Lifetime Data Analysis (2023) 29:555–584
https://doi.org/10.1007/s10985-022-09585-1

Consistent and robust inference in hazard probability and odds models with discrete-time survival data

Zhiqiang Tan[1]

Received: 23 February 2022 / Accepted: 4 December 2022 / Published online: 23 December 2022

Abstract

For discrete-time survival data, conditional likelihood inference in Cox's hazard odds model is theoretically desirable but exact calculation is numerical intractable with a moderate to large number of tied events. Unconditional maximum likelihood estimation over both regression coefficients and baseline hazard probabilities can be problematic with a large number of time intervals. We develop new methods and theory using numerically simple estimating functions, along with model-based and model-robust variance estimation, in hazard probability and odds models. For the probability hazard model, we derive as a consistent estimator the Breslow–Peto estimator, previously known as an approximation to the conditional likelihood estimator in the hazard odds model. For the hazard odds model, we propose a weighted Mantel–Haenszel estimator, which satisfies conditional unbiasedness given the numbers of events in addition to the risk sets and covariates, similarly to the conditional likelihood estimator. Our methods are expected to perform satisfactorily in a broad range of settings, with small or large numbers of tied events corresponding to a large or small number of time intervals. The methods are implemented in the R package `dSurvival`.

Keywords Breslow–Peto estimator · Mantel–Haenszel estimator · Model-robust variance estimation · Proportional hazards model

1 Introduction

Regression analysis with censored survival outcomes has been widely used and extensively studied. The subjects are covered in numerous articles and books (e.g., Andersen et al. 1993; Cox and Oaks 1984; Kalbfleisch and Prentice 1980; Therneau and Grambsch 2000). The dominant approach is to use Cox's (1972) proportional hazards models and conditional or partial likelihood inference. For continuous-time survival data, this

✉ Zhiqiang Tan
ztan@stat.rutgers.edu

1 Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA

approach is statistically desirable, where the baseline hazard function as an infinite-dimensional nuisance parameter can be eliminated through conditional inference successively given the event times, and large sample theory can be developed using counting processes. Moreover, this approach is computationally convenient, because the partial log-likelihood function is concave in a coefficient vector in the proportional hazard model.

Regression analysis with discrete-time survival data, however, has been developed to a limited extent, even though such data arise frequently in various applications (e.g., Allison 1982; Willett and Singer 2004). As remarked by Cox (1972), "Unfortunately it is quite likely in applications that the data will be recorded in a form involving ties." The number of tied events can be substantial, depending on the discrete-time units used to record the survival data. There are broadly three types of existing methods for handling discrete-time survival data. The first type is to use Cox's (1972) discrete-time version of proportional hazards models on hazard odds ratios and apply conditional inference given the numbers of events. This method retains the statistical superiority of eliminating the baseline hazard probabilities as nuisance parameters, but exact calculation is numerically intractable with a moderate or large number of ties. The second type of methods employ various ad hoc approximations to conditional likelihood estimation (Breslow 1974; Efron 1977; Peto 1972). These methods are often considered to yield satisfactory results with a small number of ties, but there remains the difficulty of handling a relatively large number of ties. Statistical properties of these methods seem to be ambiguous. In fact, the estimators of Breslow (1974) and Efron (1977) would in general be inconsistent under Cox's discrete-time proportional hazards model. The third type of methods resort to unconditional maximum likelihood over both regression coefficients and baseline hazard probabilities, either with pooled logistic regression corresponding to Cox's discrete-time model, or complementary log-log regression induced by grouping observations under Cox's continuous-time model (Prentice and Gloeckler 1978). While such methods are appropriate for a small number of time intervals, statistical performance of maximum likelihood estimation can be problematic in the presence of many time intervals, which leads to the same number of nuisance parameters.

We develop new methods and theory for regression analysis with discrete-time survival data, while accommodating a broad range of data configurations, including a small number of time intervals and large numbers of tied event times, or a large number of time intervals and small numbers of tied event times. In contrast with previous methods, we derive numerically simple estimating equations, motivated by but distinct from conditional or unconditional likelihood inference, and study model-based and model-robust statistical properties in two classes of regression models. The first model deals with how hazard probability ratios are associated with covariates, whereas the second model is Cox's discrete-time proportional hazards model on hazard odds ratios.

- We derive as a consistent estimator the Breslow–Peto estimator in the hazard probability model, even though the same estimator is known as an approximation to the partial likelihood estimator in the hazard odds model. We find that the

model-based asymptotic variance is no greater than the limit of the commonly used model-based variance estimator for the Breslow–Peto estimator.
- We propose a weighted Mantel–Haenszel estimator in Cox's hazard odds model, such that it is numerically tractable and expected to achieve similar performance as the conditional likelihood estimator. We show that the weighted Mantel–Haenszel estimating function is conditionally unbiased given the numbers of events in addition to the risk sets and covariates, similarly to the conditional likelihood estimator.
- We study both model-based and model-robust variance estimation. As a useful complement to model-based inference, model-robust variance estimation captures sampling variation of a point estimator with possible misspecification of a posited model. Moreover, the influence function obtained sheds light on the reduction of the asymptotic variance if the model is correctly specified.

See White (1982) and Manski (1988) for asymptotic theory in misspecified models, and Buja et al. (2019) for a recent discussion on model-robust variance estimation.

An important, technical advantage of our methods is that the sample estimating functions in regression coefficients are carefully constructed to achieve various unbiasedness properties, which are relevant in different asymptotic settings. First, the population estimating functions, defined as the probability limits of the sample estimating functions as the sizes of all risk sets increase to infinity, are unconditionally unbiased. Second, the sample estimating functions are conditionally unbiased successively given the risk sets and covariates. This property can be exploited to establish consistency of the point and variance estimators, while allowing some risk-set sizes bounded in probability as the sample size increases. Third, under the hazard odds model, the weighted Mantel–Haenszel estimating function is also conditionally unbiased given the numbers of events in addition to the risk sets and covariates. Consistency of the point and variance estimators can be obtained while conditioning on the numbers of events.

As a practical advantage, our methods are applicable to survival data recorded at various levels of discreteness in a unified manner. In fact, our methods are not only developed to be valid in the discrete-time setting, but also designed to *algebraically* recover the standard partial likelihood method in the case of no tied data and hence remain valid in the continuous-time setting. This unification is desirable, in freeing practitioners from deciding when to use the Breslow or Efron approximation or to use the pooled logistic regression or similar methods, depending on the level of discreteness (or the number of tied events) in the data being analyzed. From our perspective, analysis can be conducted in a statistically simple and principled manner, regardless of the discretization level: use the Breslow–Peto estimator for inference about hazard probability ratios or use the weighted Mantel–Haenszel estimator for inference about hazard odds ratios, together with the proposed variance estimators. For illustration, Fig. 1 shows how the estimates vary depending on the discretization intervals in the veteran's lung cancer data (Kalbflwisch & Prentice 1980), which are further discussed in Sect. 6. As the data are discretized in wider intervals, the Breslow–Peto estimates become more different from the conditional likelihood and weighted Mantel–Haenszel estimates.

