Abstract

For left-truncated and right-censored (LTRC) data, many researches have been done based on the assumption that the failure and truncation time are independent, which can be unrealistic in application. To take dependence into consideration, we utilize a semiparametric transformation model where the truncation time is both a truncated variable and a predictor of the time to failure. Simulation studies are conducted to investigate finite sample performance of the proposed estimator. An inverse-probability-weighted estimator is proposed for estimate the distribution of left-truncated variable. We also apply our methods to heart transplant survival data.

Key Words: Dependent truncation; Semiparametric transformation model; Inverse probability weighted estimator.

1. Introduction

Left-truncated and right-censored (LTRC) data often arise in epidemiology and individual follow-up studies (see Wang, (1991)). Their importance stems from the common use of prevalent cohort study designs to estimate survival from onset of a specified disease. Consider the following applications.

Example 1: Heart Transplant Data

The study cohorts obtained from heart transplant data (Crowley and Hu (1977)) are commonly recognized as truncated sample, the time-to-failure is truncated by the transplant time. According to the description of Crowley and Hu's paper (1977), the patients agreed to participate in the Stanford program after a medical conference at which it was decided that they were not likely to respond to other forms of therapy. Associated with each patient is a calendar date of acceptance τ_0, a date of transplantation τ_1 and a data last seen $\min(C, \tau_d)$, where τ_d is the calendar time at death and $C = \min(C_1, C_2)$ is the calendar censoring time, where $C_1 = V + d_0$ denotes the time from τ_0 to the end of study and C_2 denotes the time from τ_0 to drop-out. The survival time of interest is $T = \tau_d - \tau_0$. Let $V = \tau_1 - \tau_0$ denote the waiting time before receiving a heart transplant. If we define the population as those patients who consented to receive a heart transplant, the data set of heart recipients is a LTRC sample since the patients must survive long enough to receive a heart transplant, i.e. $T \geq V$. The patients who died before a suitable heart is found are left-truncated. In the Stanford Heart Transplant data set, there were 47 truncated observations (recipients), among which, 30 were dead ($\delta = 1$) and 17 were censored ($\delta_i = 0$). The main purpose was to explore the relationship between certain covariates (e.g. the age of surgery and mismatch sore) and T.

Following the notations in Example 1, let (T, C, V) denote the lifetime, censoring time and truncation time, respectively. Figure 1 highlights all the different times for LTRC data as described in Example 1.

Let $Z = [z_1, \ldots, z_p]^T$ represent a $p \times 1$ vector of covariates. Assume that T, V and C are continuous. Many statistical methods for LTRC data rely on the assumption that T and V are independent, which can be unrealistic in application. There are clinical evidences that a longer transplant waiting time (i.e. a larger value of V) can be worse prognosis of survivorship. When there is no covariate,

