Lecture 28—Wednesday, March 21, 2012

Topics

Comparing survivor functions across groups

The hypergeometric model is the finite-population, sampling without replacement version of the binomial model. Whereas in the binomial model the probability of a success is constant from trial to trial, in the hypergeometric model the probability changes as selections are made. The standard illustration of a hypergeometric model is choosing colored balls from an urn. Suppose there are N balls in an urn of which m of them are red and the rest are black. The probability of drawing $X = k$ red balls in a sample of size n balls is the following.

$$P(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

The first term in the numerator is the number of ways of choosing k red balls, the second term is the number of ways of choosing $n - k$ black balls, and the denominator is the number of different samples of n balls. The mean and variance of the hypergeometric distribution are the following.

$$\mu = \frac{mn}{N},\ \sigma^2 = \frac{mn}{N}\left(\frac{N-m}{N}\right)\frac{N-n}{N-1}$$

We can apply this to survival analysis as follows. For simplicity suppose there are just two groups. At failure time j there are a total of O_j failures in a risk set of N_j individuals of which n_{1j} of the individuals are from group 1. The probability of observing k failures in group 1 is the hypergeometric distribution (assuming the same failure rate in both groups).

$$P(X=k) = \frac{\binom{O_j}{k}\binom{N_j - O_j}{n_{1j}-k}}{\binom{N_j}{n_{1j}}}$$

The log-rank test resembles the Pearson X^2 test in its construction. At each failure time we obtain the expected number of deaths under a hypergeometric model for the current risk set assuming that there is a common failure rate in all the groups. We then subtract the expected number of deaths from the observed number of deaths in group 1. This is repeated at each failure time and the differences are summed. Treating the outcomes at each failure time as independent, the variance of the sum is just the sum of the hypergeometric variances at each event time. We divide the the summed differences by the square root of the variance of this sum under the hypergeometric model to obtain the test statistic of the log-rank test. When squared this test statistic has a chi-squared distribution.

In variations of this test the failure times are weighted differently in order to emphasize early events or late events. As an illustration when there are just two groups the log-rank test takes the following form.

$$Z^2 = \frac{\left[\sum_j w(t_1)\left(m_{1j} - e_{1j}\right)\right]^2}{\mathrm{Var}\left(\sum_j w(t_1)\left(m_{1j} - e_{1j}\right)\right)}$$

Here m_{ij} and e_{ij} are the observed and expected deaths for group i at failure time j. One version of the test (variously called the generalized Wilcoxon or Gehan-Wilcoxon or just the Gehan test) uses the number at risk in the group as the weight, a choice that tends to reward early survival over later survival. The log-rank test proper uses equally weighted observations which weights later survival more than the Gehan test does. The log-rank test is purely a test of significance and cannot be used to estimate the magnitude of the differences between groups.

Regression models in survival analysis

One of the goals in the regression modeling of survival data is to obtain estimates of the survivor function after adjusting for measured covariates, something that is not possible with the Kaplan-Meier estimator. There are two standard approaches to regression analysis of survival data.

1. A proportional hazards model (PH) assumes that covariates are multiplicative with respect to the hazard. Typically we model the log hazard as a linear function of covariates.
2. An accelerated failure time model (AFT) assumes covariates are multiplicative with respect to survival time. Typically we model the log survival time (or the scale parameter of a particular probability distribution) as a linear function of covariates.

In addition we have the choice of a semi-parametric or a parametric model. Parametric models can be either proportional hazards or accelerated failure time models. The most popular semiparametric model is a proportional hazards model called Cox regression.

Cox proportional hazards regression model

The Cox proportional hazards regression model is also called the Cox model, Cox regression, or just proportional hazards regression (although the latter is really a misnomer). It is extremely popular in medical fields where it is often the only kind of model that is considered. It is considered a semi-parametric model for reasons explained below. In Cox regression we model the hazard function as follows.

$$h(t, x) = h_0(t)\exp\left(\sum_{i=1}^{p}\beta_i x_i\right)$$

This is a model that is linear in the log hazard.

$$\log h(t, x) = \log h_0(t) + \sum_{i=1}^{p}\beta_i x_i$$

The hazard in Cox regression is the product of two terms.

1. $h_0(t)$ is called the baseline hazard. Notice that the baseline hazard involves t, but not x.
2. $\exp\left(\sum_{i=1}^{p}\beta_i x_i\right)$ contains the linear predictor and multiplies the baseline hazard. Notice that this term does not involve t. The assumption being made is that the individual predictors, x_i, are time-invariant.

There is a model called the extended Cox model that does allow covariates to be time-dependent.

Why is the Cox model so popular?