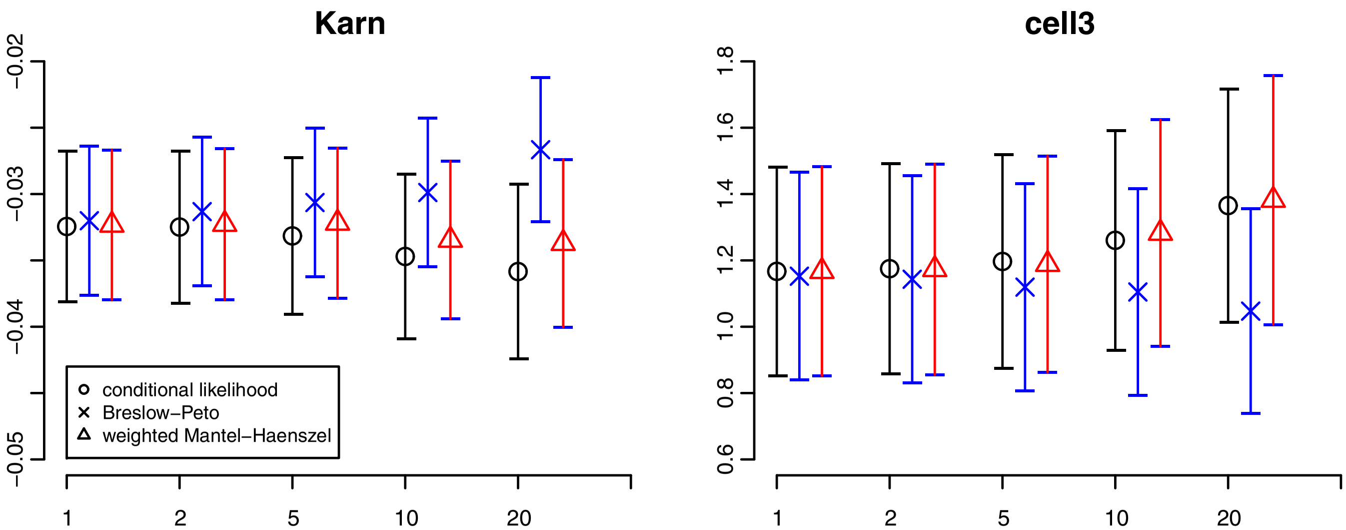

Fig. 1 Point estimates ± model-based standard errors from three methods, for the coefficients of `Karn` and `cell3` in the model in Sect. 6.1, depending on the discretization intervals in days (labeled on the x-axis) in the veteran's lung cancer data

2 Data and models

Suppose that survival data, possibly right-censored, and covariates are obtained as $\{(Y_i, \delta_i, X_i) : i = 1, \ldots, N\}$ from N individuals, where $Y_i = \min(T_i, C_i)$, $\delta_i = 1\{T_i \leq C_i\}$, T_i is an event time such as death time, C_i is a censoring time, and X_i is a covariate vector. Assume that $\{(T_i, C_i, X_i) : i = 1, \ldots, N\}$ are independent and identically distributed copies of (T, C, X), and hence $\{(Y_i, \delta_i, X_i) : i = 1, \ldots, N\}$ are independent and identically distributed copies of (Y, δ, X) with $Y = \min(T, C)$ and $\delta = 1\{T \leq C\}$. In addition, assume that the censoring and event variables, C and T, are independent conditionally on the covariate vector X.

In practice, survival data are usually recorded by grouping continuous or fine-scaled measurements, but such detailed data are not available to data analysis. While discretization of an uncensored time into an interval is straightforward, there are different options in assigning a censored time to a discrete value representing an interval. See Kaplan and Meier (1958); Thompson (1977), and Tan (2022, Supplement) for discussion on related issues. Nevertheless, assume that there are discrete values, $0 = t_0 < t_1 < \cdots < t_J < t_{J+1}$, such that (Y, δ) and (T, C) are properly defined with $C \in \{t_0, t_1, \ldots, t_J\}$ and $T \in \{t_1, \ldots, t_J, t_{J+1}\}$ and the conditionally independent censoring assumption is satisfied. An uncensored time in the interval $(t_{j-1}, t_j]$ is encoded as $Y = t_j$ and $\delta = 1$. For the censored-early option, a censored time in $[t_{j-1}, t_j)$ is encoded as $Y = t_{j-1}$ and $\delta = 0$ and the observation is included in the risk set up to time t_{j-1}. For the censored-late option, a censored time in $[t_{j-1}, t_j)$ is encoded as $Y = t_j$ and $\delta = 0$ and the observation is included in the risk set up to time t_j (Cox 1972).

We study extensions of Cox's proportional hazards model to discrete-time survival data as described above. For $j = 1, \ldots, J$, the hazard probability at time t_j given covariates $X = x$ is defined as $\pi_j(x) = P(T = t_j | T \geq t_j, X = x)$. This probability, under conditionally independent censoring, can be identified from observed data as

$$p_j(x) = P(Y = t_j, \delta = 1 | Y \geq t_j, X = x).$$

The subset $\{Y \geq t_j\}$, called the risk set at time t_j, represents individuals who are event-free (or alive) just prior to time t_j. In the following, we state probability and odds ratio models directly in terms of the event probabilities $p_j(x)$, which coincide with the hazard probabilities $\pi_j(x)$ if conditionally independent censoring holds, but otherwise remains empirically identifiable. For ease of interpretation, we treat $p_j(x)$ interchangeably with the hazard probabilities $\pi_j(x)$ whenever possible.

Consider two types of regression models on the hazard probabilities $p_j(x)$. The first places a parametric restriction on the probability ratios:

$$p_j(x) = p_j(x_0) \mathrm{e}^{(x - x_0)^{\mathrm{T}} \gamma^*}, \quad j = 1, \ldots, J, \tag{1}$$

where x_0 is a fixed vector of covariates, for example $x_0 = 0$, γ^* is an unknown coefficient vector, and the baseline hazard probabilities $\{p_j(x_0) : j = 1, \ldots, J\}$ are left to be unspecified, without further restriction other than being bounded between 0 and 1. The second model places a parametric restriction on the odds ratios:

$$\frac{p_j(x)}{1-p_j(x)} = \frac{p_j(x_0)}{1-p_j(x_0)} e^{(x-x_0)^{\mathrm{T}}\beta^*}, \quad j = 1, \ldots, J, \tag{2}$$

where β^* is an unknown coefficient vector and $\{p_j(x_0) : j = 1, \ldots, J\}$ are also left to be unspecified. In the limit of arbitrarily small time intervals, both models (1) and (2) can be seen to reduce to a Cox proportional hazards model:

$$\lambda_t(x) = \lambda_t(x_0) e^{(x-x_0)^{\mathrm{T}}\alpha^*},$$

where the survival time T is absolutely continuous with a hazard function $\lambda_t(x)$ given $X = x$, and α^* is an unknown coefficient vector. However, for discrete survival data, models (1) and (2) represent two alternative modeling approaches.