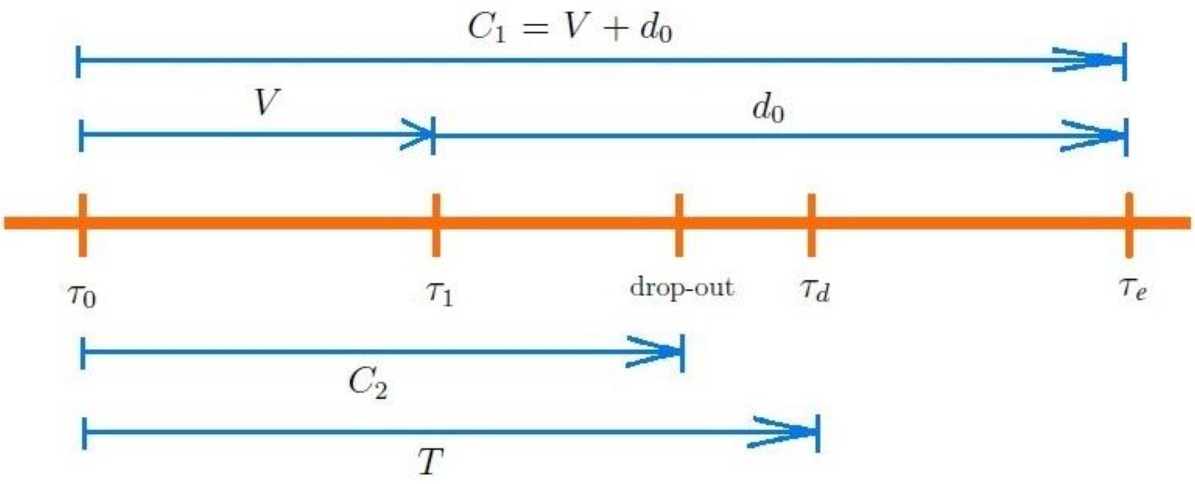

Figure 1. Schematic depiction of LTRC data described in Example 1

Chaieb et al. (2006) provided a nonparametric estimate of distribution function of T using a copula to model the joint distribution of T and V. Furthermore, Emura and Wang (2012) proposed a likelihood-based inference approach and developed a model selection method for choosing the best-fitted copula among a broad class of model alternatives. Under the assumption that C is independent of (V, T) given $V \leq T$, the copula approach can be extended to the case when right censoring is present (Beaudoin and Chaieb and (2008)). However, this assumption usually does not hold since $C_1 = V + d_0$. Mackenzie (2012) considered an alternative approach by modelling the dependence of survival time T on the truncation time V using Cox model (Cox, 1972), i.e. $P(T > t|V) = \exp[-q(V;\gamma)\Lambda_0(t)]$ if $T > V$, where $\Lambda_0(t)$ is the baseline cumulative hazard function and $q(v;\gamma)$ is some continuous function of some unknown regression parameter γ for which $q(v;0) = 1$, e.g. $q(v;\gamma) = \exp(\gamma v)$. When covariate is present, Liu and Zhang (2011) also utilize a Cox analysis with truncation variable V included as a covariate, i.e. $P(T > t|V, Z) = \exp[-\exp(\beta^T Z + \gamma k(V))\Lambda_0(t)]$, where $k(\cdot)$ is a known function.

Cox proportional hazard model is the special case of the well-known class of semiparametric liner transformation model as follows: (see Cheng et al. (1995), Cai et al. (2000) and Chen et al. (2002)):

$$S(t|Z) = g\{h(t) + \beta^T Z\}, \tag{1.1}$$

where $S(t|Z) = P(T > t|Z)$ is the survival function of T given Z, the continuous, strictly decreasing link function $g(\cdot)$ is given or specified up to a finite-dimensional parameter and $h(\cdot)$ is a completely unspecified strictly increasing function. In the case of $g(\cdot) = \exp\{-\exp(\cdot)\}$, (1.1) gives the Cox proportional hazard model and when $g(\cdot) = 1/\{1 + \exp(\cdot)\}$ it gives the proportional odds model (Bennett (1983), Murphy et al. (1997), Ying and Prentice (1991)). Furthermore, model (1.1) has an equivalent form

$$h(T) = -\beta^T Z + \epsilon,$$

where the distribution of the error ϵ is $P(\epsilon \leq x) = F_\epsilon(x) = 1 - g(x)$.

For right-censored data, based on martingale arguments, Chen et al. (2002) proposed an estimation procedure for β and $h(\cdot)$, which is easily implemented numerically and the estimator of β is the same as the Cox partial likelihood estimator in the case of the proportional hazards model. Under the independence of V and T, Shen (2011) extended Chen et al. (2002)'s approach to LTRC data. In Section 2, to take into account dependence between V and T, using the approaches of Liu and Zhang (2011) and Mackenzie (2012), we include the truncation variable V in model (1.1) as a predictor/covariate of the failure time T. The estimating procedure proposed by Chen et al. (2002) is used to obtain estimators for regression coefficients. The distribution function V is estimated using the inverse probability weighted (IPW) approach (Satten and Datta (2001) and Shen (2003)). Based on the IPW estimator, the estimator of $S(t|Z) = P(T > t|Z)$ can be obtained. In Section 3, simulation studies are conducted to investigate finite sample performance of the proposed estimator. In Section 4, we apply our methods to heart transplant survival data.

