Non-Parametric Estimation in Survival Models

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

We now discuss the analysis of survival data without parametric assumptions about the form of the distribution.

1 One Sample: Kaplan-Meier

Our first topic is non-parametric estimation of the *survival function*. If the data were not censored, the obvious estimate would be the empirical survival function

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

where I is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. The estimator is simply the proportion alive at t.

1.1 Estimation with Censored Data

Kaplan and Meier (1958) extended the estimate to *censored* data. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the distinct *ordered* times of death (not counting censoring times). Let d_i be the number of deaths at $t_{(i)}$, and let n_i be the number alive *just before* $t_{(i)}$. This is the number exposed to risk at time $t_{(i)}$. Then the Kaplan-Meier or *product limit* estimate of the survivor function is

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right).$$

A heuristic justification of the estimate is as follows. To survive to time t you must first survive to $t_{(1)}$. You must then survive from $t_{(1)}$ to $t_{(2)}$ given

that you have already survived to $t_{(1)}$. And so on. Because there are no deaths between $t_{(i-1)}$ and $t_{(i)}$, we take the probability of dying between these times to be zero. The conditional probability of dying at $t_{(i)}$ given that the subject was alive just before can be estimated by d_i/n_i. The conditional probability of surviving time $t_{(i)}$ is the complement $1 - d_i/n_i$. The overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t.

The Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times. Figure 1 shows Kaplan-Meier estimates for the treated and control groups in the famous Gehan data (see Cox, 1972 or Andersen et al., 1993, p. 22-23).

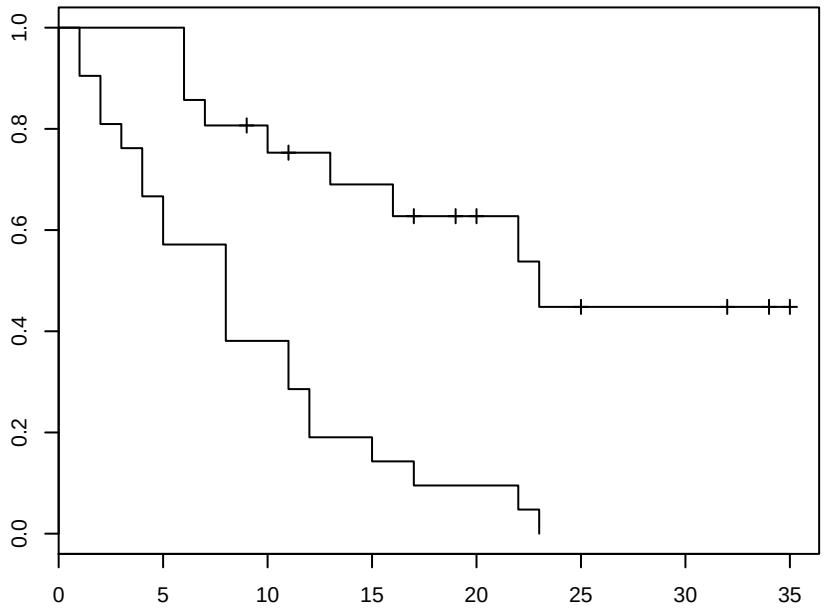

Figure 1: Kaplan-Meier Estimates for Gehan Data

If there is no censoring, the K-M estimate coincides with the empirical survival function. If the last observation happens to be a censored case, as is the case in the treated group in the Gehan data, the estimate is undefined beyond the last death.

1.2 Non-parametric Maximum Likelihood

The K-M estimator has a nice interpretation as a non-parametric maximum likelihood estimator (NPML). A rigorous treatment of this notion is beyond the scope of the course, but the original article by K-M provides a more intuitive approach. We consider the contribution to the likelihood of cases that die or are censored at time t.

- If a subject is censored at t its contribution to the likelihood is $S(t)$. In order to maximize the likelihood we would like to make this as large as possible. Because a survival function must be non-increasing, the best we can do is keep it constant at t. In other words, the estimated survival function doesn't change at censoring times.

- If a subject dies at t then this is one of the distinct times of death that we introduced before. Say it is $t_{(i)}$. We need to make the survival function just before $t_{(i)}$ as large as possible. The largest it can be is the value at the previous time of death or 1, whichever is less. We also need to make the survival at $t_{(i)}$ itself as small as possible. This means we need a discontinuity at $t_{(i)}$.

Let c_i denote the number of cases censored between $t_{(i)}$ and $t_{(i+1)}$, and let d_i be the number of cases that die at $t_{(i)}$. Then the likelihood function takes the form

$$L = \prod_{i=1}^{m} [S(t_{(i-1)}) - S(t_{(i)})]^{d_i} S(t_{(i)})^{c_i},$$

where the product is over the m distinct times of death, and we take $t_{(0)} = 0$ with $S(t_{(0)}) = 1$. The problem now is to estimate m parameters representing the values of the survival function at the death times $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$.