Model (2) is known as the discrete-time version of Cox's (1972) propositional hazard model. By comparison, model (1) seems to be previously not studied, although it can also be called a proportional hazards model because the hazard probability ratio $p_j(x)/p_j(x_0)$ is assumed to be constant in $j = 1, \ldots, J$. A potential limitation of model (1) is that the range of $p_j(x_0)$ or $p_j(x)$ as a probability between 0 and 1 may be violated for a fitted model, especially if model (1) is misspecified. The chance of such violation can be small if model (1) is correctly specified or approximately so. Examination of fitted hazard probabilities can serve as diagnosis. See Sect. 3 for further discussion.

3 Inference in hazard probability models

3.1 Point estimation

To derive a point estimator for γ^*, we rewrite model (1) as

$$P(Y = t_j, \delta = 1 | Y \geq t_j, X = x) = e^{\gamma_{0j}^* + x^{\mathrm{T}}\gamma^*}, \quad j = 1, \ldots, J, \tag{3}$$

where $\gamma_0^* = (\gamma_{01}^*, \ldots, \gamma_{0J}^*)^{\mathrm{T}}$ is a vector of unknown intercepts and γ^* is as before. Our estimators for (γ_0^*, γ^*) are defined jointly as a solution $(\hat{\gamma}_0, \hat{\gamma})$ to

$$\sum_{i:Y_i \geq t_j} \left(D_{ji} - e^{\gamma_{0j} + X_i^{\mathrm{T}}\gamma} \right) = 0, \quad j = 1, \ldots, J, \tag{4}$$

$$\sum_{j=1}^{J} \sum_{i:Y_i \geq t_j} \left(D_{ji} - e^{\gamma_{0j} + X_i^{\mathrm{T}}\gamma} \right) X_i = 0, \tag{5}$$

where $D_{ji} = 1\{Y_i = t_j, \delta_i = 1\}$, equal to 1 if $Y_i = t_j$ and $\delta_i = 1$ or 0 otherwise. Equation (4) depends only on the data from jth risk set $\{i : Y_i \geq t_j\}$, whereas Eq. (5) involves the data combined from all J risk sets. Within the jth risk set, the associated estimating functions in (γ_{0j}, γ) are $\sum_{i:Y_i \geq t_j} (D_{ji} - e^{\gamma_{0j} + X_i^{\mathrm{T}}\gamma})(1, X^{\mathrm{T}})^{\mathrm{T}}$, corresponding to quasi-likelihood score functions in model (3), viewed as a conditional moment restriction model with the Poisson logarithmic link for D_{ji} given X_i.

Solving (4) for γ_{0j} with fixed γ and substituting into (5) shows that

$$\mathrm{e}^{\hat{\gamma}_{0j}} = \frac{\sum_{i:Y_i \geq t_j} D_{ji}}{\sum_{i:Y_i \geq t_j} \mathrm{e}^{X_i^{\mathrm{T}} \hat{\gamma}}}. \tag{6}$$

and $\hat{\gamma}$ can be determined from the closed-form estimating equation

$$\sum_{j=1}^{J} \sum_{i:Y_i \geq t_j} \left(D_{ji} - \frac{\sum_{l:Y_l \geq t_j} D_{jl}}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \gamma}} \mathrm{e}^{X_i^{\mathrm{T}} \gamma} \right) X_i = 0. \tag{7}$$

By an exchange of indices i and l, Eq. (7) can be equivalently written as

$$\sum_{j=1}^{J} \sum_{i:Y_i \geq t_j} D_{ji} \left(X_i - \frac{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \gamma} X_l}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \gamma}} \right) = 0, \tag{8}$$

which is originally the estimating equation satisfied by the Breslow's (1974) and Peto's (1972) modification of the partial likelihood estimator to deal with tied event times in Cox's (continuous-time) proportional hazards model. Hence the estimator $\hat{\gamma}$ can be referred to as the Breslow–Peto estimator. Moreover, $\mathrm{e}^{\hat{\gamma}_{0j}}$ in (6) coincides with Breslow's (1974) estimator of the baseline hazard function. As a result, the difference

$$D_{ji} - \mathrm{e}^{\hat{\gamma}_{0j} + X_i^{\mathrm{T}} \hat{\gamma}} = D_{ji} - \frac{\sum_{l:Y_l \geq t_j} D_{jl}}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \hat{\gamma}}} \mathrm{e}^{X_i^{\mathrm{T}} \hat{\gamma}}$$

evaluated at $(\hat{\gamma}_{0j}, \hat{\gamma})$ is the martingale residual of ith individual at time t_j in Therneau et al. (1990), adapted to our setting of discrete survival data.

The probability ratio $p_j(x)/p_j(x_0)$ is generally closer to 1 than the odds ratios $p_j(x)(1-p_j(x_0))/\{p_j(x_0)(1-p_j(x))\}$. Hence our derivation explains the observation that the Breslow–Peto approximation often produces a conservative bias in estimating regression coefficients too close to 0 in proportional hazards models (Cox and Oaks 1984).

3.2 Model-robust inference

We study model-robust inference using $\hat{\gamma}$ with possible misspecification of model (3), similarly as in Lin and Wei (1989) for robust inference in Cox's proportional hazards model. Denote $R_{ji} = 1\{Y_i \geq t_j\}$, in addition to $D_{ji} = 1\{Y_i = t_j, \delta_i = 1\}$. Estimating Eq. (7) can be written as $\sum_{j=1}^{J} \hat{\zeta}_j(\gamma) = 0$, where

$$\hat{\zeta}_j(\gamma) = \frac{1}{n}\sum_{i=1}^n R_{ji}\left(D_{ji} - \frac{\sum_{l=1}^n R_{jl}D_{jl}}{\sum_{l=1}^n R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\bar{\gamma}}}\mathrm{e}^{X_i^{\mathrm{T}}\bar{\gamma}}\right)X_i.$$