2. The Proposed Estimators

To take into account the dependence between T and V, suppose that the survival function of T, given the left-truncated variable V and covarites Z follows the semiparametric transformation model:

$$S(t|Z,V) = g\{h(t) + \beta^T Z + \gamma k(V)\} = g\{h(t) + \theta^T \tilde{Z}\}, \tag{2.1}$$

where $\tilde{Z}_i = [Z_i, k(V_i)]^T$ and $\theta = (\beta, \gamma)^T$. Let $F(t|Z) = P(T \le t|Z)$ denote the cumulative distribution function of T given Z. Let $Q(t) = P(C \le t)$ and $G(t) = P(V \le t)$ denote the cumulative distribution functions of C and V, respectively. Suppose that the left and right endpoints of T are independent of Z. Let a_F and b_F denote the left and right endpoints of F, and similarly, define (a_G, b_G) and (a_Q, b_Q) as the left and right endpoint of V, and C, respectively. Throughout this article, for identifiabilities of $F(t|Z)$, we assume that

$$a_G = a_F = a_Q = 0, b_G \le \min(b_F, b_Q) \text{ and } b_F \le b_Q. \tag{2.2}$$

Let $(X_i, V_i, \delta_i, Z_i)$ $(i = 1, \ldots, n)$ be the observed truncated sample. Let $Y_i(t) = I_{[V_i \le t \le X_i]}$ and $N_i(t) = I_{[X_i \le t, \delta_i = 1]}$. Let $\mathcal{F}(t)$ denote the complete σ-field generated by

$$\{V_i, Z_i, Y_i(x), I_{[V_i \le X_i]}, \delta_i I_{[V_i < X_i \le t]}, I_{[V_i < X_i \le x]}, x \le t; i = 1, \ldots, n\}.$$

Let $\lambda_\epsilon(\cdot)$ and $\Lambda_\epsilon(\cdot)$ denote the hazard and cumulative hazard functions of ϵ, respectively. Let $h_0(\cdot)$, β_0 and γ_0 denote the true values of $h(\cdot)$, β and γ, respectively. Let $M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda_\epsilon(\theta_0^T \tilde{Z}_i + h_0(s))$, where $\theta_0 = (\beta_0, \gamma_0)^T$. Since $h(T) = -\theta_0^T \tilde{Z} + \epsilon$, we have

$$S(t|Z) = P(T > t|Z) = P(h(T) > h(t)|Z) = P(\epsilon > h(t) + \theta_0^T \tilde{Z}) = S_\epsilon(h(t) + \theta_0^T \tilde{Z}),$$

where $S_\epsilon(t) = g(t)$ is the survival function of ϵ. Thus, we have $\Lambda(t|Z) = \Lambda_\epsilon(\theta_0^T \tilde{Z}_i + h_0(t))$. Under model (2.1), since

$$E[dN_i(t)|\mathcal{F}(t-)] = Y_i(t) d\Lambda(t|Z) = Y_i(t) d\Lambda_\epsilon(\theta_0^T \tilde{Z}_i + h_0(t)),$$

It follows that $E[dM_i(t)|\mathcal{F}(t-)] = 0$ and $M_i(t)$ is a martingale process with respect to $\mathcal{F}(t)$. Similar to the approach of Chen et al. (2002), we consider the following two estimating equations:

$$U(\beta, h) = \sum_{i=1}^{n} \int_0^{\tau_c} \tilde{Z}_i [dN_i(t) - Y_i(t) d\Lambda_\epsilon(\theta^T \tilde{Z}_i + h(t))] = 0, \tag{2.3}$$

and

$$\sum_{i=1}^{n}[dN_i(t) - Y_i(t)d\Lambda_\epsilon(\theta^T\tilde{Z}_i + h(t))] = 0, \tag{2.4}$$

where h is a nondecreasing function satisfying $h(0) = -\infty$ and $\tau_c < b_F$ is a prespecified constant . This requirement ensures that $\Lambda_\epsilon(a + h(0)) = 0$ for any finite a.

Let $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n)^T$ and $\hat{h}(t;\hat{\theta}_n)$ denote the solution of (2.3) and (2.4). Note that $\hat{h}(t;\hat{\theta}_n)$ is a step function in t that rises at the distinct jump points of $\{I_{[X_i \le t, \delta_i = 1]}; i = 1, \ldots, n\}$. Equations (2.2) and (2.3) suggest the following iterative algorithms for computing $\hat{\theta}_n$ and $\hat{h}(t;\hat{\theta}_n)$:

Step 0: Choose an initial value of θ, denoted by $\hat{\theta}_n^{(0)}$.