Write $\pi_i = S(t_{(i)})/S(t_{(i-1)})$ for the conditional probability of surviving from $S(t_{(i-1)})$ to $S(t_{(i)})$. Then we can write

$$S(t_{(i)}) = \pi_1 \pi_2 \ldots \pi_i,$$

and the likelihood becomes

$$L = \prod_{i=1}^{m} (1 - \pi_i)^{d_i} \pi_i^{c_i} (\pi_1 \pi_2 \ldots \pi_{i-1})^{d_i + c_i}.$$

Note that all cases who die at $t_{(i)}$ or are censored between $t_{(i)}$ and $t_{(i+1)}$ contribute a term π_j to each of the previous times of death from $t_{(1)}$ to $t_{(i-1)}$. In addition, those who die at $t_{(i)}$ contribute $1 - \pi_i$, and the censored

cases contribute an additional π_i. Let $n_i = \sum_{j \geq i}(d_j + c_j)$ denote the total number exposed to risk at $t_{(i)}$. We can then collect terms on each π_i and write the likelihood as

$$L = \prod_{i=1}^{m} (1 - \pi_i)^{d_i} \pi_i^{n_i - d_i},$$

a binomial likelihood. The m.l.e. of π_i is then

$$\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}.$$

The K-M estimator follows from multiplying these conditional probabilities.

1.3 Greenwood's Formula

From the likelihood function obtained above it follows that the large sample variance of $\hat{\pi}_i$ conditional on the data n_i and d_i is given by the usual binomial formula

$$\text{var}(\hat{\pi}_i) = \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Perhaps less obviously, $\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = 0$ for $i \neq j$, so the covariances of the contributions from different times of death are all zero. You can verify this result by taking logs and then first and second derivatives of the log-likelihood function.

To obtain the large sample variance of $\hat{S}(t)$, the K-M estimate of the survival function, we need to apply the delta method twice. First we take logs, so that instead of the variance of a product we can find the variance of a sum, working with

$$K_i = \log \hat{S}(t_{(i)}) = \sum_{j=1}^{i} \log \hat{\pi}_j.$$

Now we need to find the variance of the log of $\hat{\pi}_i$. This will be our first application of the delta method. The large-sample variance of a function f of a random variable X is

$$\text{var}(f(X)) = (f'(X))^2 \text{var}(X),$$

so we just multiply the variance of X by the derivative of the transformation. In our case the function is the log and we obtain

$$\text{var}(\log \hat{\pi}_i) = (\frac{1}{\hat{\pi}_i})^2 \text{var}(\hat{\pi}_i) = \frac{1 - \pi_i}{n_i \pi_i}.$$

Because K_i is a sum and the covariances of the $\pi_i's$ (and hence of the $\log \pi_i's$) are zero, we find

$$\text{var}(\log \hat{S}(t_{(i)})) = \sum_{j=1}^{i} \frac{1-\pi_j}{n_j \pi_j} = \sum \frac{d_j}{n_j(n_j - d_j)}.$$

Now we have to use the delta method again, this time to get the variance of the survivor function from the variance of its log:

$$\text{var}(\hat{S}(t_{(i)})) = [\hat{S}(t_{(i)})]^2 \sum_{j=1}^{i} \frac{1-\hat{\pi}_j}{n_j \hat{\pi}_j}.$$

This result is known as *Greenwood's formula*. You may question the derivation because it conditions on the n_j which are random variables, but the result is in the spirit of likelihood theory, conditioning on all observed quantities, and has been justified rigourously.

Peterson (1977) has shown that the K-M estimator $\hat{S}(t)$ is consistent, and Breslow and Crowley (1974) show that $\sqrt{n}(\hat{S}(t) - S(t))$ converges in law to a Gaussian process with expectation 0 and a variance-covariance function that may be approximated using Greenwood's formula. For a modern treatment of the estimator from the point of view of counting processes see Andersen et al. (1993).

1.4 The Nelson-Aalen Estimator

Consider estimating the cumulative hazard $\Lambda(t)$. A simple approach is to start from an estimator of $S(t)$ and take minus the log. An alternative approach is to estimate the cumulative hazard directly using the Nelson-Aalen estimator:

$$\hat{\Lambda}(t_{(i)}) = \sum_{j=1}^{i} \frac{d_j}{n_j}.$$

Intuitively, this expression is estimating the hazard at each distinct time of death $t_{(j)}$ as the ratio of the number of deaths to the number exposed. The cumulative hazard up to time t is simply the sum of the hazards at all death times up to t, and has a nice interpretation as the expected number of deaths in $(0, t]$ per unit at risk. This estimator has a strong justification in terms of the theory of counting processes.

