STAT 6390: Analysis of Survival Data

Steven Chiou

Department of Mathematical Sciences,
University of Texas at Dallas

Accelerated Failure Time Model

- Regression models for survival data we studied so far:
  - Parametric models (Weibull model, exponential model).
  - Proportional hazards rate model (Cox model).
- Parametric AFT models have the common form

$$Y = \log(T) = X'\beta + \epsilon, \tag{1}$$

  where different parametric model can be specified through the distribution of ϵ.
- The semiparametric approach relaxes the assumption on ϵ.
- Last resort in the industrial testing, but the models of choice in medical research.

Semiparametric AFT model

- The parametric part of (1) is $X'\beta$.
- The non-parametric part of (1) is absence of a parametric assumption on ϵ.
- Still need to assume that ϵ_i are independent and identically distributed.
- For identifiability, the model does not contain an intercept.
- Carry out inference concerning β without deciding on a specific distribution on ϵ.

Least Squares Approach

- When there is no censoring (T_i's are completely observed, and all $\Delta_i = 1$)., the classical least-squares estimator of β is obtained by minimizing
$$\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$
in terms of β.
- The minimizer is the solution to the equation
$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - X_i'\beta) = 0,$$
where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

Least Squares Approach

- In the presence of censoring, the classical least-squares can not be used directly.
- Buckley and James (1979) proposed to replace Y_i with

$$\widehat{Y}_i(\beta) = \Delta_i \log(T_i) + (1 - \Delta_i)\left\{\frac{\int_{e_i(\beta)}^{\infty} u \mathrm{d}\widehat{S}_\beta\{e_i(\beta)}{1 - \widehat{S}_\beta\{e_i(\beta)\}} + X_i'\beta\right\}, \tag{2}$$

where $\widehat{S}_\beta(t)$ is the Kaplan-Meier estimator based on $\{e_i(\beta), \Delta_i\}$.

- The substitution (2) is a mean imputation.
- The resulting Buckly-James estimator is the solution to the following equation:

$$\sum_{i=1}^{n}(X_i - \bar{X})(\widehat{Y}_i(\beta) - X_i'\beta) = 0. \tag{3}$$

Least Squares Approach

- The Buckly-James estimating equation (3) is difficult to solve because the function is neither continuous nor (component-wise) monotone in β.
- Jin et al. (2006) proposed an iterative procedure to obtain a class of consistent and asymptotically normal estimators.
- Define

$$U(\beta, b) \sum_{i=1}^{n} (X_i - \bar{X})(\widehat{Y}_i(b) - X_i'\beta),$$

  for some constant b.
- Holding b fix, $\widehat{\beta}$ can be obtained by solving $U(\beta, b)$ for β.
- The closed form solution to β (holding b fix) is:

$$\beta = L(b) = \left\{ \sum_{i=1}^{n} (X_i - \bar{X})^{\otimes 2} \right\}^{-1} \left[\sum_{i=1}^{n} (X_i - \bar{X}) \{ \widehat{Y}_i(b) - \bar{Y}(b) \} \right],$$

  where $\bar{Y}(b) = n^{-1} \sum_{i=1}^{n} \widehat{Y}_i(b)$.

Least Squares Approach

- Jin et al. (2006) proposed to start with an initial value, $\widehat{\beta}_{(0)} \equiv b$, then iterate $\widehat{\beta}_{(m)} = L(\widehat{\beta}_{(m-1)})$, for $m \geq 1$, until convergence.
- Jin et al. (2006) also showed that for a consistent initial estimator $\widehat{\beta}_{(0)}$, $\widehat{\beta}_{(m)}$ is consistent and asymptotically normal for every $m \geq 1$.

Rank-based Approach

- The other approach in obtaining $\widehat{\beta}$ in a semiparametric AFT model is the *rank regression* approach.
- Generalizes the basic idea of the linear rank (Wilcoxon rank sum) test.
- For the ease of discussion, we will assume there is only one covariate.
- Let $Y_{(i)}$ be the sorted Y_i's.
- Let $X_{(i)}$ be the covariate value associated with the *i*th sorted Y_i's.
- A nonparametric rank-based test for the association between X and the Y_i can be based on the test statistic
- The linear rank test statistic is

$$U = \sum_{i=1}^{n} \phi_i \left(X_{(i)} - \bar{X} \right),$$

where ϕ_i is some score function attached to Y_i.

Rank-based Approach

- In the presence of censoring, the test statistic is modified to

$$U = \sum_{i=1}^{n} \phi_i \Delta_i \left(X_{(i)} - \bar{X}^* \right),$$

  where $\bar{X}^*$ denotes the average of the covariate values of all subjects at risk at time T_i.
- We need to be able to draw inference for β, therefore, we instead test whether the residuals of the AFT model are associated with the covariate.
- Define the residuals of the AFT model as $e_i(\beta) = Y_i - X_i' \widehat{\beta}$.
- We construct an estimating equation using the same procedure as before using $e_i(\beta)$.

Rank-based Approach

- In the presence of censoring and replacing Y_i with e_i, we have the test statistic

$$U(\beta) = \sum_{i=1}^{n} \phi \Delta_i \left\{ X_i - \frac{\sum_{j=1}^{n} X_j I\{e_j(\beta) \geq e_i(\beta)\}}{\sum_{j=1}^{n} I\{e_j(\beta) \geq e_i(\beta)\}} \right\},$$

  the weights, ϕ, plays the same role as the weights in the log-rank test.
- When $\phi = 1$, the resulting $U(\beta)$ corresponds to the log-rank statistics.
- When $\phi = \sum_{j=1}^{n} I\{e_j(\beta) \geq e_i(\beta)\}$ corresponds to the Gehan's statistics.

Rank-based Approach

- The estimator, $\widehat{\beta}$, can be obtained by solving $U(\beta) = 0$.
- With a general weight, it is difficult to solve the equation $U(\beta) = 0$ because $U(\beta)$ is neither continuous nor component-wise monotone in β.
- With the Gehan weight, $U(\beta)$ reduces to

$$U_G(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i (X_i - X_j) I\{e_j(\beta) \geq e_i(\beta)\}.$$

- $U_G(\beta)$ is component-wise monotone in β, but is also not continuous.
- Procedures are available to smooth $U_G(\beta)$ to facilitate the usage of the AFT model.

Reference

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Jin, Z., Lin, D., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika* **93,** 147–161.