Under suitable regularity conditions, it can be shown that $\hat{\gamma}$ converges in probability to a target value $\bar{\gamma}$, defined as a unique solution to the population version of (7) or equivalently (8):

$$0 = \sum_{j=1}^J E\left[R_j\left\{D_j - \frac{E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\mathrm{e}^{X^{\mathrm{T}}\gamma}\right\}X\right] \tag{9}$$

$$= \sum_{j=1}^J E\left[R_jD_j\left\{X - \frac{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}\right], \tag{10}$$

where $R_j = 1\{Y \geq t_j\}$, $D_j = 1\{Y = t_j, \delta = 1\}$, and $(\tilde{R}_j, \tilde{D}_j, \tilde{X})$ are defined from $(\tilde{Y}, \tilde{\delta}, \tilde{X})$ identically distributed as (Y, δ, X). Equivalently, $\bar{\gamma}$ is a unique maximizer of the objective function (which is concave in γ):

$$\sum_{j=1}^J E\left[R_jD_j\left\{X^{\mathrm{T}}\gamma - \log E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})\right\}\right]. \tag{11}$$

Moreover, $\hat{\gamma}$ can be shown to admit the asymptotic expansion

$$\hat{\gamma} - \bar{\gamma} = B(\bar{\gamma})^{-1}\sum_{j=1}^J \hat{\zeta}_j(\bar{\gamma}) + o_p(n^{-1/2}), \tag{12}$$

where $B(\gamma)$ is the negative Hessian of objective function (11), that is,

$$B(\gamma) = \sum_{j=1}^J E(\tilde{R}_j\tilde{D}_j)E\left[\frac{R_j\mathrm{e}^{X^{\mathrm{T}}\gamma}}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\left\{X - \frac{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}^{\otimes 2}\right].$$

Throughout, $x^{\otimes 2} = xx^{\mathrm{T}}$ for a vector x. From (12), the following result can be deduced, provided that the probability of survival beyond time t_J (which is the largest possible value of the censoring variable) is bounded away from 0. This boundedness condition is standard in large sample theory for survival analysis (e.g., Andersen et al. 1993, Condition VII.2.1), although further investigation can be of interest.

Proposition 1 *Assume that $P(T > t_J) \geq p_0$ for a constant $p_0 > 0$. Then $n^{1/2}(\hat{\gamma} - \bar{\gamma})$ converges in distribution to $\mathrm{N}(0, V)$ as $n \to \infty$, where $V = B(\bar{\gamma})^{-1}A(\bar{\gamma})B(\bar{\gamma})^{-1}$, $B(\gamma)$ is defined as above, $A(\gamma) = \mathrm{var}\{\sum_{j=1}^J h_j(Y, \delta, X; \gamma)\}$, and*

$$h_j(Y, \delta, X; \gamma) = R_j\left\{D_j - \frac{E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\mathrm{e}^{X^{\mathrm{T}}\gamma}\right\}\left\{X - \frac{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}.$$

Moreover, a consistent estimator of V *is* $\hat{V}_r = \hat{B}^{-1}(\hat{\gamma})\hat{A}(\hat{\gamma})\hat{B}^{-1}(\hat{\gamma})$, *where*

$$\hat{B}(\gamma) = \frac{1}{n}\sum_{j=1}^{J}\left(\sum_{l=1}^{n} R_{jl}D_{jl}\right)\sum_{i=1}^{n}\left[\frac{R_{ji}e^{X_i^{\mathrm{T}}\gamma}}{\sum_{l=1}^{n} R_{jl}e^{X_l^{\mathrm{T}}\gamma}}\left\{X_i - \frac{\sum_{l=1}^{n} R_{jl}e^{X_l^{\mathrm{T}}\gamma}X_l}{\sum_{l=1}^{n} R_{jl}e^{X_l^{\mathrm{T}}\gamma}}\right\}^{\otimes 2}\right],$$

$$\hat{A}(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{J}\hat{h}_j(Y_i, \delta_i, X_i; \gamma)\right\}^{\otimes 2},$$

and $\hat{h}_j(Y, \delta, X; \gamma)$ *is defined as* $h_j(Y, \delta, X; \gamma)$ *with* $E(\tilde{R}_j\tilde{D}_j)$, $E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})$, *and* $E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})$ *replaced by the sample averages* $n^{-1}\sum_{i=1}^{n} R_{ji}D_{ji}$, $n^{-1}\sum_{i=1}^{n} R_{ji}e^{\tilde{X}_i^{\mathrm{T}}\gamma}$, *and* $n^{-1}\sum_{i=1}^{n} R_{ji}e^{\tilde{X}_i^{\mathrm{T}}\gamma}X_i$.

Proposition 1 can be formally seen as an extension of Lin & Wei's (1989) result on the partial likelihood estimator in Cox's continuous-time model to the Breslow–Peto estimator used to handle tied event times. Lin & Wei's approach would use the Breslow–Peto modified score Eq. (8) and derive the asymptotic variance V with $h_j(Y, \delta, X; \gamma)$ defined as a correction to the modified score function

$$h_j(Y, \delta, X; \gamma) = R_j D_j\left\{X - \frac{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}\right\} - \frac{E(\tilde{R}_j\tilde{D}_j)e^{X^{\mathrm{T}}\gamma}}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}R_j\left\{X - \frac{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}.$$

From our approach, the asymptotic variance V can be equivalently derived using the estimating Eq. (7) based on the conditional moment model (3), and hence $h_j(Y, \delta, X; \gamma)$ is obtained as a correction to the associated estimating function

$$h_j(Y, \delta, X; \gamma) = R_j\left\{D_j - \frac{E(\tilde{R}_j\tilde{D}_j)e^{X^{\mathrm{T}}\gamma}}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}X - \frac{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma}\tilde{X})}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}R_j\left\{D_j - \frac{E(\tilde{R}_j\tilde{D}_j)e^{X^{\mathrm{T}}\gamma}}{E(\tilde{R}_j e^{\tilde{X}^{\mathrm{T}}\gamma})}\right\}.$$

As shown in (14), this representation is useful for simplification of the asymptotic variance V if model (3) is correct. By numerical evaluation, the variance estimator $\hat{V}_{\mathrm{r}}$ also appears to coincide with the robust variance estimator for the Breslow–Peto estimator in the R package `survival`, although no justification was provided.

3.3 Model-based inference

We study model-based inference using $\hat{\gamma}$ when model (3) is correctly specified. Under this assumption, $\hat{\gamma}$ is a consistent estimator of γ^*, with $\bar{\gamma} = \gamma^*$ satisfying the population estimating Eq. (9):

$$E\left[R_j\left\{D_j-\frac{E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma^*})}\mathrm{e}^{X^{\mathrm{T}}\gamma^*}\right\}X\right]=E\left[R_j\left\{D_j-\mathrm{e}^{\gamma^*_{0j}}\mathrm{e}^{X^{\mathrm{T}}\gamma^*}\right\}X\right]=0, \tag{13}$$

because $E\{R_j(D_j-\mathrm{e}^{\gamma^*_{0j}+X^{\mathrm{T}}\gamma^*})|X\}=0$ and $\mathrm{e}^{\gamma^*_{0j}}=E(\tilde{R}_j\tilde{D}_j)/E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma^*})$ by (3). The true value γ^* also satisfies the Breslow–Peto equation (10), by the equivalence between (9) and (10). This finding seems new. Interestingly, consistency of the Breslow–Peto estimator $\hat{\gamma}$ under model (3) is revealed more directly when defined through the new estimating Eq. (7) than through the usual Eq. (8).