The variance of $\hat{\Lambda}(t_{(i)})$ can be approximated by $\text{var}(-\log \hat{S}(t_{(i)}))$, which we obtained on our way to Greenwood's formula. Therneau and Grambsch (2000) discuss alternative approximations.

Breslow (1972) suggested estimating the survival function as

$$\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\},$$

where $\hat{\Lambda}(t)$ is the Nelson-Aalen estimator of the integrated hazard. The Breslow estimator and the K-M estimator are asymptotically equivalent, and usually are quite close to each other, particularly when the number of deaths is small relative to the number exposed.

1.5 Expectation of Life

If $\hat{S}(t_{(m)}) = 0$ then one can estimate $\mu = E(T)$ as the integral of the K-M estimate:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt = \sum_{i=1}^{m}(t_{(i)} - t_{(i-1)})\hat{S}(t_{(i)}).$$

Can you figure out the variance of $\hat{\mu}$?

2 k-Samples: Mantel-Haenszel

Consider now the problem of comparing two or more survivor functions, for example urban versus rural, or treated versus control. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the distinct times of death observed in the *total* sample, obtained by combining all groups of interest. Let

$$\begin{aligned} d_{ij} &= \text{deaths at time } t_{(i)} \text{ in group } j, \text{ and} \\ n_{ij} &= \text{number at risk at time } t_{(i)} \text{ in group } j. \end{aligned}$$

We also let d_i and n_i denote the total number of deaths and subjects at risk at time $t_{(i)}$.

If the survival probabilities are the same in all groups, then the d_i deaths at time $t_{(i)}$ should be distributed among the k groups in proportion to the number at risk. Thus, conditional on d_i and n_{ij},

$$E(d_{ij}) = d_i\frac{n_{ij}}{n_i} = n_{ij}\frac{d_i}{n_i},$$

where the last term shows that we can also view this calculation as applying an overall failure rate d_i/n_i to the n_{ij} subjects in group j.

We now proceed beyond the mean to obtain the distribution of these counts. Imagine setting up a contingency table at each distinct failure time, with rows given by survival status and columns given by group membership. The entries in the table are d_{ij} or $n_{ij} - d_{ij}$, the row totals are d_i and n_i and the column totals are n_{ij}. The distribution of the counts conditional on both the row and column totals is *hypergeometric*. (We mentioned this distribution briefly in WWS509 when we considered contingency tables with both margins fixed.) The hypergeometric distribution has mean as given above, variance

$$\text{var}(d_{ij}) = \frac{d_i(n_i - d_i)}{n_i - 1}\frac{n_{ij}}{n_i}(1 - \frac{n_{ij}}{n_i}),$$

and covariance

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

Let $\vec{d_i}$ denote the vector of deaths at time $t_{(i)}$, with mean $E(\vec{d_i})$ and var-cov matrix $\text{var}(\vec{d_i})$. We sum these over all times to obtain

$$D = \sum_{i=1}^{m}[\vec{d_i} - E(\vec{d_i})] \qquad \text{and} \qquad V = \sum_{i=1}^{m}\text{var}(\vec{d_i}).$$

Mantel and Haenszel proposed testing the equality of the k survival functions

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t)$$

by treating the quadratic form

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

as a χ^2 statistic with $k - 1$ degrees of freedom. Here V^{-} is any generalized inverse of V. Omitting the i-th group from the calculation of D and V will do; the test is invariant to the choice of omitted group. For $k = 2$ we get

$$z = \sqrt{Q} = \frac{\sum(d_{i1} - E(d_{i1}))}{\sqrt{\sum \text{var}(d_{i1})}}.$$

An approximation for $k \geq 2$ which does not require matrix inversion treats

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} denotes observed and E_{ij} expected deaths at time $t_{(i)}$ in group j, as a χ^2 statistic with $k - 1$ d.f.

The Mantel-Haenszel test can be derived as a linear rank test, and in that context is often called the *log-rank* or Savage test. Kalbfleisch and Prentice have proposed an extension to censored data of the Wilcoxon test. Other alternatives have been proposed by Gehan (1965) and Breslow (1970), but the M-H test is the most popular one.

3 Regression: Cox's Model

Let us consider the more general problem where we have a vector x of covariates. The k-sample problem can be viewed as the special case where the x's are dummy variables denoting group membership. Recall the basic model

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

and consider estimation of β without making any assumptions about the baseline hazard $\lambda_0(t)$.

3.1 Cox's Partial Likelihood

In his 1972 paper Cox proposed fitting the model by maximizing a special likelihood. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the observed distinct times of death, as before, and consider what happens at $t_{(i)}$. Let R_i denote the risk set at $t_{(i)}$, defined as the set of indices of the subjects that are alive just before $t_{(i)}$. Thus, $R_0 = \{1, 2, \ldots, n\}$.

