应用概率统计　第 37 卷
第 6 期　2021 年 12 月

Chinese Journal of Applied Probability and Statistics
Dec., 2021, Vol. 37, No. 6, pp. 627-654
doi: 10.3969/j.issn.1001-4268.2021.06.006

Survey

Statistical Analysis of Interval-Censored Failure Time Data

DU Mingyue

(*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China*)

SUN Jianguo*

(*Department of Statistics, University of Missouri, Columbia, MO 65211, USA*)

Abstract: Interval-censored failure time data are a general type of failure time or time-to-event data where the failure time of interest is known or observed only to lie in an interval or window instead of being observed exactly. They often occur in many fields, including demographical studies, epidemiological studies, medical or public health research and social science, and in different forms. A common and general set-up that naturally yields interval-censored data is the study with longitudinal or periodical follow-ups such as many clinical trials or observation studies. In this paper, after some brief discussion about the background and some commonly used models, we will review some recent advances, mainly during about last five years, on several important topics related to regression analysis as well as some issues that need more research in the analysis of interval-censored data.

Keywords: censoring mechanism; maximum likelihood estimation; observational studies; periodical follow-up; regression analysis

2020 Mathematics Subject Classification: 62N01

Citation: DU M Y, SUN J G. Statistical analysis of interval-censored failure time data [J]. Chinese J Appl Probab Statist, 2021, 37(6): 627–654.

§1. Introduction

Interval-censored failure time data are a general type of failure time or time-to-event data where the failure time of interest is known or observed only to lie in an interval or window instead of being observed exactly. They often occur in many fields, including demographical studies, epidemiological studies, medical or public health research and social science, and in different forms. Although a large literature, including four books [1–4] and several review papers [5–7], has been established for the analysis of interval-censored

*Corresponding author, E-mail: sunj@missouri.edu.

Received January 12, 2021. Revised February 20, 2021.

data, there still exist many open questions or more research is needed for many existing or new issues. The main purpose of this paper is to discuss some recent advances on several important topics related to regression analysis of interval-censored data but not to provide a comprehensive review of the recent literature.

One area that routinely yields interval-censored failure time data is medical or public health studies that entail periodic follow-ups such as clinical trials. In these situations, an individual due for pre-scheduled observations for a clinically observable change in disease or health status may miss some observations and return with a changed status. As a consequence, we only know that the true event time is greater than the last observation time at which the change has not occurred and less than or equal to the first observation time at which the change has been observed to occur. That is, we only have an interval that contains the real (but unobserved) time of occurrence of the change. Note that for the situation described above, one still observes interval–censored data even no study subject misses any pre-specified observation time or all individuals follow exactly the pre-specified visit schedule, but it is apparent that the observed data would have simple and balanced structures. On the other hand, in reality, this is clearly usually not the case and the resulting data, the focus of this paper, can have much more complicated structures.

A more specific example of interval-censored failure time data is given by Alzheimer's disease neuroimaging initiative (ADNI), a longitudinal follow-up study that started in 2004 and was designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of the Alzheimer's disease (AD) [8–10]. In the study, the participants were recruited across North America and followed and reassessed periodically to track the pathology of the disease as it progresses. Also the participants have been divided into three groups based on the levels of their cognitive conditions, cognitively normal, mild cognitive impairment and AD. Among others, one variable of interest is the time from the baseline visit date to the AD conversion. Since the participants were only examined intermittently, the AD conversion thus cannot be observed exactly and is known only to between the last examination time when the AD had not occurred and the first examination time when the AD has already occurred. In other words, we only have interval-censored data on the AD conversion.

Interval-censored failure time data occur in or can have different forms, including case I, case II and case K interval-censored data which will be described in details in the next section, and different forms mean different structures of the data. In particular, they include right-censored data, the type of the failure time data discussed most in the literature [11], as a special case. It is worth noting that the analysis of interval-censored data is quite different from and much more challenging than that of right-censored data. One such difference is that for the latter, the counting process approach could be easily adopted, which makes the analysis much easier, while this is not true for the former. A more specific difference can be seen from their regression analyses under the proportional hazards model to the described in the next section. For the latter, a simple partial likelihood function could be conveniently derived and commonly used for inference about regression parameters, while for the former, a more complicated full likelihood function has to be used in general.

One fundamental and important feature of failure time data is censoring and different formations of the data correspond to different censoring structures. In reality, one can classify censoring as either independent censoring, or dependent or informative censoring, meaning that the failure time of interest and the censoring mechanism are correlated [12, 13]. With the former, the analysis is usually performed conditional on the censoring process no matter the formats of the data. In contrast, with the latter, the analysis can be very different and also difficult as one usually has to make certain assumptions or model the censoring mechanism. In particular, for right-censored data, the modeling is relatively easy partly as only one variable is needed to describe the censoring, while for interval-censored data, as discussed below, two or more variables are usually required to characterize the censoring mechanism. As pointed out in the literature [3], in the presence of informative censoring, the analysis that ignores it may result in biased results or misleading conclusions.

The remainder of the paper is organized as follows. After a brief discussion of different data structures and some commonly used models in Section 2, we will first discuss some recent advances on regression analysis of univariate interval-censored failure time data with time-dependent covariates or in the presence of a cured subgroup in Section 3. Section 4 will also consider univariate interval-censored data as in Section 3 but with dependent or informative censoring, and Section 5 will discuss some recent advances on regression analysis of multivariate and clustered interval-censored data. Case-cohort studies are commonly used to reduce the cost on the collection of covariate information and Section 6 will review some recently developed analysis methods for them that yield interval-censored data. Variable selection has recently attracted a great deal of attention and will be the focus in Section 7 when one faces interval-censored data. At the end of each section, some directions for future research will be discussed, and Section 8 will conclude with some remarks and a brief discussion on several other topics, including the analyses of truncated and/or doubly interval-censored data and the analysis of interval-censored data with missing covariates. As mentioned above, throughout the review, we will mainly focus on the advances during about last five years and one can find more references from the cited references. Also except Section 4 or unless specifically stated, we will assume that interval censoring is non-informative.

§2. Models and Formulations of Interval-Censored Data

In this section, we will first briefly describe several commonly used regression models for the analysis of failure time data, including the Cox or proportional hazards model, the additive hazards model and the linear transformation model. Then the data structures will be discussed for the types of interval-censored data commonly seen in the literature.

Consider a failure time study and let T and X denote the failure time of interest and a vector of covariates, respectively. In this section, for simplicity, we will assume that X is time-independent. For the analysis of failure time data, without doubt, the Cox proportional hazards model[14] is the most commonly used one and has the form

$$\lambda(t \,|\, X) = \lambda_0(t)\exp(\beta' X) \tag{1}$$

with respect to the hazard function of T given X. In the above, $\lambda_0(t)$ denotes an unknown baseline hazard function and β the vector of unknown regression parameters. A main advantage of this model is its simplicity and the availability of a partial likelihood function for right-censored data. On the other hand, as pointed out in the literature, it has some disadvantages including the proportionality assumption. Also it is worth noting that the partial likelihood approach is no longer available for regression analysis of interval-censored data.

Another commonly used model for regression analysis of failure time data is the additive hazards model given by

$$\lambda(t \,|\, X) = \lambda_0(t) + \beta' X \tag{2}$$

again in terms of the hazard function of T given X. Here $\lambda_0(t)$ and β are defined as in model (1). Note that unlike model (1) which models the multiplicative effects of covariates, model (2) considers the additive effects, which can be more interesting in the fields like social science. In the case that one is more interested in the covariate effect on the cumulative distribution function (CDF), one may consider the proportional odds model given by

$$\ln\left[\frac{F(t \,|\, X)}{1 - F(t \,|\, X)}\right] = h(t) + \beta' X. \tag{3}$$

In the above, $F(t \,|\, X)$ denotes the CDF of the failure time T given covariates X, $h(t)$ an unknown monotone-increasing function, also referred to as the baseline log odds, and β represents a vector of regression parameters as in models (1) and (2).

It is apparent that the three models (1) – (3) described above are all specific models in terms of the functional form of the effects of covariates. Sometimes one may prefer a model that gives more flexibility, and one such model is the linear transformation model that can be expressed as

$$\Lambda(t \,|\, X) = G(\Lambda_0(t) \exp(\beta' X)) \tag{4}$$

in terms of the cumulative hazard function of T given X. Here G is a specific, strictly increasing transformation function, $\Lambda_0(t)$ an unknown increasing function and β defined as above. It is easy to see that the model above gives different models depending on the specification of the transformation function G. For example, the choices of $G(x) = x$ and $G(x) = \ln(1+x)$ give the proportional hazards model (1) and the proportional odds model (3), respectively.

A main advantage of linear transformation model (4) is their flexibility as they include many commonly used models as special cases as described above. Also it is well-known that sometimes an individual model such as the proportional hazards model may not fit data well and in contrast, the class of model (4) can allow for different types of covariate effects. In addition, model (4) can be rewritten as

$$\ln \Lambda_0(T) = -\beta' X + \varepsilon, \tag{5}$$

where ε denotes an error term with the distribution function $1-\exp[-G(\exp(x))]$. Thus the regression parameter β can also be interpreted as the covariate effect on a transformation of the failure time T.