There is also an interesting implication on model-based variance estimation. Under model (3), the difference $D_j-\mathrm{e}^{\gamma^*_{0j}+X^{\mathrm{T}}\gamma^*}$ has mean 0 conditionally on $R_j=1$ and X, and the individual terms $h_j(Y,\delta,X;\gamma^*)$, $j=1,\ldots,J$, are uncorrelated with each other. Then the asymptotic variance $V=B^{-1}(\gamma^*)A(\gamma^*)B^{-1}(\gamma^*)$ can be simplified such that

$$\begin{aligned} A(\gamma^*)&=\sum_{j=1}^{J}\mathrm{var}\{h_j(Y,\delta,X;\gamma^*)\} \qquad (14)\\ &=\sum_{j=1}^{J}E\left[R_jp_j(X)(1-p_j(X))\left\{X-\frac{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma^*}\tilde{X})}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma^*})}\right\}^{\otimes 2}\right], \end{aligned}$$

where $p_j(X)=\mathrm{e}^{\gamma^*_{0j}+X^{\mathrm{T}}\gamma^*}=\{E(\tilde{R}_j\tilde{D}_j)/E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\gamma^*})\}\mathrm{e}^{X^{\mathrm{T}}\gamma^*}$. A model-based estimator for the asymptotic variance V is then $\hat{V}_{\mathrm{b}}=\hat{B}^{-1}(\hat{\gamma})\hat{A}_{\mathrm{b}}(\hat{\gamma})\hat{B}^{-1}(\hat{\gamma})$ with

$$\hat{A}_{\mathrm{b}}(\gamma)=\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\left[R_{ji}\hat{p}_j(X_i;\gamma)(1-\hat{p}_j(X_i;\gamma))\left\{X_i-\frac{\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\gamma}X_l}{\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\hat{\gamma}}}\right\}^{\otimes 2}\right],$$

where $\hat{p}_j(X;\gamma)=\{(\sum_{l=1}^{n}R_{jl}D_{jl})/(\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\gamma})\}\mathrm{e}^{X^{\mathrm{T}}\gamma}$. By direct comparison, $A(\gamma^*)$ and $\hat{A}_{\mathrm{b}}(\hat{\gamma})$ are no greater than respectively $B(\gamma^*)$ and $\hat{B}(\hat{\gamma})$.

Corollary 1 *Suppose that model (3) is correctly specified. Then the asymptotic variance V for $n^{1/2}(\hat{\gamma}-\gamma^*)$ is, in the order on variance matrices, no greater than the $B^{-1}(\gamma^*)$, and the variance estimator $\hat{V}_b$ is no greater than $\hat{B}^{-1}(\hat{\gamma})$, the commonly used variance estimator for the Breslow–Peto estimator $\hat{\gamma}$.*

To accommodate small risk sets, we outline asymptotic theory conditionally on the risk sets and covariates and propose an improved model-based variance estimator. In fact, the foregoing justification of the asymptotic variance V and the variance estimators $\hat{V}_{\mathrm{r}}$ and $\hat{V}_{\mathrm{b}}$ rely on the assumption that all J risk sets are sufficiently large to ensure convergence of the sample averages $n^{-1}\sum_{l=1}^{n}R_{jl}D_{jl}$, $n^{-1}\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\gamma}$, and $n^{-1} \sum_{l=1}^{n} R_{jl} \mathrm{e}^{X_l^{\mathrm{T}} \gamma} X_l$ to their corresponding expectations for $j = 1, \ldots, J$. Alternatively, asymptotic properties of $\hat{\gamma}$ can be studied by exploiting the conditional unbiasedness of individual terms of the sample estimating function in (7) under model (3):

$$E\left\{\hat{\zeta}_j(\gamma^*)|R_{j,1:n}, X_{1:n}\right\} = 0, \quad j = 1, \ldots, J, \tag{15}$$

where $R_{j,1:n} = (R_{j1}, \ldots, R_{jn})$ and $X_{1:n} = (X_1, \ldots, X_n)$. This is a more elaborate property than unconditional unbiasedness (13). Under suitable regularity conditions similar as in fixed-design analysis of regression models, it can be shown that if model (3) is correctly specified, then $n^{1/2}(\hat{\gamma} - \gamma^*)$ converges in distribution to $\mathrm{N}(0, V_2)$ as $n \to \infty$, where $V_2 = B_2(\gamma^*)^{-1} A_2(\gamma^*) B_2(\gamma^*)^{-1}$, $B_2(\gamma) = \mathrm{plim}_{n\to\infty} \hat{B}(\gamma)$, $A_2(\gamma) = \mathrm{plim}_{n\to\infty} \sum_{j=1}^{J} v_j(\gamma)$, and $v_j(\gamma) = n \operatorname{var}\{\hat{\zeta}_j(\gamma)|R_{j,1:n}, X_{1:n}\}$, that is,

$$v_j(\gamma) = \frac{1}{n} \operatorname{var}\left\{\sum_{i=1}^{n} R_{ji}\left(D_{ji} - \frac{\sum_{l=1}^{n} R_{jl} D_{jl}}{\sum_{l=1}^{n} R_{jl} \mathrm{e}^{X_l^{\mathrm{T}} \gamma}} \mathrm{e}^{X_i^{\mathrm{T}} \gamma}\right) X_i \,\middle|\, R_{j,1:n}, X_{1:n}\right\}.$$

In the case where $P(T > t_J)$ is bounded away from 0 and all J risk sets are of sizes increasing to ∞, the asymptotic variance V_2 reduces to V in Proposition 1.

For the asymptotic variance V_2, our proposed estimator is $\hat{V}_{\mathrm{b}2} = \hat{B}(\hat{\gamma})^{-1} \hat{A}_{\mathrm{b}2}(\hat{\gamma}) \hat{B}(\hat{\gamma})^{-1}$, where $\hat{B}(\gamma)$ is as in Proposition 1, $\hat{A}_{\mathrm{b}2}(\gamma) = \sum_{j=1}^{J}\{\hat{v}_j(\gamma) + \hat{v}_j^{\mathrm{T}}(\gamma)\}/2$, and

$$\hat{v}_j(\gamma) = \frac{1}{n} \sum_{i=1}^{n} R_{ji}(1 - D_{ji}) \mathrm{e}^{X_i^{\mathrm{T}} \gamma} \frac{\sum_{l=1}^{n} R_{jl} \mathrm{e}^{X_l^{\mathrm{T}} \gamma}(X_i - X_l) \sum_{k=1}^{n} R_{jk} D_{jk} (X_i - X_k)^{\mathrm{T}}}{(\sum_{l=1}^{n} R_{jl} \mathrm{e}^{X_l^{\mathrm{T}} \gamma})^2}.$$

The matrix $\hat{v}_j(\gamma)$ is in general not symmetric, and $\{\hat{v}_j(\gamma) + \hat{v}_j^{\mathrm{T}}(\gamma)\}/2$ serves as a symmetrized version. The following properties can be established.

