# Kaplan-Meier and Cox

Pop 502 / Eco 572/ Soc 532 • Spring 2017

We consider non-parametric estimation of the survival function using cohort data. Specifically, we assume we have observations $t_1, \ldots, t_n$ of survival times as well as indicators $d_1, \ldots, d_n$ that take the value 1 if the observation ended with the event of interest and 0 otherwise.

## One-Sample: Kaplan-Meier

If there was no censoring the obvious estimate of the probability of surviving to $t$ would be the empirical survival function

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

or proportion alive at $t$.

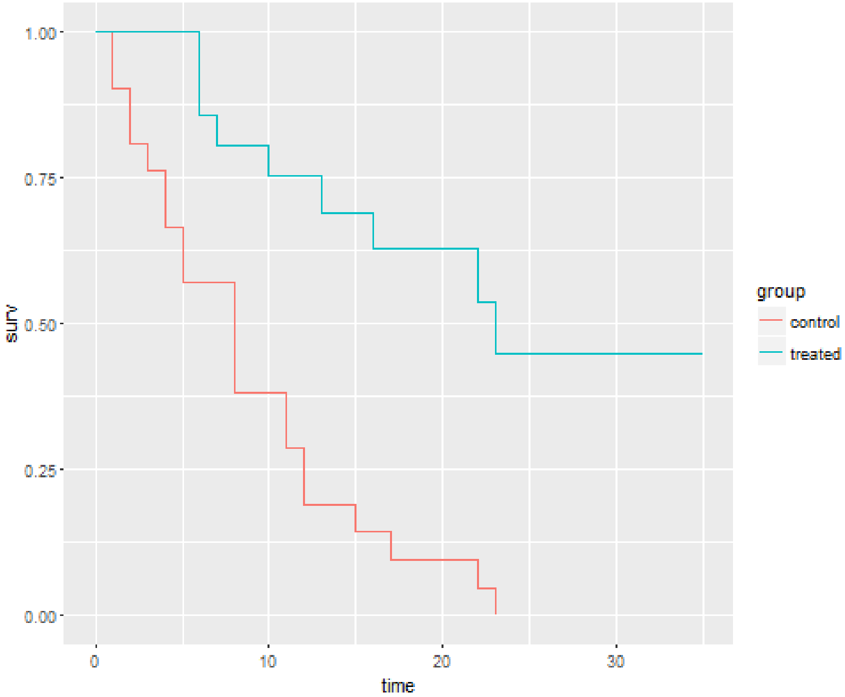
Kaplan-Meier extended the estimator to censored data. They focus on the *distinct ordered* event times (not counting censoring times), which I'll denote $t_{(i)}$. Let $d_i$ denote the number of events at $t_{(i)}$ and $n_i$ be the number alive, and hence at risk, *just before* $t(i)$. The Kaplan-Meier or *product limit* estimate is then

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

Note that $d_i/n_i$ estimates the probability of an event at $t_{(i)}$ given the number at risk (like ${}_{-n}q_x$) and one minus that or $1 - d_i/n_i$ is the probability of surviving that failure time conditional on survival up to that point, so the product is an unconditional survival probability up to $t$  (like $l_x$.)

The estimate is a step function with discontinuities at the observed failure times. If there is no censoring the estimator coincides with the empirical survival function, so it is a generalization for censored data.

In the website we compute Kaplan-Meier estimators for time in remission of leukemia patients in two groups, treated and controls. The figure below shows the estimated survival curves.  One group has no censoring and the estimate is just the proportion surviving to each duration; in the end all relapse. In the treated group we note that after 35 weeks almost half the patients remain in remission.



We can compute the standard error of the estimator using the delta method. Briefly, the method approximates the variance of a function of a random variable using a first-order Taylor series expansion, which gives

$$\mathrm{var}(f(X)) \approx [f'(X)]^2 \mathrm{var}(X)$$

In our case $\hat{S}(t)$ is a product, so we first take logs and assume independence of the conditional survival probabilities, so

$$\mathrm{var}(\log \hat{S}(t)) = \sum_{i:t_{(i)} \leq t} \mathrm{var}\ (\log p_i)$$

We estimate the variance of $p_i$ using the binomial formula, so $\mathrm{var}(p_i) = p_i q_i / n_i$ and then use the delta method to obtain

$$\mathrm{var}(\log p_i) = \frac{1}{p_i^2} \mathrm{var}(p_i) = \frac{q_i}{p_i n_i}$$

and then apply the delta method again, this time to go from $\log \hat{S}(t)$ to $S(t)$:

$$\mathrm{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{i:t_{(i)} \leq t} \frac{q_i}{p_i n_i}$$

This is known as Greenwood's formula and predates the Kaplan-Meier estimator by 32 years! It was first proposed in the context of actuarial life tables for cancer survival in 1926.

## Regression: Cox Proportional Hazards

Cox proposed a general solution to the problem of doing regression analysis with survival data without having to make strong assumptions about the shape of the hazard or force of mortality. I will use the standard statistical notation to emphasize the fact that this model has a wide range of applications beyond mortality.

The basic proportional hazards model assumes that

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

where $\lambda(t,x)$ is the hazard at time $t$ for a subject with covariate values $x$, $\lambda_0(t)$ is a *baseline* hazard that applies to everyone at time $t$ and $e^{x'\beta}$ is a *relative risk* for a subject with covariates values $x$ compared to a subject with $x = 0$.