In reality, there usually exist several types of interval-censored failure time data or several formulations are commonly used in the literature [3]. Among them, an important type is case I interval-censored data, also often referred to as current status data, meaning that each subject is observed only once for the occurrence of the failure event of interest. In consequence, the failure time T is either left- or right-censored and the observation on a study subject has the form $\{C,\, I(T \leqslant C)\}$, where C represents the observation time. One type of studies that usually produce current status data is cross-sectional studies, which are commonly used in, for example, demographical studies among others.

Corresponding to case I interval-censored data, another formulation for interval-censored data that is often seen in the literature is case II interval-censored data, which assume that there exist two observation times for each study subject. For the situation, the observation has the form $\{U,\, V,\, \delta_1 = I(T < U),\, \delta_2 = I(U \leqslant T < V)\}$ with $U < V$, where U and V denote the two observation times. A more general formulation or type of interval-censored data are case K interval-censored data, meaning that there exists a sequence of observation times for each subject. For the case, the data have the form $\{K,\, U_0 < U_1 < \cdots < U_K,\, \delta_k = I(U_{j-1} < T \leqslant U_j);\, j = 1, 2, \cdots, K\}$, where K denotes the number of observation times with the U_j's being the observation times. In practice, both K and the U_j's can be subject-dependent, and it is easy to see that many observation schemes such as these commonly used in medical follow-up or longitudinal studies can be naturally represented by this formulation.

The formulation that is used most to describe interval-censored data in practice is perhaps $I = (L, R]$ with $T \in I$, which will be referred to as general interval-censored data below. Under this formulation, it is easy to see that case I interval-censored data correspond to the situation where either $L = 0$ or $R = \infty$, while right-censored data mean either $L = R$ or $R = \infty$ for all study subjects. Also it is apparent that both case II and case K interval-censored data can be reduced to this format as often happened in reality.

§3. Analysis of Univariate Interval-Censored Data

In this section, we will discuss some recent advances on regression analysis of univariate interval-censored failure time data with the focus on the situations with the existence of time-dependent covariates or a cured subgroup. To distinguish the difference between time-independent and time-dependent covariates, we will write the time-dependent covariate as $X(t)$ in the following.

1) Analysis with Time-Dependent Covariates

In the presence of time-dependent covariates, instead of modeling the covariate effect on the hazard function like models (1) and (2), it is usually more convenient to model the covariate effect on the cumulative hazard function. Also instead of model (4), one may consider the following form

$$\Lambda(t \,|\, X) = G\Big(\int_0^t \exp[\beta' X(s)]\mathrm{d}\Lambda_0(s)\Big) \tag{6}$$

for the transformation model. Under the formulation above, a commonly used class of frailty-induced transformations is

$$G(x) = -\ln\Big[\int_0^\infty \exp(-xt) f(t)\mathrm{d}t\Big], \tag{7}$$

where $f(t)$ is the density function of a frailty variable with the support $[0, \infty)$. By setting $f(t)$ to be the gamma density with mean 1 and variance γ, one will get $G(x) =$

$\gamma^{-1}\ln(1+\gamma x)$, the class of logarithmic transformations, and it will give the class of Box-Cox transformations $G(x) = [(1+x)^{\gamma} - 1]/\gamma$ if letting $f(t)$ to be the positive stable distribution with the parameter $\gamma < 1$. It is easy to see that when covariates are time-independent, model (6) reduces to models (4) and (5).

For estimation of the effects of time-dependent covariates, two approaches are commonly used. One is the marginal maximum likelihood approach and the other is the joint modeling approach. To be more specific about them, consider a failure time study that consists of n independent subjects and in which for each subject, there exists a sequence of observation times denoted by $U_{i0} = 0 < U_{i1} < U_{i2} < \cdots < U_{iK_i}$. Also let the T_i's denote the failure times of interest and assume that the observed data have the form $O = \{O_i = (K_i,\, U_{ij},\, X_i(t)I(t \leqslant U_{iK_i}),\, \delta_{ij} = I(U_{i,j-1} < T_i \leqslant U_{ij}));\, j = 1, 2, \cdots, K_i,\, i = 1, 2, \cdots, n\}$. That is, we have case K interval-censored data. Furthermore, assume that the T_i's follow the transformation model (6). Then under the non-informative censoring assumption, meaning that $\{K_i's, U_{ij}'s\}$ are independent of the T_i's given the covariates, the observed likelihood function of β and Λ_0 has the form

$$L(\beta,\Lambda_0) = \prod_{i=0}^{n}\prod_{j=0}^{K_i}\Big\{\exp\Big[-G\Big(\int_0^{U_{ij}} \mathrm{e}^{\beta' X_i(s)}\mathrm{d}\Lambda_0(s)\Big)\Big] - \exp\Big[-G\Big(\int_0^{U_{i,j-1}} \mathrm{e}^{\beta' X_i(s)}\mathrm{d}\Lambda_0(s)\Big)\Big]\Big\}^{\delta_{ij}}. \tag{8}$$

For each i, let $L_i = \max\{U_{ij} : U_{ij} < T_i\}$ and $R_i = \min\{U_{ij} : U_{ij} \geqslant T_i\}$. Then $(L_i, R_i]$ represents the smallest interval that brackets T_i and the likelihood function above can be rewritten as

$$L(\beta,\Lambda_0) = \prod_{i=0}^{n}\Big\{\exp\Big[-G\Big(\int_0^{L_i} \mathrm{e}^{\beta' X_i(s)}\mathrm{d}\Lambda_0(s)\Big)\Big] - \exp\Big[-G\Big(\int_0^{R_i} \mathrm{e}^{\beta' X_i(s)}\mathrm{d}\Lambda_0(s)\Big)\Big]\Big\} \tag{9}$$

since for each i, only one δ_{ij} equals to one with the others being zero. For estimation of β and Λ_0, the marginal approach maximizes the likelihood function above, and in particular, Zeng et al.[15] investigated this approach with the use of the transformation given in (7). Furthermore, they showed that the maximum likelihood estimators of regression parameters are consistent and asymptotically efficient and normal. Also they developed a flexible and computationally efficient EM algorithm following Wang et al.[12], who considered the same problem but under model (1) with time-independent covariates.

The joint modeling approach treats the time-dependent covariates as longitudinal processes and is usually preferred when there also may exist measurement errors on covariates. For this, one commonly used method is to model the failure time of interest and the longitudinal covariate process jointly by using, for example, the latent variable approach and for the situation, under the Cox model, one may consider

$$\Lambda(t \mid X) = \int_0^t \exp[\beta' X(s) + B(s)] \mathrm{d}\Lambda_0(s).$$

In the above, $B(t)$ represents a function of some latent variables as well as the covariates that can be observed both exactly and with measurement errors. Among others, Yi et al.[16] developed a maximum likelihood estimation procedure under this framework and provided a MCEM algorithm. Note that many methods have been developed in the literature for joint analysis of longitudinal data and failure time data with either the failure time or the longitudinal variable as the variable of interest. However, most of them only focused on right-censored data on the failure time except Chen et al.[17]. One major difference between the methods given in Chen et al.[17] and Yi et al.[16] is that the former treats the failure time as the dropout or stopping variable and assumed that there is no more observation after the dropout. In other words, it cannot give efficient or valid estimation if there exist more observations after the failure time. In contrast, the latter takes into account all observations and also the algorithm given in Yi et al.[16] is more faster and stable than that given in Chen et al.[17].

In practice, similar to time-dependent covariates, it is apparent that covariate effects or regression coefficients could be time-dependent or time-varying too. For the situation, one could consider the models described above with replacing β by $\beta(t)$, meaning that the covariate effect may be different at different times. One example of such situations is that the effect of a treatment on a disease may take some time to kick in and then disappear after some time. Although many authors have discussed the situation for the analysis of right-censored data, there exists little research for the analysis of interval-censored data and more research on this is clearly needed. The fitting of interval-censored data to crossing hazard models is another topic for which there does not seem to exist much research except Zhang et al.[18].

2) Analysis in the Presence of a Cured Subgroup

In the standard or traditional failure time data analysis, a typical assumption is that all subjects under study would eventually experience the failure event of interest if the follow-up is sufficiently long. However, in some situations or reality, this assumption may not hold as there may exist a portion of subjects who never experience or are non-susceptible to the failure event of interest for various reasons. These individuals are usually considered to be cured or immune from the failure event and referred to as long-term survivors or cured subjects. To address the existence of a cured subgroup, two types of models or methods are commonly used, two-component mixture cure model approach and non-mixture cure model approach. The former directly models the effects of covariates on the cure rate of the population and the survival function of non-cured subjects through two separate regression models, and a drawback of this is that it does not have the usual survival model property for the whole population [19, 20]. In contrast, the latter assumes that cured subjects have infinity survival time and uses a single model to describe the survival function of the entire population [19]. Sometimes the latter model is also referred to as the promotion time cure model [21].

