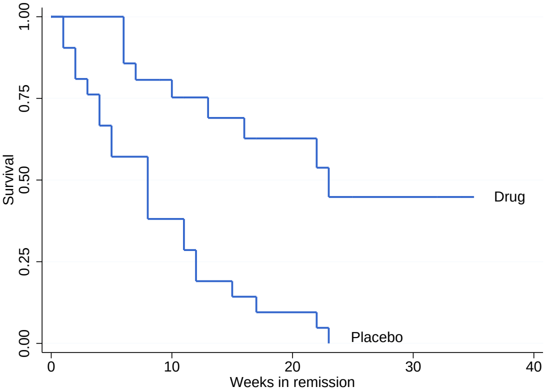
$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

where $t_1 < t_2 < \cdots < t_m$ are the distinct failure times, d_i is the number of failures at t_i, and n_i the number at risk or alive just before t_i.

The estimator is intuitively appealing, and reduces to the empirical survival function if there is no censoring or truncation.

Kaplan-Meier

The Kaplan-Meier estimator is a step function with discontinuities at the failure times. If the largest observation time is censored the curve doesn't drop to zero and is undefined after the last censoring



These are the famous Gehan data on duration of remission in leukemia patients in treated and control groups

Greenwood

Standard errors may be derived using a binomial argument and the delta method, as shown in the notes. This leads to

$$\text{var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

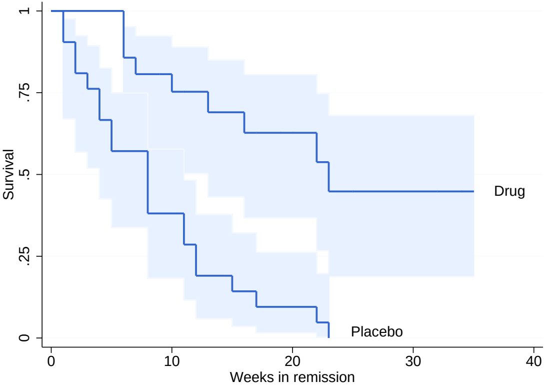
a formula derived by Greenwood for life tables in 1926! If there is no censoring/truncation it equals the standard binomial variance. This result can be used to compute pointwise confidence bands around the estimate. To avoid values outside (0,1) and improve the normal approximation it is better to work with the log-log transformation and its variance

$$\text{var}(\log(-\log \hat{S}(t))) = \frac{\text{var}(\log \hat{S}(t))}{(\log \hat{S}(t))^2}$$

which gives good results even for small samples.

Greenwood (continued)

Here are pointwise confidence bounds for the Gehan data.



Note that there is some overlap at low durations, but as we'll see, the curves are significantly different.

Nelson-Aalen

To estimate the cumulative hazard we could use $-\log \hat{S}(t)$. A direct estimator due to Nelson and Aalen is

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i}$$

This estimator is closely related to the theory of counting processes, representing the expected number of events in $(0, t]$ for a unit permanently at risk. This interpretation is particularly useful for recurrent events.

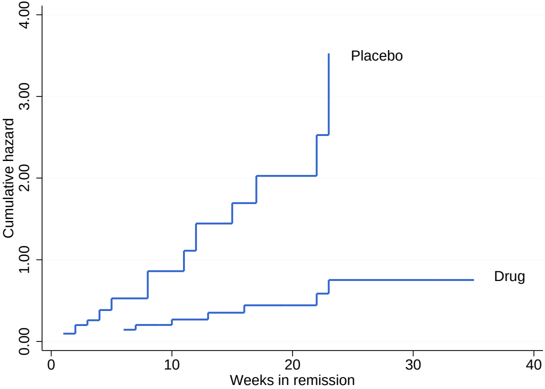
The variance of the Nelson-Aalen estimator follows from a Poisson argument

$$\text{var}(\hat{\Lambda}(t)) = \sum_{i:t_i \leq t} \frac{d_i}{n_i^2}$$

The normal approximation is improved if one works instead with the log of the cumulative hazard.

