# Newly Added: Semi-parametric Linear Regression

PubH 7450

©Wei Pan

Email: weip@biostat.umn.edu

Http: www.biostat.umn.edu/∼weip

- Background: discussed so far,
  1) semi-parametric PHM; –use partial likelihood.
  2) parametric AFT; parametric PHM can be done similarly;–use full likelihood.
- How about semi-parametric AFT?
  partial likelihood does not work; topic here.
- Discuss one of the earliest, Buckley-James estimator; intuitive, related to the EM discussed for the one-sample problem.
  Refs.: Buckley, J. and James, I. (1977). Linear regression with censored data. *Biometrika*, 66, 429-436.
  Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, 69, 521-531.
  See also Section 2.2 in Chapter 6 of Miller (1980).
  More recent ones, based on estimating functions, e.g.,
  Wei, Ying and Lin (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77,

845-851.

- Problem: given data $(T_j, \delta_j, Z_j)$, $j = 1, ..., n$,
  AFT model: $Y_j = \log X_j = Z_j'\theta + \epsilon_j$, where $\epsilon_j \overset{iid}{\sim} F_0$.
  Semi-parametric: no parametric assumption on $F_0$ (in contrast to parametric approaches where $F_0$ is assumed to be, e.g. an extreme value distr $\Longrightarrow$ $X_j$ as Weibull).
  Goal: inference on $\theta$.
- Basic idea: If $Y_j$ were observed, then could use OLSE $\hat{\theta}$ and so on.
  Challenge: $Y_j$ may be censored!
  Solution: impute!
- How?
  Estimate $E(Y_j | Y_j > \log T_j = y_j)$.
- First, if we have an initial $\hat{\theta}$, then
  $r_j = y_j - Z_j'\hat{\theta}$,

use $(r_j, \delta_j)$ to estimate $F_0$: use .....estimator, $\hat{F}_0 = \hat{F}_0(\hat{\theta})$.

- Suppose
  $\hat{p}_j = Pr_{\hat{F}_0}(\epsilon_j = r_j)$, $j = 1, 2, ...$
- The key:
  $\widehat{E}(\epsilon_j | \epsilon_j > r_j) = \dfrac{\sum_{r_k > r_j} \hat{p}_k r_k}{\sum_{r_k > r_j} \hat{p}_k}$.
  $\widehat{E}(Y_j | Y_j > \log T_j) = Z_j' \hat{\theta} + \widehat{E}(\epsilon_j | \epsilon_j > r_j)$.
- Define
  $\hat{Y}_j = \log T_j$ if $\delta_j = 1$;
  $\hat{Y}_j = \widehat{E}(Y_j | Y_j > \log T_j)$ if $\delta_j = 0$.
- Use complete data $(\hat{Y}_j, Z_j)$ to get an updated OLSE $\hat{\theta}$:
  $\hat{\theta} = [(Z - \bar{Z})'(Z - \bar{Z})]^{-1}(Z - \bar{Z})'\hat{Y}$.
- Repeat the above steps until convergence (if any).
- B-J suggested

$\widehat{Var}(\hat{\theta}_k) = \frac{\hat{\sigma}_u^2}{\sum_{\delta_j=1}(Z_{jk}-\bar{Z}_{u,k})^2}$,

$\widehat{Cov}(\hat{\theta}) = \hat{\sigma}_u^2[(Z_j - \bar{Z}_u)'\Delta(Z_j - \bar{Z}_u)]^{-1}$,

where $\hat{\sigma}_u^2 = \sum_{\delta_j=1}(y_j - \bar{y}_u - (Z_j - \bar{Z}_u)'\hat{\theta})^2/(n_u - 2)$, and

$\Delta = diag(\delta_j)$.

$u$: uncensored observations; $n_u = \sum_{j=1}^{n} \delta_j$.

No theoretical justification, but seems to work reasonably well in simulations; my view: perhaps upward-biased, why?

- Variance estimation is challenging (usually involving unknown density functions, and the estimating function is usually non-smooth); use bootstrap; under current investigation.
- Example (Miller and Halpern 1982): Stanford transplant data. Compared to other methods, the Cox and B-J estimators agreed more to each other, and were claimed to be the winners.
- Software: R function `bj()` in package `rms`.

The author (Dr Frank Harrell): "The program implements the algorithm as described in the original article by Buckley & James. Also, we have used the original Buckley & James prescription for computing variance/covariance estimator. This is based on non-censored observations only and does not have any theoretical justification, but has been shown in simulation studies to behave well. Our experience confirms this view." "The `bootcov` function may be worth using with `bj` fits, as the properties of the Buckley-James covariance matrix estimator are not fully known for strange censoring patterns."

- Example: R