Current status least-squares regressions

2021 Spring KSS

Taehwa Choi* and Sangbum Choi

May 28, 2021

Department of Statistics, Korea University
E-mail: taehwa_choi@korea.ac.kr

Survival data and censoring

- Survival data are mostly related with event time.
  - e.g. lifetime of patient, failure time of machine, duration of job employment.
- Event time can be shaded by censoring.
- Under periodic monitoring, event time may not be available (interval censoring).
- Categorized by the number of monitoring:
  - Case-1: left or right censoring (i.e., *current status*).
  - Case-2: event time lies between censoring time range.

Current status data

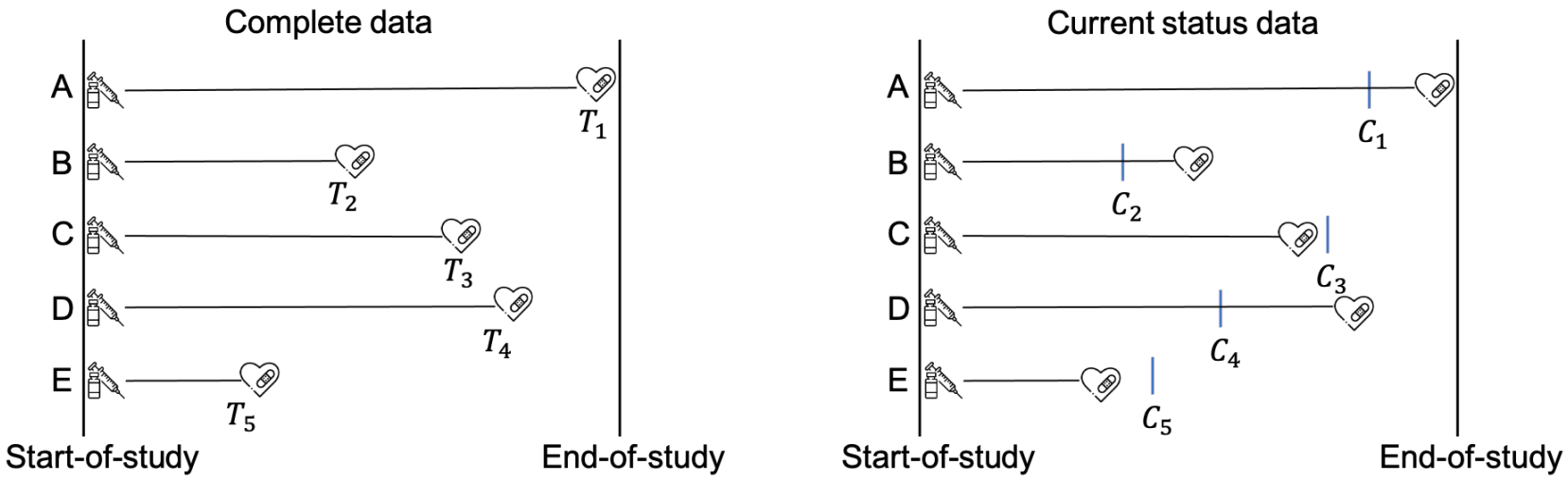

Figure 1: Illustration of current status data

- Complete data: (T, X)
- Current status data: $\{C, \Delta = I(T \leq C), X\}$

$\Rightarrow$ *There is no observed response.*

Event time data analysis

- Nonparametric estimation (Groeneboom and Wellner, 1992; Wellner and Zhan, 1997).
  - $n^{1/3}$ rate for distribution function parameter.
- Proportional hazards model (Huang, 1996, 1999)

$$\Lambda(t|X) = \Lambda_0(t)\exp(X^T\beta_0),\ \Lambda(t) = \exp\{-S(t)\},\ S(t) = P(T > t)$$

- Linear transformation model (Zhang et al., 2013; Zeng et al., 2016)

$$\log \Lambda_0(T) = X^T\beta_0 + \epsilon$$

$$\Rightarrow \textit{hazard-based model (probabilistic)}$$

Semiparametric AFT model

- Semiparametric linear model has a form of

$$Y_i \equiv \log T_i = X_i^T \beta_0 + \epsilon_i, \ (i = 1, \ldots, n)$$

  with event time T_i, covariate X_i, true regression parameter β_0 and unspecified random error $\epsilon_i \sim F$.

- Direct interpretation on event time with covariate.

- We generalize the Buckley-James method to the current status data.