Suppose first that there are no ties in the observation times, so one and only person subject failed at $t_{(i)}$. Let's call this subject $j(i)$. What is the conditional probability that this particular subject would fail at $t_{(i)}$ given the risk set R_i and the fact that exactly one subject fails at that time? Answer:

$$\frac{\lambda(t_{(i)}, x_{j(i)})dt}{\sum_{j\in R_i}\lambda(t_{(i)}, x_j)dt}.$$

We can write this probability in terms of the baseline hazard and relative risk as

$$\frac{\lambda_0(t_{(i)})e^{x'_{j(i)}\beta}}{\sum_{j\in R_i}\lambda_0(t_{(i)})e^{x'_j\beta}},$$

and we notice that the baseline hazard cancels, so the probability in question is

$$\frac{e^{x'_{j(i)}\beta}}{\sum_{j\in R_i} e^{x'_j\beta}}$$

and does not depend on the baseline hazard $\lambda_0(t)$.

Cox proposed multiplying these probabilities together over all distinct failure times and treating the resulting product

$$L = \prod_{i=1}^{m} \frac{e^{x'_{j(i)}\beta}}{\sum_{j\in R_i} e^{x'_j\beta}}$$

as if it was an ordinary likelihood. In his original paper Cox called this a "conditional likelihood" because it is a product of conditional probabilities, but later abandoned the name because it is misleading: L is not itself a conditional probability.

Kalbfleisch and Prentice considered the case where the covariates are fixed over time and showed that L is the *marginal likelihood* of the ranks of the observations, obtained by considering just the order in which people die and not the actual times at which they die.

In 1975 Cox provided a more general justification of L as part of the full likelihood—in fact, a part that happens to contain most of the information about β—and therefore proposed calling L a *partial likelihood*. This justification is valid even with time-varying covariates.

A more rigorous justification of the partial likelihood in terms of the theory of counting processes can be found in Andersen et al. (1993).

3.2 The Score and Information

The log of Cox's partial likelihood is

$$logL = \sum_i \{x_{(j(i)}\beta - \log \sum_{j\in R_i} e^{x'_j\beta}\}.$$

Taking derivatives with respect to β we find the score to be

$$\frac{\partial \log L_i}{\partial \beta_r} = x_{j(i)r} - \frac{\sum_{j\in R_i} e^{x'_j\beta} x_{jr}}{\sum_{j\in R_i} e^{x'_j\beta}}.$$

The term to the right of the minus sign is just a weighted average of x_r over the risk set R_i with weights equal to the relative risks $e^{x'_j\beta}$. Thus, we can

write the score as

$$U_r(\beta) = \frac{\partial \log L_i}{\partial \beta_r} = x_{j(i)r} - A_{ir}(\beta),$$

where $A_{ir}(\beta)$ is the weighted average of x_r over the risk set R_i.

Taking derivatives again and changing sign we find the observed information to be

$$-\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \frac{\sum e^{x_j'\beta} x_{jr} x_{js} (\sum e^{x_j'\beta}) - (\sum e^{x_j'\beta} x_{jr})(\sum e^{x_j'\beta} x_{js})}{(\sum e^{x_j'\beta})^2},$$

where all sums are over the risk set R_i. The right hand side can be written as the difference of two terms. The first term can be interpreted as a weighted average of the cross-product of x_r and x_s. The second term is the product of the weighted averages $A_{ir}(\beta)$ and $A_{is}(\beta)$. Thus we can write

$$-\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \frac{\sum e^{x_j'\beta} x_{jr} x_{js}}{\sum e^{x_j'\beta}} - A_{jr}(\beta) A_{js}(\beta).$$

You may recognize this expression as the old "desk calculator" formula for a covariance, leading to the observed information

$$I(\beta) = -\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{m} C_{irs}(\beta),$$

where $C_{irs}(\beta)$ denotes the weighted covariance of x_r and x_s over the risk set R_i with weights equal to the relative risks $e^{x_j'\beta}$ for $j \in R_i$. Calculation of the score and information is thus relatively simple. In matrix notation

$$u(\beta) = \sum_i (x_{j(i)} - A_i(\beta)) \qquad \text{and} \qquad I(\beta) = \sum_i C_i(\beta),$$

where $A_i(\beta)$ is the mean and $C_i(\beta)$ is the variance-covariance matrix of x over the risk set R_i with weights $e^{x_j'\beta}$ for $j \in R_i$.

Notably, the partial log-likelihood is formally identical with the log-likelihood for a conditional logit model.

3.3 The Problem of Ties

The development so far has assumed that only one death occurs at each distinct time $t_{(i)}$. In practice we often observe several deaths, say d_i, at $t_{(i)}$. This can happen in several ways:

- The data are *discrete*, so in fact there is a positive probability of failure at $t_{(i)}$. If this is the case one should really use a discrete model. We will discuss possible approaches below.