1. Because we model the log hazard as a linear predictor we are guaranteed that estimate of the hazard will be non-negative as it should be.
2. It is not necessary to actually specify the hazard function completely. The baseline hazard is not estimated in the Cox model because it drops out of the likelihood (see the next section) and is actually not needed when making comparisons of interest. This is why the Cox model is called semi-parametric.
3. The Cox model generally agrees with the correct parametric model when the survival times do follow a specific parametric form. Thus the Cox model is robust to model misspecification.

The Cox model is called semi-parametric because the hazard function is not fully specified.

- The Cox model contains the term $h_0(t)$ that represents the hazard when all $x_i = 0$, but otherwise it is unspecified.
- In a parametric model the hazard is fully specified. For instance, if we assume a Weibull model for the survival times, $T \sim$ Weibull(λ, p), then the baseline hazard is given by $h_0(t) = \lambda p t^{p-1}$.

Estimating the Cox model

The regression coefficients of the Cox model are estimated by maximizing a quantity known as the partial likelihood (rather than a full likelihood). Recall that the likelihood is just the probability of obtaining the data that were obtained. In a partial likelihood for survival data rather than specifying P(data), we instead construct an expression for P(those who fail). Individuals who were censored do not contribute individual terms to the partial likelihood. Thus the likelihood takes the form

$$\text{Likelihood} = L_1 \times L_2 \times \cdots \times L_k$$

in which there are k failure times. At each failure time the censored individuals do contribute to the risk set and are used in calculating the individual terms of the likelihood.

Formally the Cox partial likelihood is constructed as follows. Let $t_1, t_2, \ldots, t_n$ be the observation times for the n observations in the study and let $\delta_1, \delta_2, \ldots, \delta_n$ be indicators of the event at those times, i.e.,

$$\delta_i = \begin{cases} 1 & \text{if } t_i \text{ is a failure time} \\ 0 & \text{if } t_i \text{ is a censor time} \end{cases}$$

Using the Cox model for the hazard, the hazard for individual i is just

$$h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)$$

Now form the following ratio.

$$\frac{h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{\sum_{k\in R(t_i)} h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_{kj}\right)} = \frac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{\sum_{k\in R(t_i)} \exp\left(\sum_{j=1}^{p}\beta_j x_{kj}\right)}$$

In the denominator we are summing the hazards for all individual still alive at time t_i, i.e., members of the risk set $R(t_i)$. Notice that the baseline hazard $h_0(t)$ cancels and does not appear in the final expression. The Cox partial likelihood is the product of all such terms.

$$L = \prod_{i=1}^{n}\left[\frac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{\sum_{k\in R(t_i)} \exp\left(\sum_{j=1}^{p}\beta_j x_{kj}\right)}\right]^{\delta_i}$$

The use of δ_i as an exponent is just a convenient way of including all observations in the likelihood without having to single out the failure times. Observations that are censored have $\delta_i = 0$ and hence contribute nothing to the likelihood (their contribution to the product is one).

Some comments about the partial likelihood

1. This is not an explicit probability model in the conventional sense. The terms are the ratios of hazards and so behave like probabilities but are not really probabilities.
2. Notice that the event times $t_1, t_2, \ldots, t_n$, whether failures or censors, don't explicitly appear in the likelihood. They only arise in the calculation of the risk set. As a result, their actual values don't matter, only their relative order. The actual time of death is irrelevant. Thus the Cox model has a lot in common with conventional nonparametric approaches to statistics where ranks are analyzed rather than actual values.
3. The Cox model provides us with information about which covariates significantly affect survival, but it provides us with no estimate of the hazard or survivor function directly. After all how can it? The baseline hazard term does not appear in the likelihood. In truth there do exist

some ad hoc methods to estimate $h_0(t)$, and hence $h(t)$ and $S(t)$. What we get is a Kaplan-Meier-like estimate of the survival curve adjusted for covariates.

Hazard ratios

In logistic regression the focus is on odds ratios. A similar quantity, the hazard ratio, plays a role in Cox regression. To construct the hazard ratio we just take the ratio of the hazards of two individuals who have different values of the covariates, x.

$$\mathrm{HR} = \frac{h(t, x^*)}{h(t, x)} = \frac{h_0(t)\exp\left(\sum_{i=1}^{p} \beta_i x_i^*\right)}{h_0(t)\exp\left(\sum_{i=1}^{p} \beta_i x_i\right)} = \exp\left[\sum_{i=1}^{p} \beta_i \left(x_i^* - x_i\right)\right]$$

Now suppose

$$x_1 = \begin{cases} 1 & \text{if treatment} \\ 0 & \text{if control} \end{cases}$$

but for all other values of x, $x = x^*$. Then we have $\mathrm{HR} = \exp(\beta_1)$ for two individuals that differ only in their treatment. The hazard ratio in this instance tells us by what amount the hazard is multiplied for individuals in the treatment group relative to the control group while holding everything else constant.