Proposition 2 *(i) Suppose that model* (3) *is correctly specified. For $j = 1, \ldots, J$, $\hat{v}_j(\gamma^*)$ is a conditionally unbiased estimator for $v_j(\gamma^*)$, that is,*

$$E\{\hat{v}_j(\gamma^*)|R_{j,1:n}, X_{1:n}\} = v_j(\gamma^*).$$

Hence $\hat{V}_{b2}$ can be a consistent estimator for V_2 even if some risk sets are of sizes which are bounded in probability as $J \to \infty$ and $n \to \infty$.
(ii) Suppose that at most one event is observed in each risk set $\{i : R_{ji} = 1\}$ for $j = 1, \ldots, J$. Then $\hat{A}_{b2}(\gamma)$ is identical to $\hat{B}(\gamma)$ and the variance estimator $\hat{V}_{b2}$ is identical to $\hat{B}(\hat{\gamma})^{-1}$, the usual variance estimator for the partial likelihood estimator (i.e., the Breslow–Peto estimator in the absence of tied events).

Property (ii) in Proposition 2 shows that the variance estimator $\hat{V}_{\mathrm{b}2}$ is the same as $\hat{B}^{-1}(\hat{\gamma})$ in the extreme case where there are no tied events and the Breslow–Peto estimator reduces to the maximum partial likelihood estimator. In contrast, the variance estimator $\hat{V}_{\mathrm{b}}$ in this case remains smaller than $\hat{B}^{-1}(\hat{\gamma})$.

The variance estimator $\hat{V}_{b2}$ is an extension of a model-based variance estimator in Tan (2022) for the Breslow–Peto estimator in a probability ratio model for analysis of 2×2 tables and two-sample survival analysis. See the Supplement for details of the relationship. For 2×2 tables, the variance estimator is designed to be consistent in two asymptotic settings, either with a fixed number of large tables or with a large number of possibly sparse tables. These two settings are originally considered for Mantel–Haenszel estimation of common odds ratios in 2×2 tables (Robins et al. 1986).

3.4 Estimation of survival probabilities

We discuss estimation of survival probabilities for individuals with fixed covariates x_0. For simplicity, assume that $x_0 = 0$ in model (1); otherwise the covariates can be recentered. Then the hazard probability $p_j(x_0)$ is identified as $e^{\gamma_{0j}}$, and can be estimated as $\hat{p}_j(x_0) = e^{\hat{\gamma}_{0j}}$ by (6). The kth survival probability, defined as $P_k(x_0) = P(T > t_k | X = x_0)$, can be estimated as

$$\hat{P}_k(x_0) = \prod_{j=1}^{k} \{1 - \hat{p}_j(x_0)\} = \prod_{j=1}^{k} \left(1 - e^{\hat{\gamma}_{0j}}\right), \quad k = 1, \ldots, J. \tag{16}$$

This is a discrete version of the product-limit estimator of the baseline survival function. Unless all $\hat{p}_j(x_0)$ are sufficiently small, the estimator (16) is distinct from an alternative estimator, $e^{-\sum_{j=1}^{k} e^{\hat{\gamma}_{0j}}}$, where $\sum_{j=1}^{k} e^{\hat{\gamma}_{0j}}$ is called the cumulative hazard. The alternative estimator is often used with continuous-time data in the R package `survival`. A potential disadvantage is that the estimator (6) for the hazard probability and hence (16) for the survival probability may be negative, in general due to the fact that the right hand side of model (3) is not restricted to be no greater than 1. Such negative estimates may also occur due to estimation error, particularly in the right tail.

The standard errors for $\hat{P}_k(x_0)$ can be obtained using Taylor expansions (or the delta method) and either model-robust or model-based variance estimator for $\hat{\gamma}$. In particular, model-robust variance estimation for $\hat{P}_k(x_0)$ involves use of the influence function of $\hat{\gamma}$ depending on the data from all J risk sets. Model-based variance estimation for $\hat{P}_k(x_0)$ admits a decomposition similar to variance estimation of the cumulative hazard in Tsiatis (1981). See the Supplement for detailed derivation and formulas, assuming that all J risk sets are sufficiently large. Further investigation is needed, including formal justification of our formulas in the setting of large J and small risk sets.

4 Inference in hazard odds models

4.1 Point estimation

To derive a point estimator for β^*, we rewrite model (2) as

$$P(Y = t_j, \delta = 1 | Y \geq t_j, X = x) = \text{expit}(\beta^*_{0j} + x^{\mathrm{T}} \beta^*), \quad j = 1, \ldots, J, \tag{17}$$

where $\text{expit}(c) = \mathrm{e}^c/(1+\mathrm{e}^c)$, $\beta^*_0 = (\beta^*_{01}, \ldots, \beta^*_{0J})^{\mathrm{T}}$ is a vector of unknown intercepts and β^* is as before. Our estimators for (β^*_0, β^*) are defined jointly as a solution $(\hat{\beta}_0, \hat{\beta})$ to

$$\sum_{i:Y_i \geq t_j} \left\{ D_{ji} - (1 - D_{ji}) \mathrm{e}^{\beta_{0j} + X_i^{\mathrm{T}} \beta} \right\} = 0, \quad j = 1, \ldots, J, \tag{18}$$

$$\sum_{j=1}^{J} \frac{\sum_{l:Y_l \geq t_j} (1 - D_{jl}) \mathrm{e}^{X_l^{\mathrm{T}} \beta}}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \beta}} \sum_{i:Y_i \geq t_j} \left\{ D_{ji} - (1 - D_{ji}) \mathrm{e}^{\beta_{0j} + X_i^{\mathrm{T}} \beta} \right\} X_i = 0, \tag{19}$$

where $D_{ji} = 1\{Y_i = t_j, \delta_i = 1\}$ as in Sect. 3. Similarly as (4)–(5), Eq. (18) depends only on the data from jth risk set $\{i : Y_i \geq t_j\}$, whereas Eq. (19) involves the data combined from all J risk sets. Within the jth risk set, the associated estimating functions in (β_{0j}, β) are

$$\begin{aligned} &\sum_{i:Y_i \geq t_j} \{D_{ji} - (1 - D_{ji}) \mathrm{e}^{\beta_{0j} + X_i^{\mathrm{T}} \beta}\} (1, X^{\mathrm{T}})^{\mathrm{T}} \\ &= \sum_{i:Y_i \geq t_j} \left\{ 1 - \frac{1 - D_{ji}}{\text{expit}(-\beta_{0j} - X_i^{\mathrm{T}} \beta)} \right\} (1, X^{\mathrm{T}})^{\mathrm{T}} \end{aligned}$$

which, interestingly, corresponds to the estimating functions for calibrated estimation (Tan 2020) in logistic regression model (17) for $1 - D_{ji}$ given X_i. The j-dependent factor $\{\sum_{l:Y_l \geq t_j} (1 - D_{jl}) \mathrm{e}^{X_l^{\mathrm{T}} \beta}\}/(\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \beta})$ in (19) is introduced to achieve reduction, as discussed below, to the weighted Mantel–Haenszel estimator in two-sample survival analysis in Tan (2022) and to the maximum partial likelihood estimator in the case of only one event per risk set in Cox's continuous-time model. In addition, use of this factor is crucial for achieving conditional unbiasedness as in (28) and (32).