Under the two-component mixture cure model, the failure time of interest T is usually written as $T = YT^* + (1-Y)\infty$, where T^* denotes the failure time of a susceptible subject and Y indicates, by value 1 or 0, whether the study subject is susceptible or not. To describe the effects of covariates, one could employ a regular failure time regression model such as the model described above. For the effects of of covariates on the cure rate, the following logistic model

$$p(X) = \mathsf{P}(Y = 1 \,|\, X) = \frac{\exp(\alpha_0 + \alpha' X)}{1 + \exp(\alpha_0 + \alpha' X)} \tag{10}$$

is commonly used, where α_0 and α are unknown parameters. Note that sometimes the covariates that affect the failure time of a susceptible subject and the cure rate may be different and thus one may consider or use different covariates in the two models. Among others, Ma [20] and Hu and Xiang [19] discussed this approach for the analysis of interval-censored data. The former developed the maximum likelihood approach under models (1) and (10) and the latter proposed a sieve maximum likelihood method under a model similar to model (4) and model (10).

As mentioned above, instead of employing two separate two models in the mixture cure model, the non-mixture cure model uses a single model to describe the survival function of the entire population. Under the framework of the Cox model (1), it has the form

$$S(t \,|\, X) = \mathsf{P}(T \geqslant t) = \exp\big[- F(t) \mathrm{e}^{\beta_0 + \beta' X} \big], \tag{11}$$

where β_0 and β are unknown coefficients and F is a completely unspecified cumulative distribution function. It is easy to see that the model above inherits the proportional hazards model structure for the whole population and thus regression parameters have relatively appealing, easy interpretations. Also its value at infinity, $\exp[-\exp(\beta_0 + \beta' X)]$, represents the proportion of cured subjects. Liu and Shen [22] derived the maximum likelihood estimation approach for fitting model (11) to interval-censored data and developed an EM algorithm for the implementation of the approach.

It is apparent that as model (1), model (11) may have some limitations and corresponding to this, Li et al.[21] considered the following class of semiparametric transformation cure model

$$S(t\,|\,X) = \exp\big[-G\big(F(t)\mathrm{e}^{\beta_0+\beta' X}\big)\big], \tag{12}$$

a generalization of model (4), where G and F are defined as above. For the observed interval-censored data giving the likelihood function in (9), the new observed likelihood function has the form

$$\begin{aligned} L(\beta_0,\beta,F) = \prod_{i=1}^{n} \big\{ &\exp\big[-G\big(F(L_i)\mathrm{e}^{\beta_0+X_i^T\beta}\big)\big] - \exp\big[-G\big(F(R_i)\mathrm{e}^{\beta_0+X_i^T\beta}\big)\big]\big\}^{\delta_i} \\ &\times \exp\big[-G\big(F(L_i)\mathrm{e}^{\beta_0+X_i^T\beta}\big)\big]^{1-\delta_i}, \end{aligned} \tag{13}$$

where $\delta_i = I(R_i < \infty)$. For inference about model (12), Li et al.[21] developed the maximum likelihood estimation procedure under the transformation given in (7) and provided an EM algorithm similar to that given in Wang et al.[12] and Zeng et al.[15].

In addition to those described above, other authors who investigated the analysis of interval-censored data with a cured subgroup include Hu and Xiang[23], Kim and Jhun[24], Lam et al.[25], Lam and Wong[26], Li and Ma[27], Liu et al.[28], Xiang et al.[29] and Zhou et al.[30]. In particular, Hu and Xiang[23] considered a model that is same as model (11) except replacing $\exp(\beta_0+\beta' X)$ by $\eta(\beta_0+\beta' X)$ and Zhou et al.[30] discussed the use of the generalized odds rate mixture cure model, where η is a known link function. Also Xiang et al.[29] and Lam and Wong[26] discussed the same problem but for clustered interval-censored data, and Liu et al.[28] considered the situation with mis-measured covariates.

There exist several directions for future research related to the topic discussed here. One is that for most of the existing methods for both right-censored and interval-censored data, model (10) is commonly used to model covariate effects on the cure rate. It is apparent that sometimes it may not provide a reasonable fit or one may prefer a different model and thus one would need to develop similar or different and new estimation procedures. Also most of the methods described above apply only to time-independent covariates and it is clear that it would be useful to generalize them to time-dependent covariates or the situation that involves both time-dependent covariates and time-varying regression coefficients.

§4. Analysis of Informatively Interval-Censored Data

As discussed in the literature and above, in the presence of informative censoring, the analysis that ignores it would yield biased or even misleading estimation or results.

For the situation, unlike the non-informative case where the analysis is usually performed conditional on the censoring mechanism or observation process, one needs to model the censoring mechanism or observation process together with the failure time of interest. For this, two types of approaches are commonly used and they are the frailty or latent variable-based approach and the copula model-based approach.

Consider a failure time study that gives case I interval-censored or current status data. Let T, C and X be defined as above and assume that T and C may be correlated. Under the framework of the Cox model and assuming that covariates may have effects on C too, one may model the covariate effects on T and C as

$$\lambda(t \,|\, X, b) = \lambda_0(t)\exp(\beta' X + b), \tag{14}$$

and

$$\lambda^c(t \,|\, X, b) = \lambda_0^c(t)\exp(\gamma' X + b), \tag{15}$$

respectively, in terms of the hazard function. In the above, $\lambda_0(t)$ and $\lambda_0^c(t)$ denote the baseline hazard functions, β and γ are regression parameters as defined in model (1), and b is a latent variable with mean zero, representing the association between T and C. By assuming that T and C are independent given X and b, Li et al.[31] proposed a sieve maximum likelihood estimation procedure and established the asymptotic properties of the proposed estimators of regression parameters as well as developing a three-stage data augmentation EM algorithm.

As discussed above, instead of the Cox model, sometimes one may prefer the additive hazards model and for this, one may replace model (14) by

$$\lambda(t \,|\, X, b) = \lambda_0(t) + \beta' X + b, \tag{16}$$

and fit models (16) and (15) together to current status data. Li et al.[32] investigated this method and also developed a sieve maximum likelihood estimation approach. Of course, instead of model (14) or (16), one could also employ the linear transformation frailty model

$$\Lambda(t \,|\, X) = \mathrm{e}^b G(\Lambda_0(t)\exp(\beta' X)) \tag{17}$$

together with model (15)[2]. In practice, one may consider other models such as model (16) or (17) for the observation time variable C too.

The copula model-based approach provides another way to describe the correlation between T and C by specifying the covariate effect on the marginal models for T and C. To be specific, consider a failure time study that involves n independent subjects and gives only current status data. Let the T_i's and X_i's be defined as above and C_i the potential observation time which may depend on T_i. Also suppose that there also exists a censoring time ζ_i such as the administrative stop time and define $\widetilde{C}_i = \min(C_i, \zeta_i)$, $\Delta_i = I(C_i \leqslant \zeta_i)$ and $\delta_i = I(T_i \leqslant \widetilde{C}_i)$. Then the observed data have the form $\{O_i = (\Delta_i, \delta_i, \widetilde{C}_i, Z_i),\ i = 1, 2, \cdots, n\}$. Let F_T and F_C denote the marginal distributions of T_i and C_i, respectively, and F the joint distribution of T_i and C_i. Then there exists a copula function $C_\alpha(u, v)$ defined on $I^2 = [0,1] \times [0,1]$ such that $F(t, c) = C_\alpha\{F_T(t), F_C(c)\}$, where α represents the relationship between T_i and C_i and is often referred to as the association parameter, and $C_\alpha(u, 0) = C_\alpha(0, v) = 0$, $C_\alpha(u, 1) = u$ and $C_\alpha(1, v) = v$. It follows that

$$\mathsf{P}(T \leqslant t \,|\, C = c, Z_i) = \frac{\partial C_\alpha(u, v)}{\partial v}\Big|_{u=F_T(t), v=F_C(c)} = m_\alpha\{F_T(t), F_C(c)\}$$

and the resulting likelihood function has the form

$$\begin{aligned} L(\theta) = \prod_{i=1}^{n} \Big\{ & \big[\big(m_\alpha\{F_T(\widetilde{c}_i), F_C(\widetilde{c}_i)\}\big)^{\delta_i} \big(1 - m_\alpha\{F_T(\widetilde{c}_i), F_C(\widetilde{c}_i)\}\big)^{1-\delta_i} f_C(\widetilde{c}_i)\big]^{\Delta_i} \\ & \times \big[\big(F_T(\widetilde{c}_i) - C_\alpha\{F_T(\widetilde{c}_i), F_C(\widetilde{c}_i)\}\big)^{\delta_i} \\ & \times \big(1 - F_T(\widetilde{c}_i) - F_C(\widetilde{c}_i) + C_\alpha\{F_T(\widetilde{c}_i), F_C(\widetilde{c}_i)\}\big)^{1-\delta_i}\big]^{1-\Delta_i} \Big\}. \end{aligned} \tag{18}$$

In the likelihood function above, f_C denotes the marginal density function of the C_i's and θ represents all unknown parameters.

