Regression Analysis of Informatively Interval-censored Failure Time Data

Jianguo (Tony) Sun

Department of Statistics, University of Missouri
Columbia, MO

Joint work with Tao Hu, Ling Ma, Peijie Wang, and Hui Zhao
October 14, 2016

Outline

Case I interval-censored data:

1. Each subject is observed only once for the determination of the occurrence of the failure event of interest.
2. The observed data have the form $\{(C, I(T \leq C)\}$. That is, T is either left- or right-censored, instead of being observed exactly.
3. The areas giving such data include *demographic studies, social sciences, tumorigenicity experiments*.
4. Also often referred to as *current status data*.

Examples from tumorigenicity experiments:

1. *Failure time of interest* is usually the time until tumour onset.
2. *Observation time* is usually the time of death or sacrifice.
3. *Observed information* is the presence or absence of tumors at the the observation time.

Examples from tumorigenicity experiments:

1. *Non-lethal tumor:* The tumor onset has no effect on animal death. That is, the tumour onset time and the death time are independent.
2. *Lethal tumour:* The tumor onset will kill the animal. That is, the death time is the same as the tumour onset time.
3. *Intermediate tumour:* The tumours are between non-lethal and lethal. That is, the death time may depend on the tumour onset time.

Case II or K interval-censored data:

1. *Case II interval-censored data* —
   In general, the formulation will be $T \in (L, R]$;
   Another simpler, common formulation is
   $\{ U < V, I(0 < T \leq U), I(U < T \leq V), I(V < T) \}$.
2. *Case K interval-censored data* —
   There exists a sequence of observation times
   $U_0 = 0 < U_1 < U_2 < \cdots < U_K$ and the observed data have the form $\{ U_j, I(U_{j-1} < T \leq U_j), j = 1, ..., K \}$.
3. A common area giving such data is clinical trials or periodic follow-up studies.

Examples from HIV and AIDS studies:

1. *AIDS cohort or follow-up studies* — HIV infection time, AIDS incubation time.
   De Gruttola, Lagakos (1989); Sun, Liao, Pagano (1999); Wang, Tong, Zhao, Sun (2015).
2. *AIDS clinical trials* — Time to the AIDS diagnosis or death.
   Gómez, Espinal, Lagakos (2003); Goggins, Finkelstein, 2000; Zhou, Hu, Sun (2016).

Censoring mechanism

1. **Case I interval-censored data:**
   Noninformative censoring —

$$P(T \le t\,,\, C \le c \,|\, Z) \;=\; P(T \le t \,|\, Z)\, P(C \le c \,|\, Z)\,.$$

   Informative censoring — T and C are correlated even given covariates.
2. **Case II or K interval-censored data:**
   Noninformative censoring —

$$P(T \le t \,|\, Z, L = l, R = r, L < T \le R) = P(T \le t \,|\, Z, l < T \le r).$$

   Informative censoring — The failure time of interest T and the observation times (L, R) or U_j's may be correlated.

Some existing literature

1. **Informative case I interval-censored data:**
   Lagakos and Louis (1988),
   Zhang, Sun, and Sun (2005),
   Ma, Hu and Sun (2015),
   Zhao, Hu, Ma, Wang and Sun (2015).
2. **Informative case II or K interval-censored data:**
   Zhang, Sun, Sun, and Finkelstein (2007),
   Zhao, Hu, Ma, Wang and Sun (2015),
   Ma, Hu and Sun (2016),
   Wang, Zhao and Sun (2016).

Notation

Consider a failure time study consisting of n independent subjects.

1. T_i: Failure time of interest.
2. C_i: Observation time that may be related to T_i.
3. ζ_i: administrative or stopping time independent of T_i and C_i.
4. Z_i: p-dimensional vector of covariates.
5. Define $\tilde{C}_i = \min\{C_i, \zeta_i\}$.
6. Also define $\Delta_i = I(C_i \leq \zeta_i)$ and $\delta_i = I(T_i \leq \tilde{C}_i)$.