Solving (18) for β_{0j} with fixed β and substituting into (19) shows that

$$\mathrm{e}^{\hat{\beta}_{0j}} = \frac{\sum_{i:Y_i \geq t_j} D_{ji}}{\sum_{i:Y_i \geq t_j} (1 - D_{ji}) \mathrm{e}^{X_i^{\mathrm{T}} \hat{\beta}}}. \tag{20}$$

and $\hat{\beta}$ can be determined from the closed-form estimating equation

$$\sum_{j=1}^{J} \sum_{i:Y_i \geq t_j} \frac{D_{ji} \sum_{l:Y_l \geq t_j} (1 - D_{jl}) \mathrm{e}^{X_l^{\mathrm{T}} \beta} - (1 - D_{ji}) \mathrm{e}^{X_i^{\mathrm{T}} \beta} \sum_{l:Y_l \geq t_j} D_{jl}}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \beta}} X_i = 0. \tag{21}$$

By some rearrangement, Eq. (21) can be equivalently written as

$$\sum_{j=1}^{J} \sum_{i:Y_i \geq t_j} D_{ji} \frac{\sum_{l:Y_l \geq t_j} (1 - D_{jl}) \mathrm{e}^{X_l^{\mathrm{T}} \beta} (X_i - X_l)}{\sum_{l:Y_l \geq t_j} \mathrm{e}^{X_l^{\mathrm{T}} \beta}} = 0, \tag{22}$$

which closely resembles the Breslow–Peto estimating Eq. (8) with only the additional factor $1 - D_{jl}$ in front of $\mathrm{e}^{X_i^{\mathrm{T}} \beta}(X_i - X_l)$. In fact, in the extreme case of only one event observed in each risk set $\{i : Y_i \geq t_j\}$ for $j = 1, \ldots, J$, then Eq. (22) is easily shown to be equivalent to (8), and hence the estimator $\hat{\beta}$ numerically coincides with the Breslow–Peto or the maximum partial likelihood estimator $\hat{\gamma}$.

For two-sample survival analysis with a binary covariate X, estimating Eq. (21) or (22) can be shown to yield the weighted Mantel–Haenszel estimator proposed in Tan (2022) as an extension of Cochran's (1954) and Mantel & Haenszel's (1959) estimation of common odds ratios in analysis of 2×2 tables. See the Supplement for details. Hence the estimator $\hat{\beta}$ can also be called a weighted Mantel–Haenszel estimator.

4.2 Model-robust inference

We study model-robust inference using $\hat{\beta}$ with possible misspecification of model (17), similarly as in Sect. 3 for robust inference using $\hat{\gamma}$ in the hazard probability model. Denote, as before, $R_{ji} = 1\{Y_i \geq t_j\}$ and $D_{ji} = 1\{Y_i = t_j, \delta_i = 1\}$. Estimating Eq. (21) can be written as $\sum_{j=1}^{J} \hat{\tau}_j(\beta) = 0$, where

$$\hat{\tau}_j(\beta) = \frac{1}{n} \sum_{i=1}^{n} R_{ji} \frac{D_{ji} \sum_{l=1}^{n} R_{jl}(1 - D_{jl}) \mathrm{e}^{X_l^{\mathrm{T}} \beta} - (1 - D_{ji}) \mathrm{e}^{X_i^{\mathrm{T}} \beta} \sum_{l=1}^{n} R_{jl} D_{jl}}{\sum_{l=1}^{n} R_{jl} \mathrm{e}^{X_l^{\mathrm{T}} \beta}} X_i .$$

Under suitable regularity conditions, it can be shown that $\hat{\beta}$ converges in probability to a target value $\bar{\beta}$, defined as a unique solution to the population version of (21) or equivalently (22):

$$0 = \sum_{j=1}^{J} E\left[R_j \frac{D_j E\{\tilde{R}_j (1 - \tilde{D}_j) \mathrm{e}^{\tilde{X}^{\mathrm{T}} \beta}\} - (1 - D_j) \mathrm{e}^{X^{\mathrm{T}} \beta} E(\tilde{R}_j \tilde{D}_j)}{E(\tilde{R}_j \mathrm{e}^{\tilde{X}^{\mathrm{T}} \beta})} X \right] \tag{23}$$

$$= \sum_{j=1}^{J} E\left[R_j D_j \frac{E\{\tilde{R}_j (1 - \tilde{D}_j) \mathrm{e}^{\tilde{X}^{\mathrm{T}} \beta}\} X - E\{\tilde{R}_j (1 - \tilde{D}_j) \mathrm{e}^{\tilde{X}^{\mathrm{T}} \beta} \tilde{X}\}}{E(\tilde{R}_j \mathrm{e}^{\tilde{X}^{\mathrm{T}} \beta})} \right], \tag{24}$$

where, as in Sect. 4, $R_j = 1\{Y \geq t_j\}$, $D_j = 1\{Y = t_j, \delta = 1\}$, and $(\tilde{R}_j, \tilde{D}_j, \tilde{X})$ are defined from $(\tilde{Y}, \tilde{\delta}, \tilde{X})$ identically distributed as (Y, δ, X). Moreover, $\hat{\beta}$ can be shown to admit the asymptotic expansion

$$\hat{\beta} - \bar{\beta} = H(\bar{\beta})^{-1} \sum_{j=1}^{J} \hat{\tau}_j(\bar{\beta}) + o_p(n^{-1/2}), \tag{25}$$

where $H(\beta)$ is the negative derivative matrix in β^{T} of the right hand side of (23) or equivalently (24), that is,

$$H(\beta) = \sum_{j=1}^{J} E\left[\frac{R_j(1-D_j)\mathrm{e}^{X^{\mathrm{T}}\beta}}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}\left\{E(\tilde{R}_j\tilde{D}_j)X - E(\tilde{R}_j\tilde{D}_j\tilde{X})\right\}\right.$$
$$\left.\left\{X^{\mathrm{T}} - \frac{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}\tilde{X}^{\mathrm{T}})}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}\right\}\right].$$

See the Supplement for details. The matrix $H(\beta)$ is in general not symmetric, and hence cannot be an Hessian of a scalar objective function. From (25), the following result can be deduced, provided that the probability of survival beyond time t_J (which is the largest possible value of the censoring variable) is bounded away from 0.