For estimation, it is natural to maximize the likelihood function given in (18) if the copula function C and the association parameter α are known. Among others, Ma et al.[33] considered this approach for the situation where the T_i's and C_i's follow models (14) and (15) with $b = 0$, respectively, and developed a sieve maximum likelihood estimation procedure. Also Zhao et al.[34], Du et al.[35], Xu et al.[36] and Xu et al.[37] developed the same types of methods when instead model (14), the failure times of interest T_i's follow model (2), the generalized probit model, model (4) and the accelerated failure time model, respectively. Furthermore, Cui et al.[38] investigated the same problem as Ma et al.[33] and proposed a two-step estimation procedure that allows for the association parameter α to be estimated instead of assuming to be known.

Note that one difference between case I and case II or K interval-censored data is that for the former, only one variable C is needed to describe the censoring mechanism and in contrast, more variables are needed for the latter. To describe the latent variable-based approach for the analysis of case K interval-censored data with informative censoring, let the T_i's, X_i's, K_i's, U_{ij}'s and δ_{ij}'s be defined as above and define $N_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \leqslant t)$, a point process representing the observation process on subject i. To describe the covariate effect, assume that the hazard function of T_i has the form

$$\lambda(t \,|\, X_i, b_i) = \lambda_0(t) \exp(\beta' X_i + \tau b_i), \tag{19}$$

and $N_i(t)$ is non-homogeneous Poisson process with the intensity function

$$\lambda_h(t \,|\, X_i, b_i) = \lambda_{0h}(t) \exp(\gamma' X_i + b_i). \tag{20}$$

In the above, $\lambda_0(t)$ and β are defined as in model (1) or (14), τ and γ are scale and vector parameters, respectively, $\lambda_{0h}(t)$ denotes a completely unknown continuous baseline intensity function, and b_i is a latent variable with mean zero. By assuming that T_i and N_i are independent given X_i and b_i, one can derive the likelihood function of β, τ and $\Lambda_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s$ as

$$\begin{aligned} L(\beta, \tau, \Lambda_0) = \prod_{i=1}^{n} \Bigg\{ \prod_{j=1}^{K_i} \big\{ \exp\big[-\Lambda_0(U_{i,j-1}) \mathrm{e}^{\beta' X_i + \tau b_i} \big] - \exp\big[-\Lambda_0(U_{ij}) \mathrm{e}^{\beta' X_i + \tau b_i} \big] \big\}^{\delta_{ij}} \\ \times \big\{ \exp\big[-\Lambda_0(U_{iK_i}) \mathrm{e}^{\beta' X_i + \tau b_i} \big] \big\}^{1 - \sum\limits_{j=1}^{K_i} \delta_{ij}} \Bigg\} \end{aligned}$$

conditional on the U_{ij} and b_i's. Wang et al.[39] discussed this approach and proposed a two-step estimation procedure that first estimates model (20) and then model (19) based on maximizing the likelihood above. Chen and Shen[40] investigated the same problem as Wang et al.[39] and provided a maximum likelihood estimation procedure.

Following Wang et al.[39], Wang et al.[41] and Wang et al.[13] developed similar approaches under different models for the T_i's. The former considered a generalization of model (2) or (16) given by

$$\lambda(t \,|\, X_i, b_i) = \lambda_0(t) + \beta' X_i + \tau b_i, \tag{21}$$

while the latter proposed a generalization of model (4) given by

$$\Lambda(t \,|\, X_i, b_i) = G(\Lambda_0(t) \exp(\beta' X_i + \tau b_i)). \tag{22}$$

Here τ and the b_i's are defined as in model (19). Note that one advantage of these methods is that they do not require any assumption on or estimate the distribution of the latent variables but they may lose some efficiency. To address this, Wang et al.[42] studied the same problem under models (20) and (21) as Wang et al.[41] and proposed a sieve maximum likelihood estimation approach.

By following the work on the use of the copula model-based approach for current status data described above, Zhao et al.[43], Ma et al.[44] and Xu et al.[45] developed similar methods for the informatively interval-censored data given by the formulation $I = (L, R]$. More specifically, they considered the situation where the dependence between T and the censoring mechanism can be characterized by the correlation between T and

$W = R - L$, the length of the censoring interval. As Ma et al.[33] and others, they developed the maximum likelihood estimation procedures when T follows model (2), (1) or (4), respectively, W follows the Cox model, and the relationship between T and W can be described by a known copula function. In addition to these mentioned above, Chen and Shen[40], Liu et al.[46] and Zhao et al.[47] also discussed the analysis of interval-censored data with informative censoring and especially, Liu et al.[46] considered the case when there also exists a cured subgroup.

There exist several directions for future research on the analysis of informatively interval-censored data. One is that in the method given in Wang et al.[39] and other similar methods, the observation process $N_i(t)$ has been assumed to be a non-homogeneous Poisson process with the intensity function given in (20). It is apparent that sometimes this may not be true and thus it would be useful to relax this assumption or generalize the existing methods to more general situations. The copula model-based approach has been commonly used for modeling the relationship among correlated variables in general and for the analysis of informatively interval-censored data as discussed above. On the other hand, for the latter, the focus has mainly been on two-dimensional copula models as discussed above and it would be helpful to apply higher dimensional copula models to case K or general, informatively interval-censored data although it may not be easy. Instead of the two approaches discussed above, Zhu et al.[48] discussed a third approach, the marginal approach that avoids to model the censoring process and employs the inverse probability weighted technique, under the additive hazards model (2) and the linear transformation model (4). Among others, one issue that clearly needs more research is the comparison to the two approaches discussed above.

§5. Analysis of Multivariate and Clustered Interval-Censored Data

Multivariate or clustered failure time data occur when a study involves more than one failure times of interest that are correlated. The former usually means that the number of correlated failure times is fixed, while the number of correlated failure times may change from one cluster to another for the latter. For the analysis of these data, one key component is how to describe or model the correlation among the correlated failure times and for this, similar to the analysis of informatively interval-censored data, two types of approaches, the latent variable-based and copula model-based approaches, are commonly used. In this section, we will first discuss some recent developments on regression analysis of multivariate interval-censored data and then regression analysis of clustered interval-censored data.

To describe the latent variable-based approach for the analysis of multivariate interval-censored data, consider a failure time study consisting of K possibly correlated failure events of interest. Let T_k denote the failure time of the kth event and X_k the vector of covariates that may have effects on T_k, $k = 1, 2, \cdots, K$. Assume that there exists a latent variable b with mean zero and given X_k and b, the cumulative hazard function T_k has the form

$$\Lambda_k(t \,|\, X_k, b) = G_k(\Lambda_k(t) \exp(\beta' X_k + b)). \tag{23}$$

In the above, G_k and Λ_k are defined as G and Λ_0 in model (4) but associated with the kth failure event. Among others, Li et al.[49] discussed the model above and its fitting to multivariate current status data. In particular, they considered the situation where G_k has the form (7) and b follows a parametric distribution with an unknown parameter and derived and established the maximum likelihood estimation procedure. Wang et al.[50] and Zhou et al.[51] also considered model (23) with b following the gamma distribution and $K = 2$. The former only focused on the situation where $G_k(x) = x$ and one observes bivariate current status data, and the latter developed a sieve maximum likelihood estimation method with the use of Bernstein polynomials for the approximation of the Λ_k's for general bivariate interval-censored data. In addition, Liu and Qin[52] investigated the same problem as Wang et al.[50] but under a class of probit models, and Gao et al.[53] discussed the situation with time-dependent covariates.

The authors who recently discussed the copula model-based approach for the analysis of multivariate interval-censored data include Hu et al.[54], Sun and Ding[55], Jiang and Cook[56] and Li et al.[57]. In particular, Hu et al.[54] considered the fitting of model (23) with $b = 0$ and $G_k = x$, the marginal Cox model, to bivariate current status data and proposed a sieve maximum likelihood estimation approach with the use of Bernstein polynomials for the approximation of both the Λ_k's and the copula function. Note that unlike other copula model-based methods discussed this paper, which usually assume that the underlying copula function is known or has a parametric form, their method directly estimates it. Sun and Ding[55] considered the analysis of bivariate general interval-censored data arising from model (23) with $b = 0$ but under a class of two-parameter Archimedean copula models. Furthermore, Jiang and Cook[56] and Li et al.[57] discussed the analysis of bivariate current status data and general interval-censored data, respectively. The former considered the situation where there exists a cured subgroup, while the latter employed a three-dimensional vine copula model and investigated the case with informative censoring.