Nelson-Aalen (continued)

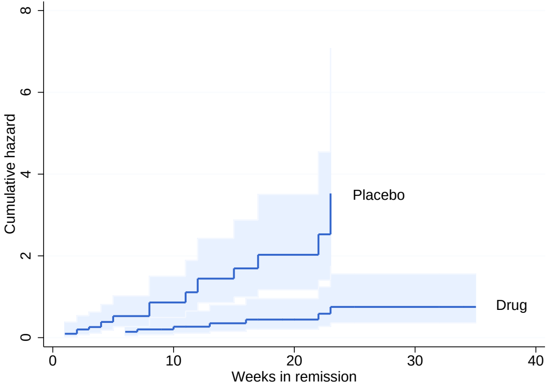
Here are the estimated cumulative hazards for the Gehan data



We see how the risk accumulates much faster in the control groups. Less clearly, the approximate linearity suggests a relatively constant risk at early durations in both groups.

Nelson-Aalen (continued)

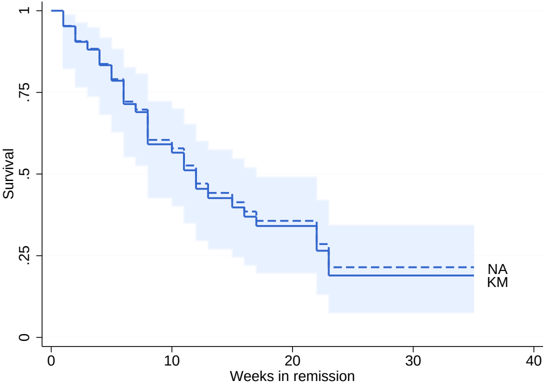
Here are 95% confidence bounds for the Nelson-Aalen estimates by group, calculated in the log-scale and then converted back to cumulative hazards.



There is a lot of uncertainty about the cumulative hazard in the control group after 20 weeks.

Kaplan-Meier vs Nelson-Aalen

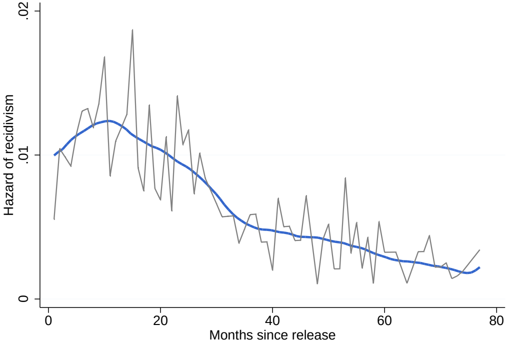
Breslow noted that one could estimate the survival as $e^{-\hat{\Lambda}(t)}$ starting from a NA estimate. This is usually very close to the KM estimate, as shown here for the pooled Gehan data.



Convention, however, is to use Kaplan-Meier for survival and Nelson-Aalen for cuulative hazards.

Estimating the Hazard

This is hard! if you difference the Nelson-Aalen estimate, or minus the log of the Kaplan-Meier estimate, you get a rough estimate with spikes at the failure times. Stata uses a kernel smoother



Alternative Epanechnikov kernel

Alternatives are splines and other smoothers. We will return to this issue when we consider flexible parametric models.

Mantel-Haenszel

We now compare two survival curves, such as treated and controls.

Imagine setting up a 2 × 2 table for each distinct failure time with the results by group. Let d_{ij} denote the number of failures and n_{ij} the number at risk at time t_i in group j

	Fail	Survive	
Treated	d_{i1}		n_{i1}
Control			n_{i2}
	d_i	$n_i - d_i$	n_i

The conditional distribution of d_{ij} given both margins, that is given the number at risk in each group and the total number of failures at t_i, is hypergeometric, with mean and variance

$$E(d_{ij}) = d_i \frac{n_{ij}}{n_i} \quad \text{and} \quad \text{var}(d_{ij}) = \frac{n_{ij}(n_i - n_{ij}) d_i (n_i - d_i)}{(n_i - 1) n_i^2}$$

Intuitively, if the survival curves were the same we would expect the number of deaths in each group to be proportional to the number at risk in each group.