The proportional hazards assumption

Notice that because the baseline hazards cancel, the hazard ratio is constant with respect to time. This is the essence of the proportional hazards assumption. We'll discuss how one might go about testing this assumption in lecture 29, but what should one do if the assumption appears to be violated?

1. Stratify by the levels of a categorical variable for which the proportionality assumption fails. In this approach a separate baseline hazard is assumed for members of each stratum but all the data are still used to obtain parameter estimates.
2. Fit separate Cox models to different time intervals. This makes sense because if the proportional hazards assumption is violated then the hazard is not constant with time.
3. Use the extended Cox model instead of the ordinary Cox model. The extended Cox model permits time-dependent covariates.

Parametric survival models

In parametric survival models an explicit probability model is chosen for the survival time distribution / hazard function. By choosing a probability model one also automatically chooses either a proportional hazards or an accelerated failure time model. Except for the Weibull (exponential) distributions, only one of these choices is possible with a given probability model. The disadvantages of the parametric approach are the following.

1. It requires stronger assumptions than the Cox regression model.
2. It can be more complicated to understand for the novice.

The advantages of the parametric approach are the following.

1. In addition to determining which variables affect survival, one also obtains smooth estimates of the survivor and hazard functions.
2. The parametric approach can easily handle any kind of censoring.

Censoring in parametric survival models

Let $f(t)$ denote the probability density for the survival distribution. Table 4 summarizes how different kinds of censored observations contribute to the parametric likelihood of failure times. Note: $F(0) = 0$.

Table 4 Contributions of censored observations to a parametric likelihood

Type	Event	Contribution to the likelihood
uncensored	$T = 2$	$f(2)$
right censored	$T > 2$	$\int_2^\infty f(t)dt = 1 - F(2)$
left censored	$T \leq 2$	$\int_0^2 f(t)dt = F(2)$
interval censored	$2 < T \leq 3$	$\int_2^3 f(t)dt = F(3) - F(2)$

Weibull regression model

While there are many potential probability models for survival distributions, the Weibull is the most commonly used and perhaps the most flexible. The Weibull survivor and hazard functions are shown below.

$$\text{survivor function: } S(t) = \exp\left(-\lambda t^{p}\right)$$

$$\text{hazard function: } h(t) = \lambda p t^{p-1}$$

Here p = shape parameter and λ (typically its log) is modeled in terms of explanatory variables. The exponential distribution is a special case of the Weibull ($p = 1$).

The Weibull distribution yields both a proportional hazards model and an accelerated failure time model depending on how things are parameterized. Having chosen one of the parameterizations it is possible to obtain the corresponding estimates for the other parameterization as Table 5 explains.

Table 5 Parameterizations for the Weibull regression model

Proportional hazards	**Accelerated failure time**
$\lambda = \exp\left(\beta_0 + \sum_i \beta_i x_i\right)$	$\lambda^{1/p} = \exp\left[-\left(\alpha_0 + \sum_i \alpha_i x_i\right)\right]$
$\log \lambda = \beta_0 + \sum_i \beta_i x_i$	$\log \lambda = -p\left(\alpha_0 + \sum_i \alpha_i x_i\right)$

From Table 5 we see that we can switch between the parameterizations using the identity $\beta_j = -\alpha_j p$. Thus when $\beta_j < 0$ in the proportional hazards parameterization (meaning the hazard is decreased by increasing the value of the predictor), it follows that $\alpha_j > 0$ in the accelerated failure time parameterization (meaning survival time is extended).

Discrete time survival analysis

An alternative approach to interval censored data is to use what's known as discrete time survival analysis. Discrete time survival analysis uses binary logistic regression with dummy variables to indicate the different survival intervals. See Singer & Willett (2003), chapters 10–12, or Kleinbaum & Klein (2005), pp. 290–292, for more details.

References

- Kleinbaum, David G. and Mitchel Klein. 2005. *Survival Analysis: A Self-learning Text.* Springer, New York.
- Singer, J. D. & Willett, J. B. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Oxford, UK.

Course Home Page

Jack Weiss
Phone: (919) 962-5930
E-Mail: jack_weiss@unc.edu
Address: Curriculum for the Environment and Ecology, Box 3275, University of North Carolina, Chapel Hill, 27599

Last Revised--March 24, 2012
URL: https://sakai.unc.edu/access/content/group/2842013b-58f5-4453-aa8d-3e01bacbfc3d/public/Ecol562_Spring2012/docs/lectures/lecture28.htm