For the analysis of clustered interval-censored data, as with the analysis of multivariate interval-censored data, one commonly used approach is the latent variable-based method. To describe this, consider a failure time study involving n clusters of subjects. Let T_{ij} and X_{ij} denote the failure time of interest of and a vector of covariates associated with subject j in the ith cluster, respectively, $j = 1, 2, \cdots, n_i$, and assume that the T_{ij}'s are independent for subjects in different clusters but dependent for those in the same cluster. Also assume that one observes case II interval-censored data given by U_{ij} and V_{ij} with $U_{ij} \leqslant V_{ij}$ and there exists a latent variable b_i such that given b_i, the T_{ij}'s with the same i are independent. Then the log likelihood function of all unknown parameters denoted by θ has the form

$$l(\theta) = \sum_{i=1}^{n} \ln \int \bigg\{ \prod_{j=1}^{n_i} [S(U_{ij} \,|\, X_{ij}, b_i; \theta) - S(V_{ij} \,|\, X_{ij}, b_i; \theta)] \bigg\} f(b_i; \eta) \mathrm{d}b_i,$$

where S denotes the survival function of T_{ij} given X_{ij} and b_i and f the density function of the b_i's with the unknown parameter η. Among others, Li et al.[58] and Lee et al.[59] discussed this approach under the following Cox frailty model

$$\lambda(t \,|\, X_{ij}, b_i) = \lambda_0(t) \exp(\beta' X_{ij} + b_i) \tag{24}$$

and a model similar to model (23), respectively. Both proposed some sieve maximum likelihood estimation procedures. Zeng et al.[60] also considered this approach under a more general class of semiparametric transformation models, generalizations of models (6) and (23). Furthermore the maximum likelihood estimation procedure proposed in Zeng et al.[60] can apply to both time-dependent covariates and the combination of multivariate and clustered interval-censored data.

Note that in the methods described above, the latent variable is assumed to follow a known distribution with some unknown parameters that are estimated along with other parameters. Sometimes one may not want to specify the distribution of the latent variable or prefer to leave the correlation among the failure times of interest arbitrary. Among others, Chen et al.[61] investigated this under the following additive hazards frailty model

$$\lambda(t \,|\, X_{ij}, b_i) = \lambda_0(t) + \beta' X_{ij} + b_i, \tag{25}$$

where the b_i's are defined as in model (24) but with an arbitrary distribution. For estimation, they provided a multivariate imputation method for general interval-censored data. Zhao et al.[62] considered the same problem but under the linear transformation model (4) and proposed a within-cluster-resampling estimation procedure. Furthermore their method allows for the presence of informative cluster size, meaning that n_i may be related to T_{ij}.

Compared to univariate interval-censored data, the literature on multivariate and clustered interval-censored data is clearly relatively limited partly due to more complicated data structures. For example, there seems to exist little literature on regression analysis of multivariate or clustered interval-censored data with time-dependent covariates and/or time-varying regression coefficients or in the presence of informative interval censoring. For the analysis of multivariate right-censored failure time data under the Cox model (1), the marginal approach is commonly used. It has the advantage of simplicity but may not be efficient. It would be useful to develop similar estimation methods for multivariate or clustered interval-censored data although may not be easy.

§6. Analysis of Case-Cohort Interval-Censored Data

Case-cohort studies are commonly performed to reduce the cost on the collection of covariate information and this is especially the case in, for example, large epidemiological cohort studies, where the assembling or collecting of covariate information on all study subjects may be expensive. Instead of collecting the information from all subjects, the case-cohort design selects a random sample or sub-cohort from the original whole cohort and collects or measures the covariate information only from the subjects in the sub-cohort or who experience the failure event of interest. Given the follow-up nature of such studies, it is easy to see that one may often only observe interval-censored data.

Several authors have recently investigated the analysis of interval-censored data arising from case-cohort studies. To describe their work, consider a failure time study giving case II interval-censored data denoted by $\{U_i, V_i, \delta_{1i}, \delta_{2i}, X_i\}$ if all covariates were observed, where U_i, V_i, δ_{1i}, δ_{2i} and X_i are defined as above but associated with subject i, $i = 1, 2, \cdots, n$. Define $\xi_i = 1$ if the covariate X_i is known or observed and 0 otherwise, $i = 1, 2, \cdots, n$. Then under the case-cohort design, the observed data have the form

$$\{(U_i, V_i, \delta_{1i}, \delta_{2i}, \xi_i X_i, \xi_i);\ i = 1, 2, \cdots, n\}.$$

Assume that the sub-cohort is selected based on the independent Bernoulli sampling with the selection probability $q \in (0, 1)$. Then the probability that the covariate X_i can be observed is given by

$$\mathsf{P}(\xi_i = 1) = \delta_{1i} + \delta_{2i} + (1 - \delta_{1i} - \delta_{2i})q, \qquad i = 1, 2, \cdots, n.$$

For estimation, a common approach is to maximize the inverse probability weighted log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} w_i \{\delta_{1i} \ln[1 - S(U_i \,|\, X_i)] + \delta_{2i} \ln[S(U_i \,|\, X_i) - S(V_i \,|\, X_i)]$$

$$+ (1 - \delta_{1i} - \delta_{2i}) \ln S(V_i \mid X_i)\}$$

assuming independent censoring. Here θ denotes all unknown parameters, S the survival function of T_i given X_i, and

$$w_i = \frac{\xi_i}{\delta_{1i} + \delta_{2i} + (1 - \delta_{1i} - \delta_{2i})q}.$$

Zhou et al.[63] discussed the method above for the situation where the T_i's follow the Cox model (1), while Du et al.[64] developed a similar method for the situation when the T_i's follow the additive hazards model (2). In addition, Du et al.[65] generalized the method given in Zhou et al.[63] to the case where interval censoring may be informative, and Zhao et al.[66] discussed the variable selection problem. More specifically, the former proposed a frailty model-based method for case II interval-censored data under models (19) and (20), and the latter considered the situation where the failure times of interest T_i's follow the Cox model (1) and the number of covariates is smaller than the sample size n.

As mentioned above, the case-cohort design collects covariate information only from the subject in a sub-cohort as well as those who experience the failure event of interest no matter they belong to the sub-cohort or not. A more general type of such designs is the so-called outcome-dependent sampling design, which aims to over-sample the subjects from the segments of the study population that are thought to be more informative in terms of the relationship between the failure time of interest and covariates or the outcome and exposure. It is easy to see that as case-cohort studies, these studies can often yield interval-censored data too. Among others, Zhou et al.[67,68] discussed the analysis of case II interval-censored data arising from the outcome-dependent sampling study under the Cox model (1) and provided some estimation methods for regression parameters.

It is apparent that much more research is needed for the analysis of interval-censored failure time data arising from case-cohort studies, especially for informatively interval-censored data. Actually there exists little literature even for the case of informatively right-censored data generated by case-cohort studies. More specifically, it would be useful to generalize some of the methods discussed in Section 4 such as that given in Wang et al.[39] to case-cohort studies. Also it is easy to see that multivariate or clustered interval-censored data can occur in case-cohort studies too and thus it would be interesting to generalize some of the methods discussed in Section 5 to the current situation. For most of the methods discussed above or in the literature, the sub-cohort is usually assumed to be selected based on the independent Bernoulli sampling and obviously this may not true or one may prefer a different sampling scheme sometimes. Then it is clear that the methods above may not be valid anymore and one would need to develop some other estimation procedures.

§7. Variable Section for Interval-Censored Data

Variable selection has recently attracted a great deal of attention with a huge amount of literature established under various contexts. This is particularly true for the analysis of failure time data and for the purpose, a general, commonly used approach has been the penalized method that maximizes an objective function minus a penalty function. To describe some recently developed methods for variable selection based on interval-censored failure time data, consider a failure time study giving general interval-censored data $\{(L_i, R_i], X_i;\, i = 1, 2, \cdots, n\}$ from n independent subjects. Let $S(t \,|\, X_i)$ denote the survival function of the failure time T_i given X_i. Then under the independent censoring assumption, the log likelihood function has the form

$$l(\beta, \theta) = \sum_{i=1}^{n} \ln[S(L_i \,|\, X_{ij}) - S(R_i \,|\, X_{ij})], \tag{26}$$

and one can perform the variable selection by maximizing the penalized likelihood function $l_p(\beta, \theta) = l(\beta, \theta) - p_\lambda(\beta)$. Here β and θ represent the regression parameters of interest and nuisance parameters, respectively, and p_λ denotes a penalty function with the tuning parameter λ.

Among others, Zhao et al.[69] discussed the approach above for the situation where the T_i's follow the Cox model (1) and proposed a broken adaptive ridge (BAR) regression procedure. For the situation, θ represents the unknown cumulative baseline hazard function Λ_0, which was approximated by the Bernstein polynomials in the method. In particular, they considered the BAR penalty function given by

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{\widetilde{\beta}_j^2}, \tag{27}$$

and proved that the resulting variable selection and estimation procedure has both the oracle property and the grouping property. In the above, β_j denotes the jth component of β, p the dimension of β, and $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \cdots, \widetilde{\beta}_p)'$ a consistent estimator of β with no zero component. Following Zhao et al.[69], Li et al.[9] generalized the method to the situation where the T_i's follow the linear transformation model (4), and as mentioned above, Zhao et al.[66] considered the same problem for interval-censored data arising from case-cohort studies.

A limitation of the method given in Zhao et al.[69] is that the number of covariates p cannot be larger than the sample size n although p can be diverging with n. To address this, Wu et al.[10] generalized it to the situation where the hazard function of T_i has the form

$$\lambda(t\,|\,X_i, Z_i) = \lambda_0(t)\exp[\beta' X_i + \psi(Z_i)], \tag{28}$$

Here as above, X_i denotes a vector of covariates associated with subject i but can be high-dimensional or with p larger than n, the focus of the variable selection, Z_i a vector of low-dimensional covariates that should always be included in the model, and $\psi(Z_i) = \sum\limits_{j=1}^{q} \psi_j(Z_{ij})$ with the ψ_j's being unknown functions and $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{iq})'$. In other words, the Z_{ij}'s are some covariates that may have non-linear effects on T_i.