A simple example may help fix ideas. Suppose there are only two groups and $x$ takes the value 1 for one group (say, treated) and 0 for the other (say, the control group). Then the model says

$$\lambda(t,x) = \begin{cases} \lambda_0(t), & \text{if } x = 0 \\ \lambda_0(t)e^{\beta}, & \text{if } x = 1 \end{cases}$$

In this case $\lambda_0(t)$ denotes the risk at time $t$ in the control group, and $e^{\beta}$ denotes the relative risk in the treated group at any given time, compared to the control group at the same time.

There are extensions of the model where the covariates may change over time, of their effects may be non-proportional, or both, but here we will focus on the simpler case.

Cox did not just contribute a model but also a way to estimate it without making any assumptions about the shape of the underlying hazard. Like previous workers, he focuses on the distinct ordered failure times $t_{(i)}$.

Suppose first that there are no ties in the observation times, so one and only one person fails at $t_{(i)}$. Let's call this person $j(i)$. Let $R_i$ denote the risk set, or indices of all subjects alive just before $t_{(i)}$. The probability that the person who failed at $t_{(i)}$ would be $j(i)$ conditional on the risk set is

$$L_i = \frac{\lambda(t_i, x_{j(i)})}{\sum_{j \in R_i} \lambda\,(t_i, x_j)}$$

If we write the risk as the product of the baseline risk times the relative risk, we find that the baseline hazard cancels out and the probability in question becomes

$$L_i = \frac{e^{x_{j(i)}'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}}$$

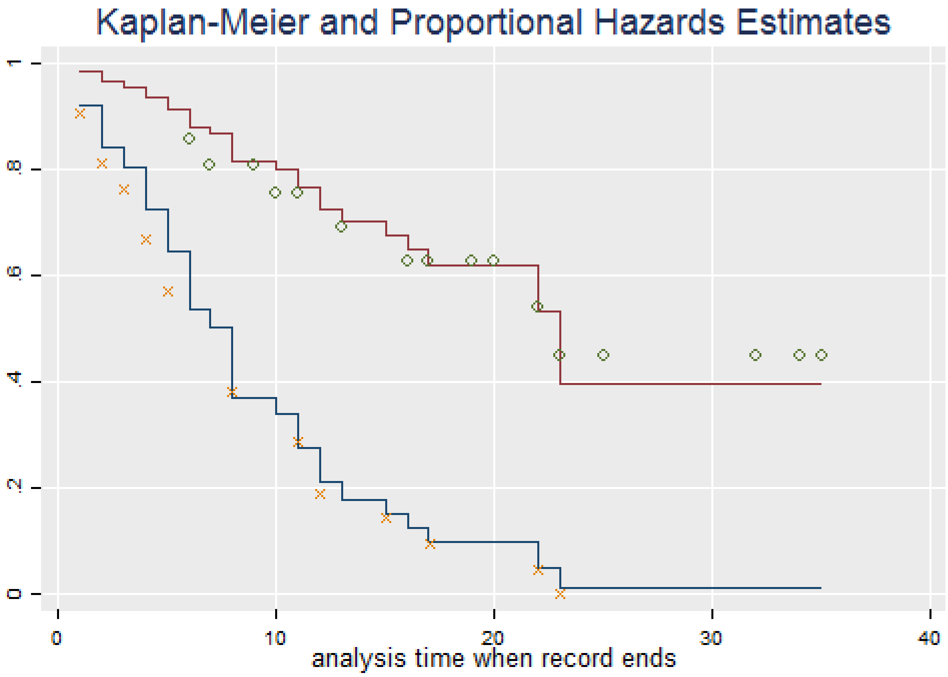
an expression that depends only on $\beta$. Cox proposed treating the product of these conditional probabilities over all distinct failure times as if it were a likelihood function, maximizing it to obtain an estimate of $\beta$.

The product is known as Cox's *partial likelihood* and the resulting estimator shares many of the optimal properties of maximum likelihood, with a small loss of efficiency compared to making full (and correct!) parametric assumptions.

Calculation of the estimate is more complicated if there are tied failure times. The numerator is simply the product of the relative risks over the $d_i$ who fail, but in principle the denominator requires considering all possible ways of selecting $d_i$ failures out of $n_i$ in the risk set, which may not be feasible. Not surprisingly, there are several approximations. The simplest one is Breslow's, which takes as denominator the sum of relative risks raised to the power $d_i$. Efron proposed a better approximation that requires only modest computational effort, and can be motivated by breaking the ties.

The website shows how to fit Cox's model to the leukemia remission data. We find a maximum partial likelihood estimate of -1.572 using Efron's method. Exponentiating this estimate we conclude that the risk of relapse is 79% lower in the treated group than in the controls at any duration of remission [exp(-1.572)= 0.208].

It is possible to obtain estimates of the baseline survival function by adapting the Kaplan-Meier logic *after* fitting a Cox model to obtain an estimate of $\beta$. The logic involves using the relative risks as weights. The figure below overlays Cox proportional-hazard estimates on the Kaplan-Meier estimates we obtained earlier, showing a good fit.



This is essentially Figure 1 in Cox's original paper. An alternative diagnostic plot in the log-log scale is shown on the website