- The data are *continuous* but have been *grouped*, so d_i represents the number of deaths in some interval around $t_{(i)}$. In this case one would probably be better off estimating a separate parameter for each interval, using a complementary-log-log binomial model or a Poisson model corresponding to a piece-wise exponential survival model, as discussed in my WWS509 notes.

- The data are *continuous* and are *not grouped*, but there are a few ties resulting perhaps from coarse measurement. In this case we can extend the argument used to build the likelihood.

Let D_i denote the set of indices of the d_i cases who failed at $t_{(i)}$. The probability that the d_i cases that actually fail would be those in D_i given the risk set R_i and the fact that d_i of them fail at $t_{(i)}$ is

$$L = \frac{\prod_{j\in D_i} e^{x_j'\beta}}{\sum_{P_i}\prod_{j\in P_i} e^{x_j'\beta}},$$

where the sum in the denominator is over all possible permutations P_i or ways of choosing d_i indices from the risk set, and the product is over the set of chosen indices, which we call a permutation. For example assume four people are at risk and two die. The deaths could be {1,2}, {1,3}, {1,4}, {2,3}, {2,4}, {3,4}. We calculate the probability of each of these outcomes. Then we divide the probability of the outcome that actually occurred by the sum of the probabilities of all possible outcomes.

This likelihood was proposed by Cox in his original paper. The numerator is easy to calculate, and has the form $\exp\{S_i'\beta\}$, where $S_i = \sum_{j\in D_i} x_j$ is the sum of the x's over the death set D_i. The denominator is difficult to calculate because the number of permutations grows very quickly with d_i.

Peto (and independently Breslow) proposed approximating the denominator by calculating the sum $\sum e^{x_j'\beta}$ over the entire risk set R_i and raising it to d_i. This leads to the much simpler expression

$$L \approx \prod_{i=1}^{m} e^{S_i'\beta}\left(\sum_{j\in R_i} e^{x_j'\beta}\right)^{d_i}.$$

The Peto-Breslow approximation is reasonably good when d_i is small relative to n_i, and is popular because of its simplicity. Efron proposed a better

approximation that requires only modest additional computational effort. Consider again our example where two out of four subjects fail. Suppose the subjects that fail are 1 and 2, and let $r_j = e^{x_j'\beta}$ denote the relative risk for the j-th subject. In continuous time one must have failed before the other, we just don't know which. The contributions to the partial likelihood would be

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

if 1 failed before 2, or

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

if 2 was the first to fail. In both cases the numerator is $r_1 r_2$. To compute the denominator Peto and Breslow add the risks over the complete risk set both times, using $(r_1+r_2+r_3+r_4)^2$, which is obviously conservative. Efron uses the average risk of subjects 1 and 2 for the second term, so he takes the denominator to be $(r_1+r_2+r_3+r_4)(0.5r_1+0.5r_2+r_3+r_4)$. This approximation is much more accurate unless d_i is very large relative to n_i. For more details see Therneau and Grambsch (2000, Section 3.3).

3.4 Tests of Hypotheses

As usual, we have three approaches to testing hypotheses about $\hat{\beta}$:

- *Likelihood Ratio Test:* given two nested models, we treat twice the difference in partial log-likelihoods as a χ^2 statistic with degrees of freedom equal to the difference in the number of parameters.

- *Wald Test:* we use the fact that approximately in large samples $\hat{\beta}$ has a multivariate normal distribution with mean β and variance-covariance matrix $\text{var}(\hat{\beta}) = I^{-1}(\beta)$. Thus, under $H_0 : \beta = \beta_0$, the quadratic form

  $$(\hat{\beta}-\beta_0)'\text{var}^{-1}(\hat{\beta})(\hat{\beta}-\beta_0) \sim \chi^2_p,$$

  where p is the dimension of β. This test is often used for a subset of β.

- *Score Test:* we use the fact that approximately in large samples the score $u(\beta)$ has a multivariate normal distribution with mean 0 and variance-covariance matrix equal to the information matrix. Thus, under $H_0 : \beta = \beta_0$, the quadratic form

  $$u(\beta_0)'I^{-1}(\beta_0)u(\beta_0) \sim \chi^2_p.$$

  Note that this test does not require calculating the m.l.e. $\hat{\beta}$.

One reason for bringing up the score test is that in the k-sample case the score test of $H_0 : \beta = 0$ based on Cox's model happens to be the same as the Mantel-Haenszel log-rank test.

Here is an outline of the proof. Assume no ties. If $\beta = 0$ then the weights used to calculate A_i and C_i are all 1, A_{ir} happens to be the proportion of the risk set R_i that comes from the r-th sample (which is the same as the expected number of deaths in that group) and C_i is a binomial variance-covariance matrix. If there are ties, Cox's approach leads to the test discussed in Section 2. Use of Peto's approximation is equivalent to omitting the factor $(n_i - d_i)/(n_i - 1)$ from the variance-covariance matrix.