Mantel-Haenzsel

The Mantel-Haenszel statistic is obtained by summing over all distinct failure times the number of observed and expected failures and the variances

$$T = \frac{D^2}{V} \quad \text{where} \quad D = \sum_i (d_{i1} - E(d_{i1})) \quad \text{and} \quad V = \sum_i \text{var}(d_{i1})$$

The asymptotic distribution of the statistic is χ^2_1.

For example in the Gehan data the treated group has 9 failures where one would expect 19.25. The variance is 6.26, leading to a highly significant χ^2 statistic of 16.79 on one d.f.

It doesn't matter which of the two groups is used in the calculation, the result is the same. The controls have 21 failures where one would expect 10.75 and the variance is the same 6.26.

Weighted Log-Rank Tests

The Mantel-Haenszel test gives equal weight to failures at each time and is optimal when the hazards are proportional. Sometimes one might have reason to give more weight to differences at earlier or later times. The table below shows some alternatives:

Test	Weight	χ^2
Mantel-Haenszel or log-rank	1	16.79
Wilcoxon-Breslow-Gehan	n	13.46
Tarone-Ware	$\sqrt{n}$	15.12
Peto-Peto-Prentice	$\tilde{S}(t_j)$	14.08
Fleming-Harrington	$\hat{S}(t_{j-1})^p(1-\hat{S}(t_{j-1}))^q$	14.45

All these test use the statistic $T = D^2/V$ where

$$D = \sum w_i(d_{ij} - E(d_{ij})) \quad \text{and} \quad V = \sum w_i^2 \text{var}(d_{ij}).$$

Peto's $\tilde{S}(t)$ is like K-M but divides by $n+1$ instead of n. For FH I set $p = 1$ and $q = 0$.

The k-sample Test

These tests extend to more than two groups. In addition to the expected values and variances we need the covariances of counts of events in groups r and s, which are given by

$$\operatorname{cov}(d_{ir}, d_{is}) = -\frac{d_i(n_i - d_i)}{n_i - 1}\frac{n_{ir}n_{is}}{n_i^2}$$

For each time t_i we now have a vector with counts of failures by group, its expected value and its variance covariance matrix. We sum over all distinct failure times and construct a quadratic form

$$Q = D'V^{-}D \quad \text{where} \quad D = \sum_i (d_i - E(d_i)) \quad \text{and} \quad V = \sum_i \operatorname{var}(d_i)$$

where V^{-} is a generalized inverse of V, obtained for example by omitting one of the groups from V.

The large sample distribution of the statistic is χ^2 with d.f. equal to the number of groups minus one. Weights are incorporated just as in the two-sample case.

The k-sample Test (continued)

Let us compute a 3-group test "by hand" using data from the Stata manual on days to tumor formation in three groups of animals exposed to carcinogenic agents.

The observed and expected counts of events in each group and the variance-covariance matrix, summed over the nine distinct failure times, are

$$O = \begin{pmatrix} 4 \\ 6 \\ 5 \end{pmatrix}, \quad E = \begin{pmatrix} 6.41 \\ 6.80 \\ 1.79 \end{pmatrix}, \quad \text{and} \quad V = \begin{pmatrix} 2.70 & & \\ -2.02 & 2.66 & \\ -0.68 & -0.64 & 1.32 \end{pmatrix}$$

The quadratic form using a Moore-Penrose generalized inverse is

$$D'V^{-}D = 8.05$$

a significant χ^2 on 2 d.f. with a p-value of 0.018. The same result is obtained by omitting the last row of D and the last row and column of V.

Software Notes

Stata can compute all of the estimates we have discussed using various subcommands of the `sts` command, including `sts list` for Kaplan-Meier and Nelson-Aalen estimates, `sts graph` for plots of the survival and hazard functions, and `sts test` for Mantel-Haenszel and weighted log-rank tests. The data must be `stset` first.