Latent time imputation

- Key idea is to *impute the latent event time*,

$$Y_i^* = \Delta_i E(Y_i | T_i \leq C_i, X_i) + (1 - \Delta_i) E(Y_i | T_i > C_i, X_i).$$

- Let $e_i(\beta) = \log C_i - X_i^T \beta$.

$$E(Y_i | T_i \leq C_i, X_i) = E[\epsilon_i | \epsilon_i \leq e_i(\beta)] + X_i^T \beta = \frac{\int_{-\infty}^{e_i(\beta)} u dF(u)}{F\{e_i(\beta)\}} + X_i^T \beta,$$

$$E(Y_i | T_i > C_i, X_i) = E[\epsilon_i | \epsilon_i > e_i(\beta)] + X_i^T \beta = \frac{\int_{e_i(\beta)}^{\infty} u dF(u)}{1 - F\{e_i(\beta)\}} + X_i^T \beta.$$

However, F is unknown in semiparametric model.

Estimation of imputed latent time

- Replace F to empirical distribution function $\hat{F}$:

$$\hat{Y}_i(\beta) = \Delta_i \left[\frac{\int_{-\infty}^{e_i(\beta)} u d\hat{F}(u)}{\hat{F}\{e_i(\beta)\}} \right] + (1 - \Delta_i) \left[\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}(u)}{1 - \hat{F}\{e_i(\beta)\}} \right] + X_i^T \beta. \tag{1}$$

$$\Rightarrow \textit{estimation of unknown } F \textit{ is crucial.}$$

- Log-likelihood of Λ for current status data:

$$\ell(\Lambda) = \sum_{i=1}^{n} \Delta_i \log F(e_i) + (1 - \Delta_i) \log\{1 - F(e_i)\}$$

NPMLE of distribution function F

- Sieve MLE (Shen, 2000): proper sieve and knot selection.
- Self-consistency equation (Turnbull, 1976): does not guarantee the NPMLE of F (Wellner and Zhan, 1997).

$$\hat{F}(t) = n^{-1} \sum_{i=1}^{n} \left[\Delta_i \frac{\int_{-\infty}^{e_i(\beta)} u d\hat{F}(u)}{\hat{F}\{e_i(\beta)\}} + (1 - \Delta_i) \frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}(u)}{1 - \hat{F}\{e_i(\beta)\}} \right].$$

- Alternative: expectation-maximization (EM) algorithm.

NPMLE of distribution function F

- Assumption: Λ is nondecreasing step-function.
  - Jump point: unique value of C_i.
  - Jump sizes: $d\Lambda_k \equiv \lambda_k$, $(k = 1, \ldots, m \leq n)$
- Note: $F = 1 - \exp(-\Lambda)$, $\Lambda(t) = \int_0^t \lambda(u)du$.
- Log-likelihood function of Λ:

$$\begin{aligned}\ell(\Lambda) &= \sum_{i=1}^{n} \Delta_i \log F(e_i) + (1 - \Delta_i) \log\{1 - F(e_i)\} \\ &= \sum_{i=1}^{n} \Delta_i \log[1 - \exp\{-\Lambda(e_i)\}] - \Delta_i \Lambda(e_i).\end{aligned}$$

EM-algorithm (I)

- Independent latent variable: $W_{ik} \sim \text{Poisson}(\lambda_k)$.
  - $\sum_{t_k \le e_i} W_{ik} \begin{cases} > 0, & \Delta_i = 1; \\ = 0, & \Delta_i = 0. \end{cases}$
- Log-likelihood function of Λ with W_{ik}:

$$\begin{aligned} \ell(\Lambda) &= \sum_{i=1}^{n} \Delta_i \log[1 - \exp\{-\Lambda(e_i)\}] - (1 - \Delta_i)\Lambda(e_i) \\ &= \sum_{i=1}^{n} \left[\sum_{k=1}^{m} \left\{ \Delta_i \log P\left(\sum_{t_k \le e_i} W_{ik} > 0 \right) \right.\right. \\ &\qquad \left.\left. + (1 - \Delta_i) \log P\left(\sum_{t_k < e_i} W_{ik} = 0 \right) \right\} \right]. \end{aligned}$$

EM-algorithm (II)

- Define complete-data log-likelihood $\ell^c(\Lambda, W)$ when W_{ik} are known.