The marginal models

1. Suppose that

$$\lambda_T(t|Z) = \lambda_1(t)\exp(Z'\beta)$$

$$\lambda_C(c|Z) = \lambda_2(c)\exp(Z'\gamma)$$

2. Let F_T and F_C denote the cdf of T and C given Z, respectively, and f_C the pdf of C given Z. Then

$$F_T(t) = 1 - \exp\{-\Lambda_T(t)\exp(Z'\beta)\},$$

$$F_C(c) = 1 - \exp\{-\Lambda_C(c)\exp(Z'\gamma)\},$$

$$f_C(c) = \lambda_2(c)\exp(Z'\gamma)\exp\{-\Lambda_C(c)\exp(Z'\gamma)\},$$

where $\Lambda_T(t) = \int_0^t \lambda_1(s)ds$ and $\Lambda_C(c) = \int_0^c \lambda_2(s)ds$.

The Copula model for the joint distribution

1. Let F denote the joint distribution function of T and C given Z.
2. Then there exists a copula function $C_\alpha(u, v)$ defined on $I^2 = [0, 1] \times [0, 1]$ such that

$$F(t, c) = C_\alpha(F_T(t), F_C(c)), \ t \geq 0, c \geq 0$$

where α specifies the degree of association between T and C.

The likelihood function

$$L(\theta) = \prod_{i=1}^{n} \Big\{ \{m_\alpha(F_T(\tilde{c}_i), F_C(\tilde{c}_i))\}^{\delta_i} \{1 - m_\alpha(F_T(\tilde{c}_i), F_C(\tilde{c}_i))\}^{1-\delta_i}$$

$$\times f_C(\tilde{c}_i) \Big\}^{\Delta_i} \Big\{ \{F_T(\tilde{c}_i) - C_\alpha(F_T(\tilde{c}_i), F_C(\tilde{c}_i))\}^{\delta_i}$$

$$\times \{1 - F_T(\tilde{c}_i) - F_C(\tilde{c}_i) + C_\alpha(F_T(\tilde{c}_i), F_C(\tilde{c}_i))\}^{1-\delta_i} \Big\}^{1-\Delta_i},$$

where $\theta = (\beta, \gamma, \Lambda_T, \Lambda_C)$ and

$$m_\alpha(F_T(t), F_C(c)) = P(T \le t | C = c, Z) = \frac{\partial C_\alpha(u, v)}{\partial v}|_{u=F_T(t), v=F_C(c)}.$$

I-spline approximation

We consider the sieve maximum likelihood approach by approximating $\Lambda_T(\cdot)$ and $\Lambda_C(\cdot)$ with monotone I-spline functions. Define the sieve space for $\theta = (\beta, \gamma, \Lambda_T, \Lambda_C)$ as

$$\Theta_n = \left\{\theta_n = (\beta, \gamma, \Lambda_{Tn}, \Lambda_{Cn})\right\},$$

where

$$\Lambda_{Tn}(t) = \sum_{j=1}^{m+k_n} \xi_j I_j(t),\ \xi_j \geq 0,\ j = 1, \ldots, m + k_n, t \in [0, u_c],$$

$$\Lambda_{Cn}(t) = \sum_{j=1}^{m+k_n} \eta_j I_j(t),\ \eta_j \geq 0,\ j = 1, \ldots, m + k_n, t \in [0, u_c]$$

with u_c being the upper bound of $\{\tilde{C}_i : i = 1, \ldots, n\}$.