R's `survival` package provides functions `survfit()` to compute Kaplan-Meier estimates, with Nelson-Aalen computed "by hand", and `survdiff()` to test equality of survival curves using Mantel-Haenszel or weighted log-rank tests. The function `Surv(t,d)` is used to specify the time variable and the failure indicator.

The logs at `http://data.princeton.edu/pop509/gehan.html` analyze the Gehan data using both Stata and R.

Cox Regression

Consider now estimating a proportional hazards model

$$\lambda(t|x) = \lambda_0(t) e^{x'\beta}$$

without making any assumptions about the baseline hazard.

Cox proposed looking at each failure time and computing a conditional probability of failure given the observations at risk at that time. If there are no ties the probability for t_i is

$$\frac{\lambda_0(t_i) e^{x_i'\beta}}{\sum_{j \in R_i} \lambda_0(t_i) e^{x_j'\beta}} = \frac{e^{x_i'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}}$$

where R_i denotes the risk set at t_i.

Note that the baseline hazard cancels out and we get a conditional probability that depends only on β.

The Partial Likelihood

Cox then proposed treating the product of these conditional probabilities as if it was a likelihood. He called it a conditional likelihood and later more correctly a *partial* likelihood

$$L = \prod_i \frac{e^{x_i'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}}$$

Kalbfleisch and Prentice showed that if the covariates are fixed over time this is the *marginal* likelihood of the ranks of the observations, where you consider just the order in which failures occur instead of the actual times.

More rigourous justification was provided later in terms of the theory of counting processes, see Andersen et al. (1993) for details. Great intuition proven right!

Notably, the partial likelihood is identical to that of a conditional logit model!

Score and Information

Maximization of this likelihood is not difficult. The log-likelihood is

$$\log L = \sum_i \{x_i'\beta - \log \sum_{j \in R_i} e^{x_j'\beta}\}$$

The score or first derivative can be shown to be

$$U(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_i \{x_i - A_i(\beta)\}$$

where $A_i(\beta)$ is the weighted mean of the covariates over the risk set using the relative risks as weights.

The observed information or negative second derivative is

$$I(\beta) = -\frac{\partial^2 \log L}{\partial \beta \beta'} = \sum_i C_i(\beta)$$

where $C_i(\beta)$ is the weighted covariance of the covariates over the risk set, with the relative risks as weights.

The Problem of Ties

So far we assumed no ties. If there are lots of ties the data are probably discrete or have been grouped, and Cox regression is not appropriate. Otherwise one can adjust the partial likelihood using one of four methods

1. Cox's exact partial likelihood looks at all possible ways of selecting d_i failures out of the risk set R_i. This is computationally very intensive.
2. The "exact" marginal likelihood of the ranks can be computed by numerical integration. Not as difficult, but still demanding.
3. Breslow's approximation treats the tied failures as coming from the same risk set. This is the quickest method.
4. Efron's approximation considers all possible ways of breaking the ties and adjusts the risk set accordingly. This turns out to be reasonably fast and remarkably accurate.

The Problem of Ties (continued)

Consider a simple example where the risk set consists of $\{1, 2, 3, 4\}$ and $\{1, 2\}$ are observed to fail. Let i have relative risk $r_i = e^{x_i'\beta}$.

Cox looks at all possible ways of choosing two to fail and computes

$$\frac{r_1 r_2}{r_1 r_2 + r_1 r_3 + r_1 r_4 + r_2 r_3 + r_2 r_4 + r_3 r_4}$$

Breslow keeps the risk set constant after a failure, writing

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)^2}$$

Efron notes that if $\{1\}$ fails first the risk set becomes $\{2, 3, 4\}$, but if $\{2\}$ fails first it becomes $\{1, 3, 4\}$ so he averages

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)(\frac{r_1 + r_2}{2} + r_3 + r_4)}$$

I strongly recommend Efron when an exact calculation is not feasible. Stata uses Breslow and R uses Efron as the default.

Cox and the Gehan Data

This is the actual example used in Cox's 1972 paper.

Fitting a proportional hazards model using the exact partial likelihood (option `exactp` in Stata and `ties="exact"` in R gives an estimate of 1.628, equivalent to a relative risk or hazard ratio of 5.095.

