Survival Analysis with R

Background

In the class on essential statistics we covered basic categorical data analysis – comparing proportions (risks, rates, etc) between different groups using a chi-square or fisher exact test, or logistic regression. For example, we looked at how the diabetes rate differed between males and females. In this kind of analysis you implicitly assume that the rates are constant over the period of the study, or as defined by the different groups you defined.

But, in longitudinal studies where you track samples or subjects from one time point (e.g., entry into a study, diagnosis, start of a treatment) until you observe some outcome *event* (e.g., death, onset of disease, relapse), it doesn't make sense to assume the rates are constant. For example: the risk of death after heart surgery is highest immediately post-op, decreases as the patient recovers, then rises slowly again as the patient ages. Or, recurrence rate of different cancers varies highly over time, and depends on tumor genetics, treatment, and other environmental factors.

Definitions

Survival analysis lets you analyze the rates of occurrence of events over time, without assuming the rates are constant. You model the *time until an event occurs*,[1] or compare the time-to-event between different groups, or how time-to-event correlates with quantitative variables.

Hazard: the instantaneous event (death) rate at a particular time point t. Survival analysis doesn't assume the hazard is constant over time. The *cumulative hazard* is the total hazard experienced up to time t.

Survival function: Probability an individual survives (or, the probability that the event of interest does not occur) up to and including time t. It's the probability that the event (e.g., death) hasn't occured yet. T is the time of death. $Pr(T > t)$ is the probability that the time of death is greater than some time t.

$$S(t) = Pr(T > t)$$

Kaplan-Meier cuve: illustrates the survival function. It's a step function illustrating the cumulative survival probability over time. The curve is horizontal over periods where no event occurs, then drops vertically corresponding to a change in the survival function at each time an event occurs.

Censoring: occurs when when you track the sample/subject through the end of the study and the event never occurs. This could happen due to the sample/subject dropping out of the study for reasons other than death, or some other loss to followup. The sample is *censored* in that you only know that the individual survived up to the loss to followup, but you don't know anything about survival after that.[2]

Proportional hazards assumption: The main goal of survival analysis is to compare the survival functions in different groups, e.g., leukemia patients as compared to cancer-free controls. If you followed both groups until everyone died, both survival curves would end at 0%, but one group might have survived on average a lot longer than the other group. Survival analysis does this by comparing the *hazard* at different times over the observation period. Survival analysis doesn't assume that the hazard is constant, but *does* assume that the *ratio* of hazards between groups is constant over time. This workshop does *not* cover methods to deal with non-proportional hazards, or interactions of covariates with the time to event.

Proportional hazards regression a.k.a. **Cox regression** is the most common approach to assess the effect of different variables on survival.

[1] In the medical world, we typically think of *survival analysis* literally – tracking time until death. But, it's more general than that – survival analysis models time until an *event* occurs (*any* event). This might be death of a biological organism. But it could also be the time until a hardware failure in a mechanical system, time until recovery, time someone remains unemployed after losing a job, time until a ripe tomato is eaten by a grazing deer, time until someone falls asleep in a workshop, etc. *Survival analysis* also goes by *reliability theory* in engineering, *duration analysis* in economics, and *event history analysis* in sociology.

[2] This describes the most common type of censoring – *right censoring*. *Left censoring* occurs when the "start" is unknown, such as when an initial diagnosis or exposure time is unknown.

Cox PH Model

Kaplan-Meier curves are good for visualizing differences in survival between two categorical groups,[3] but they don't work well for assessing the effect of *quantitative* variables like age, gene expression, leukocyte count, etc. Cox PH regression can assess the effect of both categorical and continuous variables, and can model the effect of multiple variables at once.[4]

Cox PH regression models the natural log of the hazard at time t, denoted $h(t)$, as a function of the baseline hazard ($h_0(t)$) (the hazard for an individual where all exposure variables are 0) and multiple exposure variables x_1, x_1, ..., x_p. The form of the Cox PH model is:

$$log(h(t)) = log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

If you exponentiate both sides of the equation, and limit the right hand side to just a single categorical exposure variable (x_1) with two groups ($x_1 = 1$ for exposed and $x_1 = 0$ for unexposed), the equation becomes:

$$h_1(t) = h_0(t) + e^{\beta_1 x_1}$$

Rearranging that equation lets you estimate the **hazard ratio**, comparing the exposed to the unexposed individuals at time t:

$$HR(t) = \frac{h_0(t)e^{\beta_1}}{h_0(t)} = e^{\beta_1}$$

This model shows that **the hazard ratio is e^{β_1},** and remains constant over time t (hence the name *proportional hazards regression*). The β values are the regression coefficients that are estimated from the model, and represent the *loghazardratio* for each unit increase in the corresponding predictor variable. The interpretation of the hazards ratio depends on the measurement scale of the predictor variable, but in simple terms, a positive coefficient indicates worse survival and a negative coefficient indicates better survival for the variable in question.

Survival analysis in R

The core survival analysis functions are in the **survival** package. The survival package is one of the few "core" packages that comes bundled with your basic R installation, so you probably didn't need to `install.packages()` it. But, you'll need to load it like any other library when you want to use it. The core functions we'll use out of the survival package include:

- `Surv()`: Creates a survival object.
- `survfit()`: Fits a survival curve using either a formula, of from a previously fitted Cox model.
- `coxph()`: Fits a Cox proportional hazards regression model.

`Surv()` creates the response variable, and typical usage takes the time to event,[^time2] and whether or not the event occured (i.e., death vs censored). `survfit()` creates a survival curve that you could then display or plot. `coxph()` implements the regression analysis, and models specified the same way as in regular linear models, but using the `coxph()` function.

[3] And there's a chi-square-like statistical test for these differences called the log-rank test that compare the survival functions categorical groups.

[4] See the multiple regression section of the essential statistics lesson.