An example of I-spline base functions

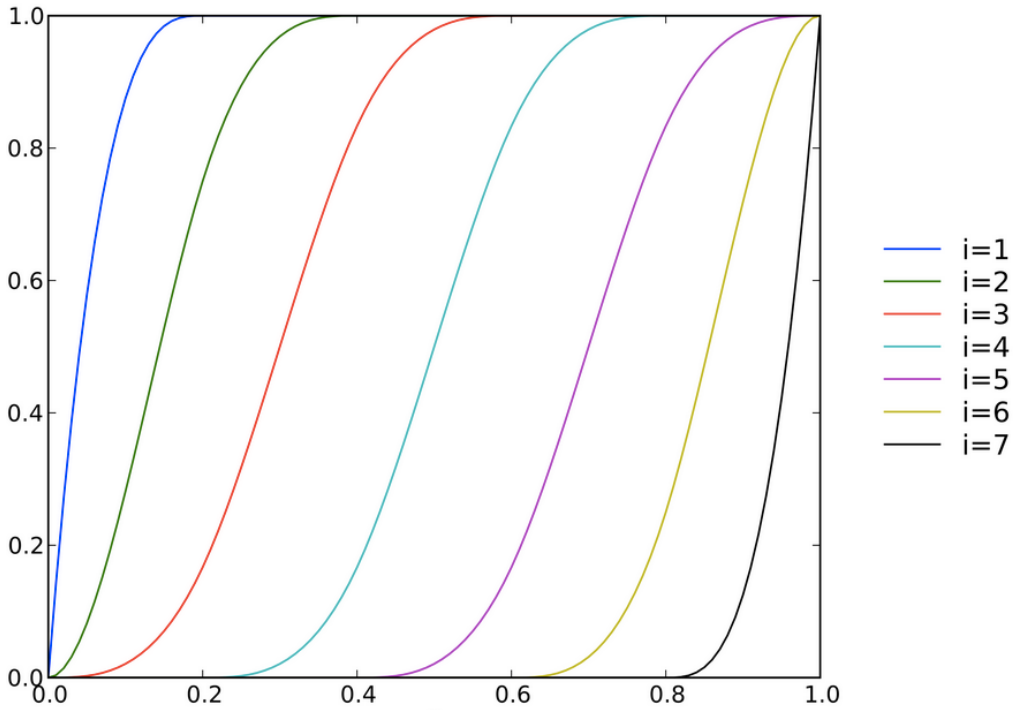

Figure : The *I*-spline family of order 3 with 4 interior knots

Sieve maximum likelihood estimators

1. Define the estimator $\hat{\theta} = (\hat{\beta}, \hat{\gamma}, \hat{\Lambda}_{Tn}(\cdot), \hat{\Lambda}_{Cn}(\cdot))$ as the value of θ that maximizes the log likelihood function $l(\theta) = \log(L(\theta))$ over Θ_n.
2. Under some regularity conditions, $\hat{\beta}$ and $\hat{\gamma}$ are strongly consistent and $\|\hat{\Lambda}_{Tn} - \Lambda_{T0}\|_2 \longrightarrow 0$ $\|\hat{\Lambda}_{Cn} - \Lambda_{C0}\|_2 \longrightarrow 0$ almost surely.
3. Also we have

$$n^{1/2}\left((\hat{\beta} - \beta_0)', (\hat{\gamma} - \gamma_0)'\right)' \rightarrow N(0, \Sigma)$$

and $(\hat{\beta}', \hat{\gamma}')'$ are semiparametrically efficient.

4. The covariance matrix Σ can be estimated by the inverse of the observed information matrix by treating Λ_{Tn} and Λ_{Cn} as finite-dimensional nuisance parameters.

Some remarks

1. In the proposed method, it has been assumed that the copula function C_α and the association parameter α are known (Zheng and Klein, 1995).
2. For the selection of m and k_n, one can try different values and compare the results, or employ the Akaike information criterion (AIC).
3. One application of the proposed approach is for the sensitivity analysis on how various levels of the association may affect the analysis results (Lagakos and Louis, 1988).

Notation

Consider a failure time study consisting of n independent subjects.

1. T_i: Failure time of interest.
2. Observation times $\{U_{i0} < U_{i1} < \cdots < U_{iK_i}\}$.
3. Z_i: p-dimensional vector of covariates.
4. The observed data have the form

$$\{O_i = (U_{ij}, \delta_{ij} = I(U_{ij-1} < T_i \leq U_{ij}), Z_i, j = 1, ..., K_i); i = 1, ..., n\}.$$