At any time since remission the risk of relapse for patients in the control group is 5 times the risk of treated patients. The drug is remarkably effective.

The computing logs use an indicator for treated, so the coefficient has the opposite sign: -1.628, which translates into a risk ratio of 0.196.

This means that the drug reduces the risk of relapse by 80% at any given duration since remission.

Obviously the choice of reference category is up to the researcher. Here I follow the original paper.

Ties and the Gehan Data

Here's how the treatment of ties affects the results:

	Breslow	Efron	Marginal	Partial
e^{β}	4.52	4.82	4.94	5.09
β	1.51	1.57	1.60	1.63
z	3.68	3.81	3.79	3.76

As you can see, Efron comes much closer to the exact partial likelihood.

Cox himself got $\hat{\beta} = 1.65$ or a risk ratio of 5.21 by evaluating the exact partial likelihood *by hand* over a grid of values!

The table above also shows the usual z-tests obtained as the ratio of the estimate to its standard error. This is just one of several possible tests.

Testing Hypotheses

There are three ways to test hypotheses in Cox models

1. Likelihood ratio tests, comparing the partial likelihoods of nested models. Usually requires fitting two models
2. Wald tests, based on the large sample distribution of partial likelihood estimates

   $$\hat{\beta} \sim N(\beta, I^{-1}(\beta))$$

   Can be computed by fitting just one model.
3. Score tests, based on the large sample distribution of the first derivative of the partial likelihood

   $$U(\beta) \sim N(0, I(\beta))$$

   Sometimes can be calculated without fitting any models!

These tests are all asymptotically equivalent, but usually we prefer likelihood ratio tests. I mention the score test because it happens to be equivalent to the Mantel-Haenszel test!

Testing Hypotheses (continued)

Here are the results of these tests using the exact partial likelihood:

- the likelihood ratio, which compares the model with two groups with the null, is 16.8 on one d.f. (This is our preferred test.)
- the Wald test is $z = 3.76$, and is equivalent to a χ^2 of 14.1 on one d.f. based on the asymptotic normality of the estimator
- the score test is based on the fact that the score has a normal distribution with variance given by the information matrix. Under the hypothesis of no group difference the score is 10.25 and the information is 6.2570, yielding a chi-squared of 16.79. Looks familiar? Details of the calculation appear in Table 2 of Cox's paper.

The group difference is clearly not due to chance.

Baseline Survival

We now consider estimation of the baseline survival given a partial likelihood estimator of β.

An argument similar to Kaplan and Meier's leads to estimating the baseline survival when there are no ties as

$$\hat{S}_0(t) = \prod_{i:t_i \leq t} (1 - \frac{e^{x_i'\hat{\beta}}}{\sum_{j \in R_i} e^{x_j'\hat{\beta}}})^{e^{-x_i'\hat{\beta}}}$$

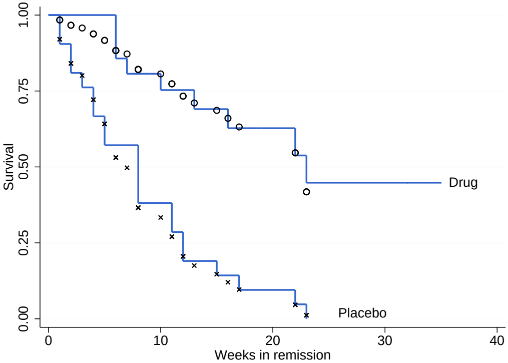
which is just like K-M but with the relative risks as weights. The exponential on the right scales the probability for observation i into a baseline.

If there are ties an iterative procedure is required, but the underlying logic is the same. See the notes for details.

The resulting estimator is a step function with drops at the observed failure times.

Baseline Survival (continued)

The figure below reproduces Figure 1 in Cox's original paper, showing the estimated survival for the treated and control groups under proportional hazards, overlaid on separate K-M estimates



Stata's `stcoxkm` can do essentially this plot (but is not as pretty :)