# Lecture Six: Cox Propotional Hazards Models (I)

1. **The strengths and limitations of the non-parametric methods**

   (a) It's kind of preliminary data analysis, and does not require specific assumption about survival time. It's very useful in the analysis of a single sample of survival data, or in the comparison of two or more populations of survival time.

   (b) It does not provide estimate of the size of treatment difference; and it's not flexible enough for more complex data that include many *explanatory variables, or covariates* which may have an impact on the time that the patient survives.

   (c) Estimation of $S(t)$ is not very accurate at tail (i.e.: KM method).

2. **Terms and Definitions**

   (a) Terminology: dependent and independent (stat), variate and covariate (biostat), response and confounder (epidemiologist), endogenous and exogenous (economist).

   (b) *Confounding effects* and *interaction effects*:
   Confounding effects are contributed by various factors such as race and gender that cannot be seperated by the design under study; an interaction effect between factors is a joint effect with one or more contributing factors, the objective of a statistical interaction investigation is to conclude whether the joint contribution of two or more factors is the same as the sum of the contributions from each factor when considered alone.

3. **Proportional hazards models: Modeling the hazard function**

   In survival analysis, we focus on the hazard function most of the time. Notice the relationship between the survival function and hazard (or cumulative hazard function), other estimates, such as survival function, can be obtained if we have estimate of hazard function.

   (a) The proportional hazard model (Cox, 1972) is defined as

   $$h_1(t) = \psi h_0(t),$$

i. $\psi$ is the relative hazard (relative risk) or hazard ratio between two groups.

ii. $\psi$ is independent of time, but may be a function of covariates.

iii. Relation of survival functions:

$$S_1(t) = [S_0(t)]^{\psi},$$

comparison: $S_1(t)$ versus $S_0(t)$ when $\psi < 1, = 1$, or $> 1$.

iv. In terms of log cumulative hazard

$$log(H_1(t)) = log(\psi) + log(H_0(t)),$$

v. Example: One covariate - standard treatment vs new treatment. $X_1 = 0$ or 1.

$$h_i(t) = e^{\beta x_i} h_0(t),$$

where $\psi = exp(\beta)$.

(b) The general PH model: Assume p covariates $X_1, X_2, \ldots, X_p$,

$$h_i(t) = exp(\eta_i) h_0(t),$$

where

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi},$$

i. $h_0(t)$ is unspecified baseline hazard function which corresponds $\mathbf{x} = 0$.

ii. The baseline/reference group can be chosen either biologically or mathematically through transformation of covariates in a study. for example, if we define

$$\tilde{x}_1 = x_1 - \bar{x}_1,$$

then, the baseline group is the one with mean value of $x_1$.

iii. The hazard ratio, which is $e^{\eta}$, is constant over time.

iv. The impact of a covariate on the survival of the study sample is assumed to change the hazard only in proportion to the baseline hazard.

v. If we are primarily interested in the impact of covariates, including treatment, on survival, we don't have to know $h_0$.

vi. Proportinal hazard: A practical assumption, not a universal truth. How about additive hazard model

$$h_i(t) = h_0(t) + exp(\beta' \mathbf{x})?$$

or

$$h_i(t) = h_0(t) + \beta' \mathbf{x}?$$

(see **aareg()** in Splus 6.1.2; and Stat. in Med. 1993, 12:1569-1538 by Aalen)

4. Including covariates into the linear component

There are two types of variable on which a hazard function may depend, namely *variates and factors*. A variate is a variable takes numererical values (continuous scale of measurement), for example, age, systolic BP; A factor is a variable which takes a limited set of values known as the levels of the factor, for example, sex, treatment. A factor variable can be nominal or ordinal.

(a) Brief review of linear regression.

(b) Including a variate:

i. How to interpret the coefficient?

ii. Is the effect of the variate really in linear trend?

(c) Including a factor:

i. Coding and main effects.

ii. Interpretation of coefficient

(d) Including an interaction:

i. The **hierarchic principle**: Interactions should not be included unless the corresponding main effects are present (read section 3.2.3 at p62).

ii. Quantitative and qualitative (Ref: Biometrics, pp 361-372 (1985) by Gail and Simon).

iii. Coding and interpretation (hazard ratio for one of the factors involved in the interaction will depend on the level of the other).

iv. Checking: Those include testing and graphic examination (eg. interaction.plot() in Splus).

v. Including a mixed term:
Coding (use the table at p64 to illustrate).

5. Proportional hazard regression

   (a) Baseline hazard $h_0$ is unspecified because of 'not of direct interest'.

   (b) Covariates $\mathbf{x}$ is constant over time (can be relaxed)

   (c) if $\beta_k < 0$, $x_k$ is protective factor; if $\beta_k > 0$, $x_k$ is a risk factor; and if $\beta_k = 0$, then $x_k$ is not associated with survival.

   (d) Intercept is not included in the *linear component* of the model because it can be absorbed into the baseline hazard.

6. Checking the PH assumption
   For Cox model, besides assumptions of independent observations and independent random censoring mentioned in non-parametric methods, there is addtional PH assumption.

   (a) For A categorical covariate: The curves of log cumulative plot (i.e.: $log(-log(S(t)))$ vs $log(t)$) of KM estimate for all strata are parallel.

   (b) For a continuous covariate: Use methods based on residuals or statistic test (to be discussed later), or categorized continous covariate if categorization is meaningful.