Other authors who have studied variable selection for interval-censored data include Hu et al.[70], Scolas et al.[71], Sun et al.[72], Wu and Cook[73], Xu et al.[74] and Yi et al.[75]. In particular, Wu and Cook[73] considered the same problem as Zhao et al.[69] but under the Cox model with the baseline hazard function being a piecewise constant function. Sun et al.[72] and Xu et al.[74] proposed some variable selection procedures for interval-censored data where there may exist a cured subgroup, and Yi et al.[75] considered the variable selection under the context of joint analysis of longitudinal data and interval-censored data. Note that all of the methods mentioned above apply only to either low- or high-dimensional situations but would not be valid for ultra-high-dimensional covariates. To address the latter situation, Hu et al.[70] developed a model-free or nonparametric screening and feature selection procedure based on the idea of cumulative residuals for interval-censored data. In particular, they proved that the method has the sure independent screening property and will tend to rank the active or significant covariates above the inactive or non-significant ones in terms of their association with the failure time of interest.

More research is needed for variable selection under the context of interval-censored failure time data. One topic is that in all methods described above, each covariate is treated individually but sometimes there may exist some known group structures among covariates. For the situation, it is apparent that some group variable selection procedures that can take into account the group structure should be developed. Another topic that has not been investigated is variable selection for the model with interaction terms[76]. For example, sometimes one may need to consider the following quadratic Cox model

$$\lambda(t\,|\,X) = \lambda_0(t)\exp\Big(\sum_{j=1}^{p}\beta_j X_j + \sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\beta_{ij} X_{j_1} X_{j_2}\Big),$$

where β_j and X_j denote the jth component of β and X, respectively, and p the dimension of β and X. It is easy to see that the methods discussed above cannot apply to this situation. A third direction for future research is to generalize some of the methods above to interval-censored data with time-dependent covariates and/or time-varying regression coefficients as well as to multivariate and clustered interval-censored data.

§8. Discussion and Concluding Remarks

As mentioned above, this paper aims to discuss some recent advances on several important topics related to regression analysis of interval-censored failure time data but not give a comprehensive review of recent literature on interval-censored data. There exist several topics or types of interval-censored data on which some recent advances have been made but not discussed above, including the analysis of truncated and/or doubly interval-censored data, the analysis of interval-censored data with missing or auxiliary covariates, nonparametric estimation of a survival function based on interval-censored data, and the Bayesian approach for the analysis of interval-censored data.

In the preceding sections, the discussion has focused on the failure time variable that starts or measures the time from zero to the occurrence of an event of interest. Sometimes the failure time of interest may measure or represent the elapse time between two successive events such as the onset of a disease and the death due to the disease. One will observe doubly interval-censored data if the observations on either or both events suffer interval censoring, and among others, one area that often produces such data is epidemiological studies on disease progression such as HIV and AIDS. For the situation, left truncation may often occur together too and thus one has to face left-truncated and doubly interval-censored data. Among others, Wang et al.[77] and Wang et al.[78] recently discussed regression analysis of such data under the additive hazards model (2) and the Cox model (1), respectively. For inference, the former developed an efficient maximum likelihood estimation method, while the latter proposed a pairwise pseudo-likelihood estimation approach. In addition, Wang et al.[79] and Wu et al.[80] also investigated the situation when there exists a cured subgroup, and Gao and Chan[81] considered a special situation that yields length-biased and interval-censored data.

It is worth to note that in the literature, the term doubly censored data is sometimes also used to denote the type of failure time data where the failure time variable of interest T is either left- or right-censored if $T \leqslant L_i$ or $T > R$, respectively, and exactly observed if $L < T \leqslant R$ with $L < R$. Among others, Li et al.[82] and Li et al.[83] discussed regression analysis of univariate and multivariate doubly censored data under the semiparametric transformation model (6) and a model similar to model (23), respectively. For inference, both developed nonparametric maximum likelihood estimation procedures. Instead of doubly censored data, in practice, one may face doubly truncated data where L and R serve as left- and right-truncated variables rather than censoring variables. Among others, Ying et al.[84] and Liu et al.[85] recently considered regression analysis of such data and proposed weighted rank estimation procedures for regression parameters under a class of linear models and transformation models, respectively.

Missing covariates often occur in failure time studies and in particular, missing covariates and interval-censored data can often occur together in various settings including demographic, epidemiological, financial, medical and sociological studies. For the analysis of data with missing covariates, a naive approach is to analyze only the subjects with complete covariates but many authors have pointed out that this not only may be inefficient but also could yield biased estimators. Although many methods have been developed for the analysis of right-censored failure time data with missing covariates, there exists limited literature on the analysis of interval-censored data with missing covariates. The authors who investigated the problem include Li et al.[86] and Du et al.[87], and both focused on case II interval-censored data arising from the additive hazards model (2) and the linear transformation model (4), respectively. The former provided several inverse probably weighted estimation procedures when covariates may be missing at random, while the latter proposed a two-step estimation procedure for non-ignorable missing covariates. The asymptotic properties of the proposed estimators of regression parameters were established in both situations. In addition, Chen et al.[88] discussed regression analysis of multivariate current status data with auxiliary covariates.

The other authors who have recently discussed the analysis of interval-censored data include He et al.[89], Gamage et al.[90], Li et al.[91], Lin et al.[92], Ou et al.[93], Shen[94], Wang et al.[95], Wang et al.[96], Wu and Cook[97], Xu et al.[98] and Zhang and Zhao[99]. In particular, Shen[94] and Wang et al.[95] discussed nonparametric estimation of a survival function based on interval-censored data in the presence of informative censoring and left truncation, respectively, and Lin et al.[92] and Zhang and Zhao[99] investigated the use of the Bayesian approach and the empirical likelihood approach for regression analysis of interval-censored data under the Cox model (1) and the linear transformation model (4), respectively. Also Li et al.[91] considered the situation where there may exist misclassification in the analysis of current status data, Ou et al.[93] discussed quantile regression of interval-censored data, and Xu et al.[98] studied joint analysis of interval-censored data and panel count data. Furthermore, Szabo et al.[100] investigated the fitting of the accelerated hazards model to interval-censored data, and Zhang and Zhang[101] studied the spatial modeling of interval-censored data.

References

[1] BOGAERTS K, KOMÁREK A, LESAFFRE E. *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS* [M]. Boca Raton: Chapman and Hall/CRC, 2017.

[2] CHEN D G, SUN J G, PEACE K E. *Interval-Censored Time-to-Event Data: Methods and Applications* [M]. New York: Chapman and Hall/CRC, 2012.

[3] SUN J G. *The Statistical Analysis of Interval-Censored Failure Time Data* [M]. New York: Springer, 2006.

[4] VAN DEN HOUT A. *Multi-State Survival Models for Interval-Censored Data* [M]. New York: Chapman and Hall/CRC, 2016.

[5] MA L, FENG Y Q, CHEN D G, et al. Interval-censored time-to-event data and their applications in clinical trials [M] // YOUNG W R, CHEN D G. (eds.) *Clinical Trial Biostatistics and Biopharmaceutical Applications*. New York: Chapman and Hall/CRC, 2014: 307–334.

[6] SUN J G, LI J L. Interval censoring [M] // KLEIN J P, VAN HOUWELINGEN H C, IBRAHIM J G, et al. (eds.) *Handbook of Survival Analysis*. New York: Chapman and Hall/CRC, 2013: 369–390.

[7] SUN J G, ZHOU Q N, CHEN D G. Clinical trials: interval-censored failure time data [M] // CHOW S C. (ed.) *Encyclopedia of Biopharmaceutical Statistics*. 4th ed. Boca Raton: Chapman and Hall/CRC, 2018: 589–596.

[8] LI K, CHAN W, DOODY R S, et al. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data [J]. *J Alzheimers Dis*, 2017, **58(2)**: 361–371.

[9] LI S W, WU Q W, SUN J G. Penalized estimation of semiparametric transformation models with interval-censored data and application to Alzheimer's disease [J]. *Stat Methods Med Res*, 2020, **29(8)**: 2151–2166.

[10] WU Q W, ZHAO H, ZHU L, et al. Variable selection for high-dimensional partly linear additive Cox model with application to Alzheimer's disease [J]. *Stat Med*, 2020, **39(23)**: 3120–3134.

[11] KALBFLEISCH J D, PRENTICE R L. *The Statistical Analysis of Failure Time Data* [M]. 2nd ed. New York: Wiley, 2002.

[12] WANG L M, MCMAHAN C S, HUDGENS M G, et al. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data [J]. *Biometrics*, 2016, **72(1)**: 222–231.

[13] WANG P J, ZHAO H, DU M Y, et al. Inference on semiparametric transformation model with general interval-censored failure time data [J]. *J Nonparametr Stat*, 2018, **30(3)**: 758–773.

[14] COX D R. Regression models and life-tables (with discussion) [J]. *J Roy Statist Soc Ser B*, 1972, **34(2)**: 187–220.