Step 1: Let $t_1 < t_2 < \cdots < t_{n_d} < \tau_c$ denote the distinct uncensored points. Obtain $\hat{h}^{(0)}(t_1;\hat{\theta}_n^{(0)})$ by solving

$$\sum_{i=1}^{n} Y_i(t_1)\Lambda_\epsilon(\theta^T\tilde{Z}_i + h(t_1)) = 1,$$

with $\theta = \hat{\theta}_n^{(0)}$. Then, obtain $\hat{h}(t_k)$ for $k = 2, \ldots, n_d$, one-by-one by solving the equation

$$\sum_{i=1}^{n} Y_i(t_k)\Lambda_\epsilon(\theta^T\tilde{Z}_i + h(t_k)) = 1 + \sum_{i=1}^{n} Y_i(t_k)\Lambda_\epsilon(\theta^T\tilde{Z}_i + h(t_k-)),$$

with $\theta = \hat{\theta}_n^{(0)}$.

Step 2: Obtain a new estimate of θ by solving (2.2) with $h(t_k) = \hat{h}^{(0)}(t_k;\hat{\theta}_n^{(0)})$.

Step 3: Set $\hat{\theta}_n^{(0)}$ to be the estimate obtained in Step 2 and repeat Steps 1 and 2 until prescribed convergence criteria are met.

Consider the special case of the Cox model, in which $\lambda_\epsilon = \exp(t)$. It then follows from (2.3) and (2.4) that the estimator $\hat{\theta}_n$ satisfies the following equation:

$$\sum_{i=1}^{n}\int_0^{\tau_c}\left\{\tilde{Z}_i - \frac{\sum_{j=1}^{n}\tilde{Z}_jY_j(t)\exp(\theta^T\tilde{Z}_j)}{\sum_{j=1}^{n}Y_j(t)\exp(\theta^T\tilde{Z}_j)}\right\}dN_i(t) = 0,$$

which is precisely the Cox partial likelihood score equation with truncation variable included as a covariate (see Liu and Zhang (2011)). This can be shown as follows:

Proof:

Since

$$d\Lambda_\epsilon(h(t)+\theta^T\tilde{Z}_i)) = \exp(\theta^T\tilde{Z}_i)d(\exp(h(t)),$$

we have

$$\sum_{i=1}^{n}[dN_i(t) - Y_i(t)\exp(\theta^T\tilde{Z}_i)d(\exp(h(t))] = 0,$$

which implies that

$$d(\exp(h(t)) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} Y_i(t)\exp(\theta^T\tilde{Z}_i)}.$$

By (2.3), we obtain

$$\sum_{i=1}^{n}\int_0^{\tau_c} \tilde{Z}_i[dN_i(t) - Y_i(t)\exp(\theta^T\tilde{Z}_i)d(\exp(h(t))] = 0.$$

Hence,

$$\sum_{i=1}^{n}\int_0^{\tau_c} \tilde{Z}_i\bigg[dN_i(t) - Y_i(t)\exp(\theta^T\tilde{Z}_i)\frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{j=1}^{n} Y_j(t)\exp(\theta^T\tilde{Z}_j)}\bigg] = 0,$$

i.e.

$$\sum_{i=1}^{n}\int_0^{\tau_c}\bigg\{\tilde{Z}_i - \frac{\sum_{j=1}^{n}\tilde{Z}_jY_j(t)\exp(\theta^T\tilde{Z}_j)}{\sum_{j=1}^{n} Y_j(t)\exp(\theta^T\tilde{Z}_j)}\bigg\}dN_i(t) = 0.$$

The proof is complete.

Equations (2.3) and (2.4) can be solved iteratively using a similar algorithm proposed by Chen et al. (2002).

For any vector x, let $x^{\otimes 2} = xx^T$. Similar to Proposition of Chen et al. (2002), under suitable regularity conditions, we have the following proposition.

Theorem 1. Under assumption (2.2) and regularity conditions (Fleming and Harrington, 1991), we have that $n^{\frac{1}{2}}(\hat{\theta}_n - \theta) \to N(0, \Sigma_{\hat{\theta}_n})$ in distribution, as $n \to \infty$, where $\Sigma_{\hat{\theta}_n} = \Sigma_2^{-1}\Sigma_1(\Sigma_2^{-1})^T$

$$\Sigma_1 = E\bigg[\int_0^{\tau_c}[\tilde{Z}_1 - \mu_z(t;\theta_0)]^{\otimes 2}\lambda_\epsilon(h_0(t)+\theta_0^T\tilde{Z}_1)Y_1(t)]dh_0(t)\bigg],$$