- E-step: conditional expectation of $\ell^c(\Lambda, W)$ given data;

$$E\{\ell^c(\Lambda, W)\} = \sum_{i=1}^{n}\sum_{k=1}^{m} I(t_k \le e_i)\{\log \lambda_k \hat{E}(W_{ik}) - \lambda_k\},$$

  where

$$\hat{E}(W_{ik}) = \Delta_i \left\{ \frac{\lambda_k I(t_k \le e_i)}{1 - e^{-\sum_{t_k \le e_i} \lambda_k}} \right\} + \lambda_k I(t_k > e_i). \qquad (2)$$

- M-step: $\arg\max_{\lambda_k} E\{\ell^c(\Lambda, W)\}$

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} I(t_k \le e_i)\hat{E}(W_{ik})}{\sum_{i=1}^{n} I(t_k \le e_i)}, \; k = 1, \ldots, m. \qquad (3)$$

Parameter estimation (I)

- Estimating equation for β:

$$U_n(\beta, b) = \sum_{i=1}^{n} (X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b) - (X_i - \bar{X})^T \beta\}, \quad (4)$$

  where $\bar{Y}(b) = n^{-1} \sum_{i=1}^{n} \hat{Y}_i(b)$ and $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.

- However, $U_n(\beta, \beta) = 0$ is discontinuous about β.

- Jin et al. (2006)'s algorithm (Nelder-Mead simplex):

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

  with $a^{\otimes 2} = aa^T$.

Parameter estimation (II)

Step 1. Initiate β, denoted as $\hat{\beta}_{(0)}$.

Step 2. For kth stage ($k \geq 1$), estimate $\hat{F}_{(k)}(t; \hat{\beta}_{(k-1)})$;

Step 3. Compute $\hat{Y}(\hat{\beta}_{(k-1)})$ and $\hat{\beta}_{(k)} = L_n(\hat{\beta}_{(k-1)})$;

Step 4. Set $k \leftarrow k + 1$, and repeat steps (2) and (3) until convergence.

Asymptotic results

Theorem (Consistency)
$(\hat{\beta}, \hat{F})$ are strongly consistent estimators of (β_0, F_0)

Theorem (Rate of convergence)
$d\{(\hat{\beta}, \hat{F}), (\beta_0, F_0)\} = \|\hat{\beta} - \beta_0\| + \|\hat{F} - F_0\|_v = O_p(n^{-1/3})$, where $\|F\|_v = \sup_{h \in \mathcal{H}} |\int h(t) dF(t)|$ with $\mathcal{H} = \{h : \|h\|_{BV[0,\tau]} \leq 1\}$.

Theorem (Asymptotic normality)
$n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a zero-mean normal random vector.

Simulation set-up

- Latent event times are generated from

$$\log T = 0.5X_1 - 0.5X_2 + \epsilon.$$

- $X_1 \sim \text{Uniform}(-1, 1)$ and $X_2 \sim \text{Bernoulli}(0.5)$.
- ϵ follows (i) $N(0, 1)$, (ii) Gumbel$(-0.5, 0.5)$ and (iii) log of Gamma$(1.5, 1)$.
- Monitoring time $C \sim \text{Uniform}(0.1, c_0)$ for $c_0 > 0.1$.

Simulation results (I)

Table 1: Simulation results for current status data.

			Buckley-James				Complete-data			
Error	n	Par	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Normal	150	β_1	−0.017	0.125	0.120	0.915	0.000	0.073	0.071	0.943
		β_2	0.021	0.138	0.134	0.935	0.005	0.081	0.082	0.950
	300	β_1	0.000	0.089	0.087	0.946	0.006	0.047	0.050	0.962
		β_2	−0.005	0.095	0.097	0.954	−0.011	0.061	0.058	0.931
EV	150	β_1	−0.057	0.129	0.128	0.901	−0.004	0.091	0.091	0.960
		β_2	0.060	0.141	0.143	0.930	0.001	0.105	0.105	0.946
	300	β_1	−0.037	0.097	0.090	0.912	−0.002	0.068	0.064	0.925
		β_2	0.030	0.107	0.102	0.910	0.000	0.074	0.074	0.942
Gamma	150	β_1	−0.038	0.182	0.181	0.925	−0.003	0.139	0.137	0.942
		β_2	0.031	0.206	0.204	0.945	−0.005	0.155	0.158	0.954
	300	β_1	−0.024	0.129	0.130	0.949	−0.002	0.098	0.096	0.947
		β_2	0.026	0.150	0.145	0.919	0.003	0.112	0.111	0.952

Simulation results (II)