[15] ZENG D L, MAO L, LIN D Y. Maximum likelihood estimation for semiparametric transformation models with interval-censored data [J]. *Biometrika*, 2016, **103(2)**: 253–271.

[16] YI F T, TANG N S, SUN J G. Regression analysis of interval-censored failure time data with time-dependent covariates [J]. *Comput Statist Data Anal*, 2020, **144**: 106848 (14 pages).

[17] CHEN C M, SHEN P S, TSENG Y K. Semiparametric transformation joint models for longitudinal covariates and interval-censored failure time [J]. *Comput Statist Data Anal*, 2018, **128**: 116–127.

[18] ZHANG H, WANG P J, SUN J G. Regression analysis of interval-censored failure time data with possibly crossing hazards [J]. *Stat Med*, 2018, **37(5)**: 768–775.

[19] HU T, XIANG L M. Partially linear transformation cure models for interval-censored data [J]. *Comput Statist Data Anal*, 2016, **93**: 257–269.

[20] MA S G. Mixed case interval censored data with a cured subgroup [J]. *Statist Sinica*, 2010, **20(3)**, 1165–1181.

[21] LI S W, HU T, ZHAO X Q, et al. A class of semiparametric transformation cure models for interval-censored failure time data [J]. *Comput Statist Data Anal*, 2019, **133**: 153–165.

[22] LIU H, SHEN Y. A semiparametric regression cure model for interval-censored data [J]. *J Amer Statist Assoc*, 2009, **104(487)**: 1168–1178.

[23] HU T, XIANG L M. Efficient estimation for semiparametric cure models with interval-censored data [J]. *J Multivariate Anal*, 2013, **121**: 139–151.

[24] KIM Y J, JHUN M. Cure rate model with interval censored data [J]. *Stat Med*, 2008, **27(1)**: 3–14.

[25] LAM K F, WONG K Y, ZHOU F F. A semiparametric cure model for interval-censored data [J]. *Biom J*, 2013, **55(5)**: 771–788.

[26] LAM K F, WONG K Y. Semiparametric analysis of clustered interval-censored survival data with a cure fraction [J]. *Comput Statist Data Anal*, 2014, **79**: 165–174.

[27] LI J L, MA S G. Interval-censored data with repeated measurements and a cured subgroup [J]. *J R Stat Soc Ser C Appl Stat*, 2010, **59(4)**: 693–705.

[28] LIU Y Q, HU T, SUN J G. Regression analysis of interval-censored failure time data with cured subgroup and mismeasured covariates [J]. *Comm Statist Theory Methods*, 2020, **49(1)**: 189–202.

[29] XIANG L M, MA X M, YAU K K W. Mixture cure model with random effects for clustered interval-censored survival data [J]. *Stat Med*, 2011, **30(9)**: 995–1006.

[30] ZHOU J, ZHANG J J, LU W B. Computationally efficient estimation for the generalized odds rate mixture cure model with interval-censored data [J]. *J Comput Graph Statist*, 2018, **27(1)**: 48–58.

[31] LI S W, HU T, WANG P J, et al. Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments [J]. *Comput Statist Data Anal*, 2017, **110**: 75–86.

[32] LI H Q, ZHANG H, SUN J G. Estimation of the additive hazards model with current status data in the presence of informative censoring [J]. *Stat Interface*, 2019, **12(2)**: 321–330.

[33] MA L, HU T, SUN J G. Sieve maximum likelihood regression analysis of dependent current status data [J]. *Biometrika*, 2015, **102(3)**: 731–738.

[34] ZHAO S S, HU T, MA L, et al. Regression analysis of informative current status data with the additive hazards model [J]. *Lifetime Data Anal*, 2015, **21(2)**: 241–258.

[35] DU M Y, HU T, SUN J G. Semiparametric probit model for informative current status data [J]. *Stat Med*, 2019, **38(12)**: 2219–2227.

[36] XU D, ZHAO S S, HU T, et al. Regression analysis of informative current status data with the semiparametric linear transformation model [J]. *J Appl Stat*, 2019, **46(2)**: 187–202.

[37] XU D, ZHAO S S, SUN J G. Regression analysis of dependent current status data with the accelerated failure time model [J/OL]. *Comm Statist Simulation Comput*, 2020 [2020-07-30]. https://doi.org/10.1080/03610918.2020.1797795.

[38] CUI Q, ZHAO H, SUN J G. A new copula model-based method for regression analysis of dependent current status data [J]. *Stat Interface*, 2018, **11(3)**: 463–471.

[39] WANG P J, ZHAO H, SUN J G. Regression analysis of case K interval-censored failure time data in the presence of informative censoring [J]. *Biometrics*, 2016, **72(4)**: 1103–1112.

[40] CHEN C M, SHEN P S. Semiparametric regression analysis of failure time data with dependent interval censoring [J]. *Stat Med*, 2017, **36(21)**: 3398–3411.

[41] WANG S Y, WANG C J, WANG P J, et al. Semiparametric analysis of the additive hazards model with informatively interval-censored failure time data [J]. *Comput Statist Data Anal*, 2018, **125**: 1–9.

[42] WANG S Y, WANG C J, WANG P J, et al. Estimation of the additive hazards model with case K interval-censored failure time data in the presence of informative censoring [J]. *Comput Statist Data Anal*, 2020, **144**: 106891 (15 pages).

[43] ZHAO S S, HU T, MA L, et al. Regression analysis of interval-censored failure time data with the additive hazards model in the presence of informative censoring [J]. *Stat Interface*, 2015, **8(3)**: 367–377.

[44] MA L, HU T, SUN J G. Cox regression analysis of dependent interval-censored failure time data [J]. *Comput Statist Data Anal*, 2016, **103**: 79–90.

[45] XU D, ZHAO S S, HU T, et al. Regression analysis of informatively interval-censored failure time data with semiparametric linear transformation model [J]. *J Nonparametr Stat*, 2019, **31(3)**: 663–679.

[46] LIU Y Q, HU T, SUN J G. Regression analysis of current status data in the presence of a cured subgroup and dependent censoring [J]. *Lifetime Data Anal*, 2017, **23(4)**: 626–650.

[47] ZHAO H, CUI Q, SUN J G. A copula model approach for the additive hazards model with dependent current status data [J]. *Sci Sin Math*, 2019, **49(9)**: 1261–1272. (in Chinese)

[48] ZHU Y Y, CHEN Z Q, LAWLESS J F. Semiparametric analysis of interval-censored failure time data with outcome-dependent observation schemes [J/OL]. *Scand J Stat*, 2020 [2020-12-21]. https://doi.org/10.1111/sjos.12511.

[49] LI S W, HU T, ZHAO S S, et al. Regression analysis of multivariate current status data with semiparametric transformation frailty models [J]. *Statist Sinica*, 2020, **30(2)**: 1117–1134.

[50] WANG N C, WANG L M, MCMAHAN C S. Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm [J]. *Comput Statist Data Anal*, 2015, **83**: 140–150.

[51] ZHOU Q N, HU T, SUN J G. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data [J]. *J Amer Statist Assoc*, 2017, **112(518)**: 664–672.

[52] LIU H, QIN J. Semiparametric probit models with univariate and bivariate current-status data [J]. *Biometrics*, 2018, **74(1)**: 68–76.

[53] GAO F, ZENG D L, COUPER D, et al. Semiparametric regression analysis of multiple right- and interval-censored events [J]. *J Amer Statist Assoc*, 2019, **114(527)**: 1232–1240.

[54] HU T, ZHOU Q N, SUN J G. Regression analysis of bivariate current status data under the proportional hazards model [J]. *Canad J Statist*, 2017, **45(4)**, 410–424.

[55] SUN T, DING Y. Copula-based semiparametric regression method for bivariate data under general interval censoring [J]. *Biostatistics*, 2021, **22(2)**: 315–330.

[56] JIANG S, COOK R J. A mixture model for bivariate interval-censored failure times with dependent susceptibility [J]. *Stat Biosci*, 2020, **12(1)**: 37–62.

[57] LI H Q, MA C C, LI N, et al. A vine copula approach for regression analysis of bivariate current status data with informative censoring [J]. *J Nonparametr Stat*, 2020, **32(1)**: 185–200.

[58] LI J L, TONG X W, SUN J G. Sieve estimation for the Cox model with clustered interval-censored failure time data [J]. *Stat Biosci*, 2014, **6(1)**: 55–72.

[59] LEE C Y, WONG K Y, LAM K F, et al. Analysis of clustered interval-censored data using a class of semiparametric partly linear frailty transformation models [J/OL]. *Biometrics*, 2020 [2020-11-02]. https://doi.org/10.1111/biom.13399.

[60] ZENG D L, GAO F, LIN D Y. Maximum likelihood estimation for emiparametric regression models with multivariate interval-censored data [J]. *Biometrika*, 2017, **104(3)**: 505–525.

[61] CHEN L, SUN J G, XIONG C J. A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model [J]. *Comput Statist Data Anal*, 2016, **103**: 242–249.

[62] ZHAO H, MA C C, LI J L, et al. Regression analysis of clustered interval-censored failure time data with linear transformation models in the presence of informative cluster size [J]. *J Nonparametr Stat*, 2018, **30(3)**: 703–715.