$$\Sigma_2 = E\bigg[\int_0^{\tau_c}[\tilde{Z}_1 - \mu_z(t;\beta_0)]\tilde{Z}_1^T\dot{\lambda}_\epsilon(h_0(t)+\theta_0^T\tilde{Z}_1)Y_1(t)]dh_0(t)\bigg],$$

where

$$\mu_z(t) = \frac{E[\tilde{Z}_1 \lambda_\epsilon(h_0(X_1) + \theta_0^T \tilde{Z}_1) Y_1(t) B(t; X_1)]}{E[\lambda_\epsilon(h_0(t) + \theta_0^T \tilde{Z}_1) Y_1(t)]},$$

where

$$B(t, s) = \exp\biggl(\int_s^t \frac{E[\dot{\lambda}_\epsilon(h_0(x) + \theta_0^T \tilde{Z}_1) Y_1(x)]}{E[\lambda_\epsilon(h_0(x) + \theta_0^T \tilde{Z}_1) Y_1(x)} dh_0(x) \biggr).$$

Proof: The proof is similar to Appendix of Chen et al. (2002) and is omitted.

Note that Σ_1 and Σ_2 can be consistently estimated by

$$\hat{\Sigma}_1 = n^{-1} \sum_{i=1}^n \int_0^{\tau_c} [\tilde{Z}_i - \bar{Z}(t; \hat{\theta}_n)]^{\otimes 2} \lambda_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(t; \hat{\theta}_n)) Y_i(t) d\hat{h}(t; \hat{\theta}_n),$$

and

$$\hat{\Sigma}_2 = n^{-1} \sum_{i=1}^n \int_0^{\tau_c} [\tilde{Z}_i - \bar{Z}(t; \hat{\theta}_n)] \tilde{Z}_i^T \dot{\lambda}_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(t; \hat{\theta}_n)) Y_i(t) d\hat{h}(t; \hat{\theta}_n),$$

respectively, where $\dot{\lambda}_\epsilon(x) = d\lambda_\epsilon(x)/dx$,

$$\bar{Z}(t; \hat{\theta}_n) = \sum_{i=1}^n \frac{\tilde{Z}_i \lambda_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(t; \hat{\theta}_n)) Y_i(t) \hat{B}(t, X_i)}{\sum_{i=1}^n \lambda_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(t; \hat{\theta}_n)) Y_i(t)},$$

$$\hat{B}(t, s) = \exp\biggl(\int_s^t \frac{\sum_{i=1}^n \dot{\lambda}_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(x; \hat{\theta}_n)) Y_i(x)}{\sum_{i=1}^n \lambda_\epsilon(\hat{\theta}_n^T \tilde{Z}_i + \hat{h}(x; \hat{\theta}_n)) Y_i(x)} d\hat{h}(x; \hat{\theta}_n) \biggr).$$

Hence, a consistent estimator of $\Sigma_{\hat{\theta}_n}$ is given by $\hat{\Sigma}_{\hat{\theta}_n} = \hat{\Sigma}_2^{-1} \hat{\Sigma}_1 (\hat{\Sigma}_2^{-1})^T$.

The survival function of T given Z is given by

$$S(t|Z) = \int_{a_G}^{b_G} S(t|Z, v) G(dv) = \int_0^{b_G} g\{h(t) + \beta^T Z + \gamma k(v)\} G(dv).$$

For heart transplant data, the left-truncated variable is the transplant waiting time, which is determined by the donor searching process. The information about the amount of time spent on the donor searching is very important in efficiently allocating resources to assist patients in finding donors. Next, we discuss the estimator of the distribution of left-truncated variable, i.e. $G(v)$. Let $p = P(V \leq T)$. Under model (2.1), we have

$$E\biggl[\frac{I_{[V_i \leq t]}}{g\{h(V_i) + \beta^T Z_i + \gamma k(V_i)\}} \biggr] = \int_z \int_{v \leq t} p^{-1} \frac{S(v|v, z)}{g\{h(v) + \beta^T z + \gamma k(v)\}} dG(v) dA(z) = p^{-1} G(t),$$

where $A(z)$ is the cumulative distribution function of Z. Let $t \to \infty$. It follows that $E[1/(g\{h(V_i) + \beta^T Z_i + \gamma k(V_i)\})] = p^{-1}$. Hence, p can be estimated by

$$\hat{p}(\hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)) = n\left[\sum_{i=1}^{n} \frac{1}{g\{\hat{h}(V_i; \hat{\theta}_n) + \hat{\beta}_n^T Z_i + \hat{\gamma}_n V_i\}}\right]^{-1},$$

and $G(t)$ can be estimated by

$$\hat{G}(t; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)) = \hat{p}(\hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n))n^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le t]}}{g\{\hat{h}(V_i; \hat{\theta}_n) + \hat{\beta}_n^T Z_i + \hat{\gamma}_n k(V_i)\}}.$$