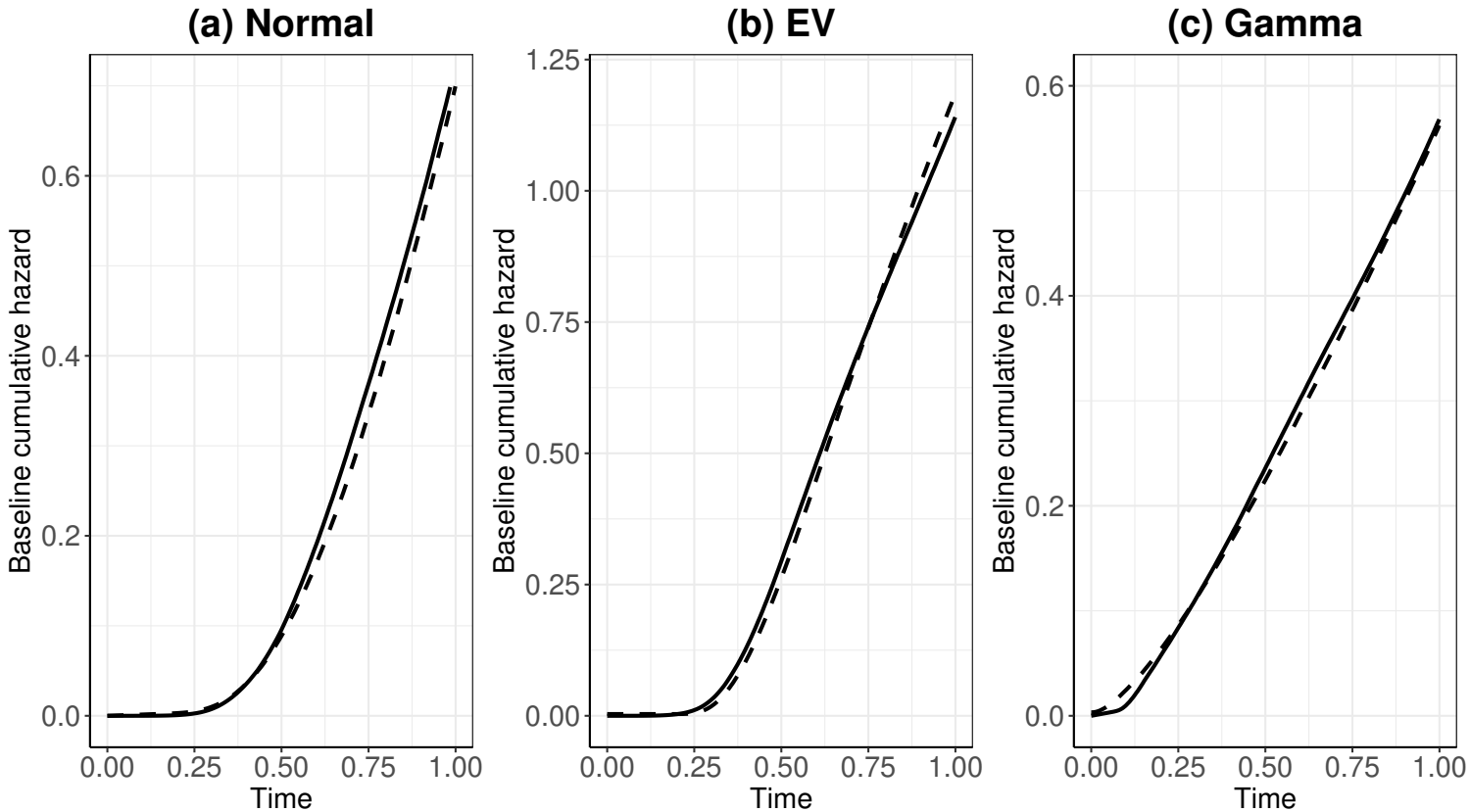

Figure 2: Baseline hazard function curves with $n = 300$. Dashed and solid lines are true and estimated cumulative hazard functions, respectively.

Discussion

- Generalization of Buckely-James method to current status data.
- Similar performance to the complete-case analysis.
- Ongoing researches
  - Extension to general interval-censoring:

$$(0, L), \quad (L, R), \quad (R, \infty).$$

  - Another semiparametric regression: rank-based model.

References (I)

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*, volume 19. Springer, New York.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Stat.*, 24(2):540–568.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Stat. Sin.*, 9:501–519.

Jin, Z., Lin, D., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93(1):147–161.

Shen, X. (2000). Linear regression with current status data. *J. Am. Stat.*, 95(451):842–852.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 38(3):290–295.

References (II)

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *J. Am. Stat.*, 92(439):945–959.

Zeng, D., Mao, L., and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–271.

Zhang, B., Tong, X., Zhang, J., Wang, C., and Sun, J. (2013). Efficient estimation for linear transformation models with current status data. *Commun. Stat-Theor. M.*, 42(17):3191–3203.