All three tests are asymptotically equivalent. The quality of the normal approximations depends on the sample size, the distribution of the cases over the covariate space, and the extent of censoring.

3.5 Time-Varying Covariates

A nice feature of the Cox model and partial likelihood is that it extends easily to the case of time-varying covariates. Note that the partial likelihood is built by considering only what happens at each failure time, so we only need to know the values of the covariates at the distinct times of death.

One use of time-varying covariates is to check the assumption of proportionality of hazards. In his original paper Cox analyzes a two-sample problem and introduces an auxiliary covariate

$$z_i = \begin{cases} 0, \text{in group 0} \\ t - c, \text{in group 1} \end{cases}$$

where c is an arbitrary constant close to the mean of t, chosen to avoid numerical instability. If the coefficient of z is 0, the assumption of proportionality of hazards is adequate. A positive value indicates that the ratio of hazards for group 1 over group 0 increases over time. A negative coefficient suggests a declining hazard ratio, a common occurrence.

Another use of time-varying covariates is to represent variables that simply change over time. In a study of contraceptive failure, for example, one may treat frequency of intercourse as a time-varying covariate. Another example of a time-varying covariate is education in an analysis of age at marriage.

Note that $x(t)$ may represent the actual value of a variable at time t, or any index based on the individual's history up to time t. In a study of the effects of breastfeeding on post-partum amenorrhea, for example, $x(t)$ could represent the total number of suckling episodes in the week preceding t.

The partial likelihood function with time-varying covariates does not have an interpretation as a marginal likelihood of the ranks. For further details see Cox and Oakes (1984, Chapter 8).

3.6 Estimating the Baseline Survival

Interest so far has focused on the regression coefficients β. We now consider how to estimate the baseline hazard $\lambda_0(t)$, which dropped out of the partial likelihood.

Kalbfleisch and Prentice (1980, Section 4.3) use an argument similar to the derivation of the Kaplan-Meier estimate, noting that the hazard should assign mass only to the discrete times of death. Let π_i denote the conditional survival probability at time $t_{(i)}$ for a baseline subject. To obtain the conditional probability for a subject with covariates x we would need to raise π_i to $e^{x'\beta}$. This leads to a likelihood of the form

$$L = \prod_{i=1}^{m} \prod_{j \in D_i} (1-\pi_i)^{e^{x_j'\beta}} \prod_{j \in R_i - D_i} \pi_i^{e^{x_j'\beta}}.$$

Meier suggested maximizing this likelihood with respect to both the π_i and β. A simpler approach is to plug-in the estimate $\hat{\beta}$ from the partial likelihood, and maximize the resulting expression with respect to the π_i only. If there are no ties, this gives

$$\hat{\pi}_i = \left(1 - \frac{e^{x_{j(i)}'\hat{\beta}}}{\sum_{j \in R_i} e^{x_j'\hat{\beta}}}\right)^{e - x_{j(i)}'\hat{\beta}}.$$

Think of this as follows. With no covariates, our estimate would be $\hat{\pi}_i = 1 - d_i/n_i$ with $d_i = 1$, which is the K-M estimate. With covariates we do the same thing, except that we weight each case by its relative risk. The resulting survival probability is then raised to $e^{-x_{j(i)}'\beta}$ to turn it into a *baseline* probability.

If there are ties one has to solve

$$\sum_{j \in D_i} \frac{e^{x_j'\hat{\beta}}}{1 - \pi_i^{e^{x_j'\hat{\beta}}}} = \sum_{j \in R_i} e^{x_j'\hat{\beta}}$$

iteratively. A suitable starting value is

$$\log \pi_i = -\frac{d_i}{\sum_{j \in R_i} e^{-x_j'\hat{\beta}}}.$$

The estimate of the baseline survival function is then a step function

$$\hat{S}_0(t) = \prod_{i:t_{(i)}<t} \hat{\pi}_i.$$

Cox and Oakes (1984, Section 7.8) describe a simpler estimator that extends the Nelson-Aalen estimate of the cumulative hazard to the case of covariates. The estimator can be described somewhat heuristically as follows. Treat the baseline hazard as 0 except at failure times. The expected number of deaths at $t_{(i)}$ can be obtained by summing the hazards over the risk set:

$$E(d_i) = \sum_{j \in R_i} \lambda_0(t_{(i)}) e^{x_j'\beta}.$$

Equating the observed and expected number of deaths at $t_{(i)}$ leads us to estimate $\lambda_i = \lambda_0(t_{(i)})$ as

$$\hat{\lambda}_i = \frac{d_i}{\sum e^{x_j'\beta}},$$

where the sum is over the risk set R_i. The cumulative hazard and survival functions are then estimated as

$$\hat{\Lambda}_0(t) = \sum_{i:t_{(i)}<t} \hat{\lambda}_i, \qquad \text{and} \qquad \hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}.$$

If there are no covariates these reduce to the ordinary Nelson-Aalen and Breslow estimators described earlier.