Models

1. Define $N_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \leq t)$, $i = 1, ..., n$.
2. Suppose that there exists a latent variable μ_i and given Z_i and μ_i, the hazard function of T_i has the form

$$\lambda_T(t|Z_i, \mu_i) = \lambda_0(t) \exp(Z_i^T \beta_1 + \mu_i \beta_2) .$$

3. Also Suppose that given Z_i and μ_i, $N_i(t)$ is a nonhomogeneous Poisson process with the intensity function

$$\lambda_N(t|Z_i, \mu_i) = \lambda_{0N}(t) \exp(Z_i^T \alpha + \mu_i) .$$

4. Huang and Wang (2004), Sun (2006).

Likelihood function

Under the assumptions above, the full likelihood function has the form

$$L(\beta_1, \beta_2, \alpha, \Lambda_0, \Lambda_{0N}, g)$$

$$= \int L_T(\beta_1, \beta_2, \Lambda_0 | \mu_i' s)\, L_N(\alpha, \Lambda_{0N} | \mu_i' s)\, g(\mu_1, ..., \mu_n) \prod_{i=1}^{n} d\mu_i \,.$$

Step 1: Estimation of α and Λ_{0N}

1. Let the $s_{(l)}$'s be the ordered and distinct values of the observation times $\{U_{ij}\}$, $d_{(l)}$ the number of the observation times equal to $s_{(l)}$, and $R_{(l)}$ the number of observation times satisfying $U_{ij} \leq s_{(l)} \leq \tau_i$ among all subjects.
2. Then one can estimate α and Λ_{0N} by

$$\sum_{i=1}^{n} w_i \widetilde{z}_i \left(K_i \widehat{\Lambda}_{0h}^{-1}(\tau_i) - E(e^{u_i}) \exp(z_i^T \alpha) \right) = 0 ,$$

$$\widehat{\Lambda}_{0N}(t) = \prod_{s_{(l)} > t} (1 - \frac{d_{(l)}}{R_{(l)}}) ,$$

   where $\widetilde{z}_i^T = (1, z_i^T)$ and the w_i's are some weights.
3. Huang and Wang (2004).

Step 2: Estimation of $\beta = (\beta_1^T, \beta_2)^T$ and Λ_0

1. Let $0 = t_0 < t_1 < \ldots < t_{q_n} = \tau_0$ denote a partition of the observation period $[0, \tau_0]$ and define

$$\Lambda_n(t) = I_1(t)\, e^{\gamma_1} \frac{t - t_0}{t_1 - t_0} + \sum_{l=2}^{q_n} I_l(t) \left(\sum_{k=1}^{l-1} e^{\gamma_k} + e^{\gamma_l} \frac{t - t_{l-1}}{t_l - t_{l-1}} \right),$$

where $I_l(t) = I(t_{l-1} < t \leq t_l)$ and $\gamma = (\gamma_1, \ldots, \gamma_{q_n})^T$ are unknown parameters.

2. Define $(\widehat{\beta}, \widehat{\Lambda}_n)$ or $\widehat{\theta} = (\widehat{\beta}^T, \widehat{\gamma}^T)^T$ of $\theta^T = (\beta^T, \gamma^T)$ as the values that maximize $L_T(\beta, \Lambda_n | \widehat{u}_i' s)$, where

$$\widehat{u}_i = \log \left\{ \frac{K_i}{\widehat{\Lambda}_{0h}(\tau_i) \exp(z_i^T \widehat{\alpha})} \right\}.$$

Step 2: Estimation of $\beta = (\beta_1^T, \beta_2)^T$ and Λ_0

1. It can be shown that under some regularity conditions, $\widehat{\beta}$ is consistent and $\sqrt{n}(\widehat{\beta} - \beta_0)$ converges to the multivariate normal distribution with mean zero.
2. Variance estimation: the bootstrap procedure.

Some remarks

1. Developed a sieve estimation procedure for regression analysis of general interval-censored data in the presence of informative interval censoring.
2. The method can be seen as a generalization of that given in Huang and Wang (2004).
3. Limitations or future research: Nonhomogeneous Poisson process; Model checking; Derivation of the asymptotic properties of $\hat{\Lambda}_n(t)$.

Analysis of a tumorigenicity study

1. Groups of 50 male and female F344/N rats and B6C3F$_1$ mice were exposed to chloroprene at concentrations of $0, 12.8, 32$ or 80 ppm by inhalation for 2 years.
2. Each animal was examined for various tumors at its death time. Some animals died naturally during the study and those who survived at the end of the study were sacrificed.
3. Objective: Compare the tumour growth rates between different dose groups.
4. We'll focus on a specific type of tumour (A/B A) for male and female B6C3F$_1$ mice in the control (0 ppm, $Z = 0$) and high dose (80 ppm, $Z = 1$) groups.