Thus, $S(t|Z_i)$ can be estimated by

$$\hat{S}(t; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)|Z_i) = \int_0^{b_G} g\{\hat{h}(t; \hat{\theta}_n) + \hat{\beta}_n^T Z_i + \hat{\gamma}_n v\}\hat{G}(dv; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)).$$

The asymptotic normality of $n^{1/2}[\hat{G}(t; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)) - G(t)]$ can be established by the following expression:

$$\begin{aligned} n^{1/2}[\hat{G}(t; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)) - G(t)] &= n^{1/2}[\hat{G}(t; \hat{\theta}_n; \hat{h}(\cdot; \hat{\theta}_n)) - n^{1/2}[\hat{G}(t; \theta_0; h_0(\cdot))] \\ &\quad + n^{1/2}[\hat{G}(t; \theta_0; h_0(\cdot)) - G(t)]. \end{aligned}$$

3. Simulation Studies

Case 1: Proportional Odds Model

We generated T following the proportional odds model with $h(t) = \log(t/10)$ and T has the survivorship function

$$P(T > t|V, Z) = \frac{1}{1 + \exp\{\log(t/10) + V + z_1\}},$$

where z_1 is a Bernoulli random variable with probability 0.5. Hence, $\theta_0 = (\beta_0 = 1, \gamma_0 = 1)^T$. Note that under this set-up, the p^{th} percentile of T at (V, z_1) is $t_p = 10\exp\{\log((1-p)/p) - (V + z_1)\}$, which decreases as V or z_1 increases. The left-truncation variable V was generated from exponential distribution with mean $\theta_g = 0.5, 1.0, 2.0$, and right censoring variable C was generated from $D + V$, where D is exponentially distributed with mean $\theta_d = 10, 100$. Sample size is set at $n = 200, 400$. The replication time is 1000. The values of τ_a and τ_b are set at the smallest and largest values of X_i's, respectively. For each simulated dataset, we obtained the estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n)^T$. Using $\hat{\Sigma}_{\hat{\theta}_n}$, we also calculated the estimated standard deviations of $\hat{\beta}_n$ and $\hat{\gamma}_n$. Table 1 shows the simulated biases, simulated standard deviations (std), and the estimated standard deviation (estd). Table 1 also shows the proportion of left-truncation $P(T \leq V)$ (denoted by q) and right-censoring (denoted by $p_c = P(\delta_i = 0)$).

Case 2: Cox Model

We generated T following the proportional hazards model with $h(t) = t$ and $\theta_0 = (\beta_0 = -1, \gamma_0 = -0.5)^T$. The resulting T has the survivorship function

$$P(T > t|V, z_1) = e^{-e^{t - z_1 - 0.5V}},$$

where z_1 is a Bernoulli random variable with probability 0.5 equal to 5 and 10. Note that under this set-up, the p^{th} percentile of T at (z_1, V) is $t_p = \log(-\log p) + z_1 + 0.5V$, which is a linear function of z_1 and V, and increases as z_1 or V increases. The V and C are generated as described in Case 1. The values of (θ_g, θ_d) are set at (6,7.5), (6,28), (10,6.8), (10,25), (16,6) and (16,22). Sample size is set at $n = 200, 400$. The replication time is 1000. Table 2 shows the simulation results of the estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$.