Having obtained estimates of the baseline hazard and survival, we can obtain fitted hazards and survival functions for any value of x. This task is pretty straightforward when we have time-fixed covariates, as all we need to do is multiply the baseline hazard by the relative risk, or raise the baseline survival to the relative risk. With time-varying covariates things get somewhat more complicated, as we have to pick up the appropriate hazard for each distinct failure time depending on the values of the covariates at that point.

3.7 Martingale Residuals

Residuals play an important role in checking linear and generalized linear models. Not surprisingly, the concept has been extended to survival models. A lot of this work relies heavily on the terminology and notation of counting processes. We will try to convey the essential ideas in a non-technical way.

Instead of focusing on the i-th individual's survival time T_i, we will introduce a function $N_i(t)$ that counts events over time. In survival models $N_i(t)$ will be zero while the subject is alive and then it will become one. (In more general event-history models $N_i(t)$ counts the number of occurrences of the event to subject i by time t.) While survival time T_i is a random *variable*, $N_i(t)$ is a stochastic *process*, a function of time.

We will also introduce a function to track the i-th individual's exposure. $Y_i(t)$ will be one while the individual is in the risk set and zero afterwards. Note that $Y_i(t)$ can become zero due to death or due to censoring.

To complete the model we add a hazard function $\lambda_i(t)$ representing the i-th individuals risk at time t. In a Cox model $\lambda_i(t) = \lambda_0(t)\exp\{x_i'\beta\}$.

In the terminology of counting processes, the process $N_i(t)$ is said to have *intensity* $Y_i(t)\lambda_i(t)$. The intensity is just the risk if the subject is exposed, and is zero otherwise. The probability that $N_i(t)$ will jump in a small interval $[t, t+dt)$ conditional on the entire history of the process is given by $\lambda_i(t)Y_i(t)dt$, and is proportional to the intensity of the process and the width of the interval.

A key feature of this formulation is that

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(t)Y_i(t)dt$$

is a *martingale*, a fact that can be used to establish the properties of tests and estimators using martingale central limit theory. A martingale is essentially a stochastic process without drift. Given two times $0 < t_1 < t_2$ the expectation $E(M(t_2))$ given the history up to time t_1 is simply $M(t_1)$. In other words martingale increments have mean zero. Also, martingale increments are uncorrelated, although not necessarily independent.

The integral following the minus sign in the above equation is called the *compensator* of $N_i(t)$. You may think of it as the conditional expected value of the counting process at time t. Subtracting the compensator turns the counting process into a martingale. This equation suggest immediately the first type of residual one can use in Cox models, the so-called *Martingale Residual*:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(t)e^{x_i'\hat{\beta}}d\hat{\Lambda}_0(t)$$

where $\hat{\Lambda}_0(t)$ denotes the Nelson-Aalen estimator of the baseline hazard. Because this is a discrete function with jumps at the observed failure times, the integral in the above equation should be interpreted as a sum over all j such that $t_{(j)} < t$. Usually the residual is computed at $t = \infty$ (or the largest

observed survival time), in which case the martingale residual is

$$\hat{M}_i = d_i - e^{x_i'\hat{\beta}}\hat{\Lambda}_0(t_i),$$

where d_i is the usual death indicator and t_i is the observation time (to death or censoring) for the i-th individual.

One possible use of residuals is to find outliers, in this case individuals who lived unusually short or long times, after taking into account the relative risks associated with their observed characteristics.

Martingale residuals are just one of several types of residuals that have been proposed for survival models. Others include deviance, score and Schoenfeld residuals. For more details on counting processes, martingales and residuals see Therneau and Grambsch (2000), especially Section 2.2 and Chapter 4.

3.8 Models for Discrete and Grouped Data

We close with a brief review of alternative approaches for discrete and continuous grouped data, expanding slightly on the WWS509 discussion.

Cox (1972) proposed an alternative version of the proportional hazards model for *discrete* data. In this case the hazard is the conditional probability of dying at time t given survival up to that point, so that

$$\lambda(t) = \Pr\{T = t | T \geq t\}.$$

Cox's discrete logistic model assumes that the *conditional* odds of surviving $t_{(i)}$ are proportional to some baseline odds, so that

$$\frac{\lambda(t,x)}{1-\lambda(t,x)} = \frac{\lambda_0(t)}{1-\lambda_0(t)} e^{x'\beta}.$$

Note that taking logs on both sides results in a logit model, where the logit of the discrete hazard is linear on β.