Analysis of a tumorigenicity study

1. Assumed the proportional hazards models for the tumour onset time T and natural death time C;
2. Used quadratic I-spline functions for approximating $\Lambda_T(\cdot)$ and $\Lambda_C(\cdot)$;
3. Considered Gumbel, FGM and Frank copulas used in simulation studies for the analysis;
4. AIC was used for the selection of Kendall's τ, k_n.

Results

Table 4: Estimated dose effects under the FGM model.

	$\hat{\beta}$			$\hat{\gamma}$			
τ	estimate	SEE	$p-$value ($\times 10^8$)	estimate	SEE	p-value ($\times 10^{11}$)	AIC
			$k_n = 3$				
-2/9	1.9980	0.3494	1.0717	1.3625	0.1986	0.6837	1083.126
-1/9	2.1313	0.3485	0.0966	1.3953	0.2006	0.3492	1080.084
0	2.2795	0.3952	0.8003	1.4112	0.2005	0.1929	1077.819
1/9	2.4151	0.3744	0.0111	1.4162	0.2002	0.1492	1076.560
2/10	2.4611	0.3603	0.0008	1.4148	0.2000	0.1514	1076.206
2/9	2.4654	0.3573	0.0005	1.4140	0.2000	0.1561	1076.219
			$k_n = 5$				
-2/9	2.0258	0.3571	1.3988	1.3177	0.1988	3.3907	1063.924
-1/9	2.0999	0.3481	0.1620	1.3558	0.2005	1.3657	1059.394
0	2.2785	0.3946	0.7744	1.3750	0.2004	0.6791	1057.002
1/9	2.4104	0.3744	0.0121	1.3809	0.2001	0.5142	1055.971
2/10	2.4566	0.3601	0.0009	1.3828	0.1998	0.4490	1055.449
2/9	2.4615	0.3572	0.0006	1.3830	0.1997	0.4401	1055.346

Result summary

1. The results are different and depends on the possible correlation between the tumor onset time and the observation time.
2. All the $p-$values in this table, for both β and γ are significant with any reasonable level of significance, which implies that mice in the high dose group have significantly shorter tumor onset time and death time.

Analysis of an AIDS clinical trial

1. A clinical trial (ACTG 359) for comparison of 6 different antiretroviral treatment regimens for AIDS patients.
2. Variable of interest: the first time at which the number of RNA copies, a measure of the viral load level, drops below the threshold of 500 viral copies/ml.
3. The patient blood samples were supposed to be collected at month 1, 2, 3, 4, 6, 8, 10 and 12 for the determination of their RNA copy counts.

Analysis of an AIDS clinical trial

1. Observed data: 271 AIDS patients whose numbers of RNA copies were measured at least once during the 12 months follow-up period in addition to their initial numbers of RNA copies.
2. For each patient: the number of observation times K_i's and observation times U_{ij}'s, and the event indicators δ_{ij}'s.
3. Define $Z_i = 0$ if the initial number of RNA copies of the subject is below 20000 viral copies/ml and 1 otherwise.

Analysis of an AIDS clinical trial

1. $\hat{\beta}_1 = -1.3568$ with SD $(\hat{\beta}_1) = 0.1908$.
2. $\hat{\beta}_2 = 2.3024$ with SD $(\hat{\beta}_2) = 0.5233$.
3. $\hat{\alpha} = -0.1485$ with SD $(\hat{\alpha}) = 0.0341$.

Summary

1. Investigated regression analysis of interval-censored failure time data with informative censoring mechanism.
2. Developed some sieve estimation procedures with the use of smooth functions.
3. Limitations: Selection of the copula model and the association parameter; Data from non-proportional hazards models.

Future research

1. Interval-censored data arising from other models.
2. Bivariate or multivariate interval-censored data.
3. Interval-censored competing risk data.
4. Clustered interval-censored data.

Thank You!