[63] ZHOU Q, ZHOU H, CAI J. Case-cohort studies with interval-censored failure time data [J]. *Biometrika*, 2017, **104(1)**: 17–29.

[64] DU M Y, LI H Q, SUN J G. Additive hazards regression for case-cohort studies with interval-censored data [J]. *Stat Interface*, 2020, **13(2)**: 181–191.

[65] DU M Y, ZHOU Q N, ZHAO S S, et al. Regression analysis of case-cohort studies in the presence of dependent interval censoring [J]. *J Appl Stat*, 2021, **48(5)**: 846–865.

[66] ZHAO H, WU Q W, GILBERT P B, et al. A regularized estimation approach for case-cohort periodic follow-up studies with an application to HIV vaccine trials [J]. *Biom J*, 2020, **62(5)**: 1176–1191.

[67] ZHOU Q N, CAI J W, ZHOU H B. Outcome-dependent sampling with interval-censored failure time data [J]. *Biometrics*, 2018, **74(1)**: 58–67.

[68] ZHOU Q N, CAI J W, ZHOU H B. Semiparametric inference for a two-stage outcome-dependent sampling design with interval-censored failure time data [J]. *Lifetime Data Anal*, 2020, **26(1)**: 85–108.

[69] ZHAO H, WU Q W, LI G, et al. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression [J]. *J Amer Statist Assoc*, 2020, **115(529)**: 204–216.

[70] HU Q, ZHU L, LIU Y Y, et al. Nonparametric screening and feature selection for ultrahigh-dimensional case II interval-censored failure time data [J]. *Biom J*, 2020, **62(8)**: 1909–1925.

[71] SCOLAS S, EI GHOUCH A, LEGRAND C, et al. Variable selection in a flexible parametric mixture cure model with interval-censored data [J]. *Stat Med*, 2016, **35(7)**: 1210–1225.

[72] SUN L Q, LI S W, WANG L M, et al. Variable selection in semiparametric nonmixture cure model with interval-censored failure time data: an application to the prostate cancer screening study [J]. *Stat Med*, 2019, **38(16)**: 3026–3039.

[73] WU Y, COOK R J. Penalized regression for interval-censored times of disease progression: selection of HLA markers in psoriatic arthritis [J]. *Biometrics*, 2015, **71(3)**: 782–791.

[74] XU Y, ZHAO S S, HU T, et al. Variable selection for generalized odds rate mixture cure models with interval-censored failure time data [J]. *Comput Statist Data Anal*, 2021, **156**: 107115 (17 pages).

[75] YI F T, TANG N S, SUN J G. Simultaneous variable selection and estimation for joint models of longitudinal and failure time data with interval censoring [J/OL]. *Biometrics*, 2020 [2020-10-08]. https://doi.org/10.1111/biom.13387.

[76] LI C, SUN J G. Variable selection for high-dimensional quadratic Cox model with application to Alzheimer's disease [J]. *Int J Biostat*, 2020, **16(2)**: 20190121 (10 pages).

[77] WANG P J, TONG X W, ZHAO S S, et al. Efficient estimation for the additive hazards model in the presence of left-truncation and interval censoring [J]. *Stat Interface*, 2015, **8(3)**: 391–402.

[78] WANG P J, LI D N, SUN J G. A pairwise pseudo-likelihood approach for left-truncated and interval-censored data under the Cox model [J/OL]. *Biometrics*, 2020 [2020-10-15]. https://doi.org/10.1111/biom.13394.

[79] WANG P J, TONG X W, SUN J G. A semiparametric regression cure model for doubly censored data [J]. *Lifetime Data Anal*, 2018, **24(3)**: 492–508.

[80] WU Y, CHAMBERS C D, XU R H. Semiparametric sieve maximum likelihood estimation under cure model with partly interval censored and left truncated data for application to spontaneous abortion [J]. *Lifetime Data Anal*, 2019, **25(3)**: 507–528.

[81] GAO F, CHAN K C G. Semiparametric regression analysis of length-biased interval-censored data [J]. *Biometrics*, 2019, **75(1)**: 121–132.

[82] LI S W, HU T, WANG P J, et al. A class of semiparametric transformation models for doubly censored failure time data [J]. *Scand J Stat*, 2018, **45(3)**: 682–698.

[83] LI S W, HU T, TONG T J, et al. Semiparametric regression analysis of multivariate doubly censored data [J]. *Stat Model*, 2020, **20(5)**: 502–526.

[84] YING Z L, YU W, ZHAO Z Q, et al. Regression analysis of doubly truncated data [J]. *J Amer Statist Assoc*, 2020, **115(530)**: 810–821.

[85] LIU T Q, YUAN X H, SUN J G. Weighted rank estimation for nonparametric transformation models with doubly truncated data [J]. *J Korean Statist Soc*, 2021, **50(1)**: 1–24.

[86] LI H Q, ZHANG H, ZHU L, et al. Estimation of the additive hazards model with interval-censored data and missing covariates [J]. *Canad J Statist*, 2020, **48(3)**: 499–517.

[87] DU M Y, LI H Q, SUN J G. Regression analysis of censored data with nonignorable missing covariates and application to Alzheimer disease [J]. *Comput Statist Data Anal*, 2021, **157**: 107157 (15 pages).

[88] CHEN Y R, FENG Y Q, SUN J G. Regression analysis of multivariate current status data with auxiliary covariates under the additive hazards model [J]. *Comput Statist Data Anal*, 2015, **87**: 34–45.

[89] HE B H, LIU Y Y, WU Y S, et al. Semiparametric efficient estimation for additive hazards regression with case II interval-censored survival data [J]. *Lifetime Data Anal*, 2020, **26(4)**: 708–730.

[90] GAMAGE P W W, CHAUDARI M, MCMAHAN C S, et al. An extended proportional hazards model for interval-censored data subject to instantaneous failures [J]. *Lifetime Data Anal*, 2020, **26(1)**: 158–182.

[91] LI S W, HU T, SUN J G. Regression analysis of misclassified current status data [J]. *J Nonparametr Stat*, 2020, **32(1)**: 1–19.

[92] LIN X Y, CAI B, WANG L M, et al. A Bayesian proportional hazards model for general interval-censored data [J]. *Lifetime Data Anal*, 2015, **21(3)**: 470–490.

[93] OU F S, ZENG D L, CAI J W. Quantile regression models for current status data [J]. *J Statist Plann Inference*, 2016, **178**: 112–127.

[94] SHEN P S. Nonparametric estimators of survival function under the mixed case interval-censored model with left truncation [J]. *Lifetime Data Anal*, 2020, **26(3)**: 624–637.

[95] WANG C J, SUN J G, WANG D H, et al. Nonparametric estimation of interval-censored failure time data in the presence of informative censoring [J]. *Acta Math Appl Sin Engl Ser*, 2017, **33(1)**: 107–114.

[96] WANG P J, ZHOU Y, SUN J G. A new method for regression analysis of interval-censored data with the additive hazards model [J]. *J Korean Statist Soc*, 2020, **49(4)**: 1131–1147.

[97] WU Y, COOK R J. Assessing the accuracy of predictive models with interval-censored data [J/OL]. *Biostatistics*, 2020 [2020-03-14]. https://doi.org/10.1093/biostatistics/kxaa011.

[98] XU D, ZHAO H, SUN J G. Joint analysis of interval-censored failure time data and panel count data [J]. *Lifetime Data Anal*, 2018, **24(1)**: 94–109.

[99] ZHANG Z G, ZHAO Y C. Empirical likelihood for linear transformation models with interval-censored failure time data [J]. *J Multivariate Anal*, 2013, **116**: 398–409.

[100] SZABO Z, LIU X Y, XIANG L M. Semiparametric sieve maximum likelihood estimation for accelerated hazards model with interval-censored data [J]. *J Statist Plann Inference*, 2020, **205**: 175–192.

[101] ZHANG Y, ZHANG B. Semiparametric spatial model for interval-censored data with time-varying covariate effects [J]. *Comput Statist Data Anal*, 2018, **123**: 146–156.

区间删失失效时间数据的统计分析

杜明月 孙建国

(香港理工大学应用数学系, 香港) (密苏里大学统计系, 哥伦比亚, MO 65211, 美国)

摘　要: 区间删失失效时间数据是失效时间或事件发生时间数据的一般类型. 对于这类数据, 其感兴趣的失效时间仅已知或被观测落在某个区间内而不是被精确地观测到. 此类数据经常出现在很多领域, 例如人口统计研究, 流行病学研究, 医学或公共健康研究和社会科学, 以及其他不同的领域. 纵向或者定期随访研究是产生区间删失数据的常见领域, 例如许多临床试验或者观察性研究. 在本文中, 我们将在简要地讨论背景和一些常用模型后, 主要回顾一些过去五年时间左右, 在几个重要的与回归分析相关的主题上的近期进展, 以及区间删失数据分析中需要更多研究的一些问题.

关键词: 删失机制; 极大似然估计; 观察性研究; 定期随访; 回归分析

中图分类号: O212.1