Do not confuse this model with the proportional odds model, where the *unconditional* odds of survival (or odds of dying) are proportional for different values of x.

If the $\lambda_0(t)$ are small so that $1 - \lambda_0(t)$ is close to one, this model will be similar to the proportional hazards model, as the odds are close to the hazard itself.

A nice property of this model is that the partial likelihood turns out to be identical to that of the continuous-time model. To see this point note

that under this model the hazard at time t_i for an individual with covariate values x is

$$\lambda(t_i, x) = \frac{\theta_i e^{x'\beta}}{1 + \theta_i e^{x'\beta}},$$

where $\theta_i = \lambda_0(t_i)/(1 - \lambda_0(t_i))$ denotes the baseline conditional odds of surviving t_i. Suppose there are two cases exposed at time t_i, labelled 1 and 2. The probability that 1 dies and 2 does not is

$$\lambda(t_i, x_1)(1 - \lambda(t_i, x_2)) = \frac{\theta_i e^{x_1'\beta}}{1 + \theta_i e^{x_1'\beta}} \frac{1}{1 + \theta_i e^{x_2'\beta}}.$$

The probability that 2 dies and 1 does not has a similar structure, with $\theta_i e^{x_2'\beta}$ in the numerator and the same denominator. When we divide the probability of one of these outcomes by the sum of the two, the denominators cancel out, as do the θ_i. Thus, the conditional probability is

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

which is exactly the same as in the continuous case. (You may wonder why we did not consider the $1-\lambda$'s in the continuous case. It turns out we didn't have to. In you repeat the continuous-time derivation including terms of the form λdt for deaths and $1 - \lambda dt$ for survivors you will discover that the dt's for deaths cancel out but those for survivors do not, and as $dt \to 0$ the terms $1 - \lambda dt \to 1$ and drop from the likelihood.)

There is an alternative approach to discrete data that is particularly appropriate when you have *grouped continuous times*. See Kalbfleisch and Prentice (1980), Section 2.4.2 and Section 4.6.1. Suppose we wanted a discrete time model that preserves the relationship between survivor functions in the continuous time model, namely

$$S(t, x) = S_0(t)^{e^{x'\beta}}.$$

In the discrete case we have $S(t, x) = \prod_{u<t}(1 - \lambda(u, x))$, so we must have

$$1 - \lambda(t_i, x) = (1 - \lambda_0(t_i))^{e^{x'\beta}}.$$

Solving for the hazard we get

$$\lambda(t_i, x) = 1 - (1 - \lambda_0(t_i))^{e^{x'\beta}}.$$

The linearizing transformation for this model is the complementary log-log link, $\log(-\log(1-\lambda))$.

To see why this model is uniquely appropriate for grouped data suppose we only observe failures in intervals

$$[0=\tau_0,\tau_1),[\tau_1,\tau_2),\ldots,[\tau_{k-1},\tau_k=\infty).$$

Suppose the hazard is continuous and satisfies a standard proportional hazards model. The probability of surviving interval i for a subject with covariates x is

$$\Pr\{T>\tau_i|T>\tau_{i-1},x\}=\frac{S(\tau_i,x)}{S(\tau_{i-1},x)}.$$

Writing this in terms of the baseline survival we get

$$\left(\frac{S_0(\tau_i)}{S_0(\tau_{i-1})}\right)^{e^{x'\beta}}=\left(e^{-\int_{\tau_{i-1}}^{\tau_i}\lambda_0(t)dt}\right)^{e^{x'\beta}}.$$

In view of this result, we define the baseline hazard in interval (τ_{i-1},τ_i) as

$$\lambda_{0i}=1-e^{-\int_{\tau_{i-1}}^{\tau_i}\lambda_0(t)dt}.$$

The hazard for an individual with covariates x in the same interval then becomes

$$\lambda_i(x)=1-(1-\lambda_{0i})^{e^{x'\beta}},$$

and can be linearized using the c-log-log transformation.

This is the only discrete model appropriate for grouped data from a continuous-time proportional hazards model. Unfortunately, it cannot be estimated non-parametrically using a partial likelihood argument. (If you try to construct a partial likelihood you will discover that the λ_{0i}'s do not drop out of the likelihood.)

In practice, both the logit and the complementary log-log discrete models can be estimated via ordinary likelihood techniques by treating the baseline hazard at each discrete failure time (or interval) as a separate parameter to be estimated, provided of course one has enough failures at each time (or interval). The resulting likelihood is the same as that of a binomial model with logit or c-log-log link, so either model can be easily fitted with standard software.

One difficulty with discrete models generated by grouping continuous data has to do with censored cases that may have been exposed part of the interval. The only way to handle these cases is to censor them at the

beginning of the interval, which throws away some information. This is not a problem with piece-wise exponential models based on the Poisson equivalence, because one can easily take into account partial exposure in the offset.