Table 1 (Case 1). Simulated biases, std and estd of $\hat{\beta}_n$ and $\hat{\gamma}_n$

						$\hat{\beta}_n$		$\hat{\gamma}_n$		
θ_g	θ_d	q	p_c	n	bias	std	estd	bias	std	estd
0.5	10	0.14	0.46	200	-0.015	0.372	0.341	-0.068	0.617	0.566
0.5	10	0.14	0.46	400	-0.010	0.219	0.200	-0.014	0.384	0.357
0.5	100	0.14	0.13	200	-0.017	0.353	0.326	-0.022	0.566	0.524
0.5	100	0.14	0.13	400	0.008	0.202	0.193	0.015	0.357	0.338
1.0	10	0.30	0.44	200	-0.061	0.518	0.518	-0.125	0.744	0.744
1.0	10	0.30	0.44	400	-0.035	0.275	0.251	-0.067	0.495	0.450
1.0	100	0.30	0.13	200	0.046	0.443	0.443	0.086	0.726	0.675
1.0	100	0.30	0.13	400	0.023	0.258	0.245	0.007	0.472	0.453
2.0	10	0.50	0.43	200	-0.045	0.370	0.338	-0.132	0.765	0.697
2.0	10	0.50	0.43	400	-0.035	0.275	0.262	-0.067	0.525	0.493
2.0	100	0.50	0.12	200	-0.043	0.343	0.318	0.082	0.746	0.684
2.0	100	0.50	0.12	400	0.007	0.329	0.313	0.023	0.517	0.489

Table 2 (Case 2). Simulated biases, std and estd of $\hat{\beta}_n$ and $\hat{\gamma}_n$

						$\hat{\beta}_n$		$\hat{\gamma}_n$		
θ	θ_d	q	p_c	n	bias	std	estd	bias	std	estd
6	7.5	0.15	0.44	200	-0.042	0.379	0.350	-0.032	0.226	0.208
6	7.5	0.15	0.44	400	-0.024	0.272	0.258	-0.018	0.181	0.173
6	28	0.15	0.16	200	-0.010	0.360	0.334	-0.012	0.214	0.187
6	28	0.15	0.16	400	-0.004	0.252	0.238	-0.005	0.175	0.164
10	6.8	0.30	0.44	200	-0.041	0.396	0.364	-0.032	0.252	0.227
10	6.8	0.30	0.44	400	-0.018	0.274	0.262	-0.011	0.166	0.154
10	25	0.30	0.16	200	-0.046	0.417	0.382	-0.030	0.240	0.217
10	25	0.30	0.16	400	0.019	0.259	0.244	-0.012	0.149	0.140
16	6	0.45	0.44	200	-0.066	0.437	0.407	-0.053	0.285	0.249
16	6	0.45	0.44	400	-0.025	0.273	0.254	-0.019	0.172	0.164
16	22	0.45	0.16	200	-0.049	0.428	0.394	-0.035	0.276	0.253
16	22	0.45	0.16	400	0.011	0.269	0.256	-0.013	0.163	0.154

Based on the results of Tables 1 and 2, we have the following conclusions:

(1) Given proportion of left-truncation q, the standard deviations of both $\hat{\beta}_n$ and $\hat{\gamma}_n$ increase as proportion of right-censoring (p_c) increases. Given proportion of right censoring p_c, the standard deviations of both $\hat{\beta}_n$ and $\hat{\gamma}_n$ increase as proportion of left-truncation increases. When both truncation and censoring are heavy (i.e. $q = 0.50; p_c = 0.43$ for case 1 and $q = 0.45; p_c = 0.44$ for case 2), the biases of $\hat{\gamma}_n$ can be large.

(2) When $n = 200$, the estimated standard deviation underestimates the empirical standard deviation. However, when $n = 400$, the estimated standard deviation is close to the empirical standard deviation for most of the cases considered.

4. Analysis of Heart Transplant Data

The proposed estimators were applied to the Heart Transplant Data (Crowley and Hu (1977)) described in Example 2. The main purpose was to explore the relationship between certain covariates and the cause of death due to transplant rejection. We consider proportional odds and Cox regression analysis for all the patients and the patients over 45 years old. The age of surgery (z_1), mismatch score (z_2) and left-truncated variable V are included in the model (2.1) with $k(v) = v$. Table 3 lists the estimated parameters $\hat{\beta}_{1n}$, $\hat{\beta}_{2n}$ and $\hat{\gamma}_n$ for age, mismatch score and V, respectively.

Table 3. The estimated parameters $\hat{\beta}_n$ and $\hat{\gamma}_n$

all the patients	(proportional odds model)	
$\hat{\beta}_{1n}$(p-value)	$\hat{\beta}_{2n}$(p-value)	$\hat{\gamma}_n$ (p-value)
2.231(0.106)	0.079(0.0013)	0.141(0.224)
patients over 45	(proportional odds model)	
$\hat{\beta}_{1n}$(p-value)	$\hat{\beta}_{2n}$(p-value)	$\hat{\gamma}_n$ (p-value)
1.758(0.104)	0.264(0.011)	0.471(0.077)
all the patients	(Cox regression model)	
$\hat{\beta}_{1n}$(p-value)	$\hat{\beta}_{2n}$(p-value)	$\hat{\gamma}_n$ (p-value)
0.045(0.093)	0.025(0.0002)	0.006(0.258)
patients over 45	(Cox regression model)	
$\hat{\beta}_{1n}$(p-value)	$\hat{\beta}_{2n}$(p-value)	$\hat{\gamma}_n$ (p-value)
0.087(0.082)	0.028(0.003)	0.009(0.108)

5. Conclusion

In this article, to take dependence into account, the truncation variable V is included in transformation model as a predictor/covariate of the failure time T. Using the approach of Chen et al. (2002), we obtain estimators for regression coefficients. Furthermore, we also propose an inverse-probability-weighted estimator to estimate the distribution of left-truncated variable, i.e. the distribution of transplant waiting time. Simulation results indicate that the proposed estimators perform adequately. Further research is required to develop goodness-of-fit tests for the proposed model.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Methods Based on Counting Processes.* New York: Springer.

Beaudoin, D. and Chaieb L. L. (2008). Archimedean copula model selection under dependent truncation. *Statis. Med.*, **27**, 4440-4454.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Med.*, **2**, 273-277.

Cai, T., Wei, L. J. and Wicox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika*, **87**, 867-878.

Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika*, **91**, 277-290.

Chaieb, L. L., Rivest, L.-P. and Abdous, B. (2006), Estimating survival under a dependent truncation. *Biometrika*, 93, 655-669.

Chen, K., Jin, Z and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659-668.

Cox, D. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* **B**, **34**, 187-220.

The CSHA working group (1994). Canadian study of health and aging: study methods and prevalence of dementia. *J. Can. Med. Asso.*, **150**, 899-913.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Amer. Statist. Assoc.*, **94**, 496-509.

Fleming, T. R. and Harrington, D. P. (1991). Counting Processes and Survival Analysis. New York: Wiley.

Hyde, J., (1980). Survival analysis with incomplete observations. In Biostatistics Casebook, R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses, eds, New York: John Wiley and Sons, pp. 31-46.

Emura, T. and Wang, W. (2012). Nonparametric maximum likelihood estimation for dependent truncation data based on copulas, *J. Multi. Anal.*, **110**, 171-188.

Klein, J. P. and Moeschberger, M. L. (1997). Survival analysis: Tenchniques for censored and truncated data. Springer.

Lai, T. L. and Ying, Z. (1991), Estimating a distribution function with truncated and censored data. The Annal of Statistics, 19, 417-422.

Liu, Y. and Zhang, X. (2011). Analysis of dependently truncated sample using inverse probability weighted estimator. George State University, Department of Mathematics and Statistics, Mathematical Thesis, 8-1-2011.

Mackenzie, T. (2012). Survival curve estimating with dependent left truncation data using Cox model. *Inter. J. of Biostat.*, **8**, 1-18.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1996). Maximum likelihood estimation in the proportional odds model. *J. Am. Statist. Assoc.*, **92**, 968-976.

Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics*, **58**, 64-70.

Satten, G. A. and Datta S (2001), The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist. Ass.*, **55**, 207-210.

Shen, P. S. (2003), The product-limit estimate as an inverse-probability-weighted average *Communi. in Statist. Theory and Methods*, **32**, No 6, 1119-1133.

Shen, P. S. (2011). Semiparametric analysis of transformation models with left-truncated and right-censored data. *Comp. Stat.*, **26**, 521-537.

Tsai, W.-Y., Jewell, N. P., Wang, M.-C. (1987). A note on the product-limit estimate under right censoring and left truncation. *Biometrika*, **74**, 883-886.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *J. Amer. Statist. Ass.*, **84**, 742-748.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Amer. Statist. Ass.*, **86**, 130-143.

Ying, S and Prentice, R. (1999). Semiparametric inference in the proportional odds regression model. *J. Am. Statist. Assoc.*, **92**, 968-976.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression model with censored data. *J. R. Statist. Soc. B*, **69**, 507-564.