Proposition 3 *Assume that $P(T > t_J) \geq p_0$ for a constant $p_0 > 0$. Then $n^{1/2}(\hat{\beta} - \bar{\beta})$ converges in distribution to $\mathrm{N}(0, \Sigma)$ as $n \to \infty$, where $\Sigma = H(\bar{\beta})^{-1}G(\bar{\beta})H(\bar{\beta})^{\mathrm{T}^{-1}}$, $H(\beta)$ is defined as above, $G(\beta) = \mathrm{var}\{\sum_{j=1}^{J} g_j(Y, \delta, X; \beta)\}$, and*

$$g_j(Y, \delta, X; \beta)$$
$$= R_j \frac{D_j E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}) - (1-D_j)\mathrm{e}^{X^{\mathrm{T}}\beta}E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}$$
$$\left\{X - \frac{E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}\tilde{X})}{E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}\right\}$$
$$- \frac{E(\tilde{R}_j\tilde{D}_j\tilde{X})E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}) - E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}\tilde{X})E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}$$
$$\times \left\{\frac{R_j\mathrm{e}^{X^{\mathrm{T}}\beta}}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})} - \frac{R_j(1-D_j)\mathrm{e}^{X^{\mathrm{T}}\beta}}{E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta})}\right\},$$

which is denoted as $g_{j1}(Y, \delta, X; \beta) + g_{j2}(Y, \delta, X; \beta)$. Moreover, a consistent estimator of Σ is $\hat{\Sigma}_r = \hat{H}^{-1}(\hat{\beta})\hat{G}(\hat{\beta})\hat{H}^{\mathrm{T}^{-1}}(\hat{\beta})$, where

$$\hat{H}(\beta) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\left[\frac{R_{ji}(1-D_{ji})\mathrm{e}^{X_i^{\mathrm{T}}\beta}}{\sum_{l=1}^{n} R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta}}\left\{\sum_{l=1}^{n} R_{jl}D_{jl}(X_i - X_l)\right\}\right.$$
$$\left.\left\{X_i^{\mathrm{T}} - \frac{\sum_{l=1}^{n} R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta}X_l^{\mathrm{T}}}{\sum_{l=1}^{n} R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta}}\right\}\right],$$

$$\hat{G}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{J}\hat{g}_j(Y_i, \delta_i, X_i; \beta)\right\}^{\otimes 2},$$

and $\hat{g}_j(Y, \delta, X; \beta)$ *is defined as* $g_j(Y, \delta, X; \beta)$ *with all expectations replaced by the corresponding sample averages.*

From Proposition 3, the influence function of $\hat{\beta}$ is $H^{-1}(\bar{\beta})\sum_{j=1}^{J} g_j(Y, \delta, X; \bar{\beta})$. Here $g_j(Y, \delta, X; \beta)$ consists of two terms. The first term, $g_{j1}(Y, \delta, X; \beta)$, can be seen as a correction to the jth population estimating function in (23), to account for the variation in substituting the estimator $\mathrm{e}^{\hat{\beta}_{0j}}$ for $E(\tilde{R}_j\tilde{D}_j)/E\{\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta}\}$ in the sample estimating Eq. (21). The second term, $g_{j2}(Y, \delta, X; \beta)$, is involved to further account for substituting the factor $\{\sum_{l=1}^{n} R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}\}/(\sum_{l=1}^{n} R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta})$ for the corresponding population quantity. A similar interpretation of $g_j(Y, \delta, X; \beta)$ can also be obtained as a correction to the jth population estimating function in (24).

4.3 Model-based inference

We study model-based inference using $\hat{\beta}$ when model (17) is correctly specified. Under this assumption, $\hat{\beta}$ is a consistent estimator of β^*, with $\bar{\beta} = \beta^*$ satisfying the population estimating Eq. (23):

$$E\left[R_j\frac{D_jE\{\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*}\} - (1-D_j)\mathrm{e}^{X^{\mathrm{T}}\beta^*}E(\tilde{R}_j\tilde{D}_j)}{E(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*})}X\right] = 0, \tag{26}$$

because $E[R_j\{D_j - (1-D_j)\mathrm{e}^{\beta^*_{0j}+X^{\mathrm{T}}\beta^*}\}|X] = 0$ and $\mathrm{e}^{\beta^*_{0j}} = E(\tilde{R}_j\tilde{D}_j)/E\{\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*}\}$ by (17). Equivalently, the true value β^* also satisfies Eq. (24).

Considerable simplification can be obtained for the model-based asymptotic variance for $\hat{\beta}$ in Proposition 3. Under model (17), $g_j(Y, \delta, X; \beta^*)$ reduces to $g_{j1}(Y, \delta, X; \beta^*)$ only, because $g_{j2}(Y, \delta, X; \beta^*) \equiv 0$ due to (26). Moreover, the difference $D_j - (1-D_j)\mathrm{e}^{\beta^*_{0j}+X^{\mathrm{T}}\beta^*}$ has mean 0 conditionally on $R_j = 1$ and X, and the individual terms $g_j(Y, \delta, X; \beta^*)$, $j = 1, \ldots, J$, are uncorrelated with each other. Then the asymptotic variance $\Sigma = H^{-1}(\beta^*)G(\beta^*)H^{\mathrm{T}^{-1}}(\beta^*)$ can be calculated such that

$$\begin{aligned} G(\beta^*) &= \sum_{j=1}^{J}\mathrm{var}\{g_{j1}(Y, \delta, X; \beta^*)\} \\ &= \sum_{j=1}^{J} E\left[R_j\mathrm{e}^{\beta^*_{0j}+X^{\mathrm{T}}\beta^*}\frac{E^2(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*})}{E^2(\tilde{R}_j\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*})}\left\{X - \frac{E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*}\tilde{X})}{E(\tilde{R}_j(1-\tilde{D}_j)\mathrm{e}^{\tilde{X}^{\mathrm{T}}\beta^*})}\right\}^{\otimes 2}\right]. \end{aligned} \tag{27}$$

See the Supplement for details. A model-based estimator for the asymptotic variance Σ is then $\hat{\Sigma}_{\mathrm{b}} = \hat{H}^{-1}(\hat{\beta})\hat{G}_{\mathrm{b}}(\hat{\beta})\hat{H}^{-1}(\hat{\beta})$, where $\hat{G}_{\mathrm{b}}(\beta)$ is defined as

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\left[R_{ji}\mathrm{e}^{\beta_{0j}+X_i^{\mathrm{T}}\beta}\frac{(\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta})^2}{(\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta})^2}\right.$$
$$\left.\left\{X_i-\frac{\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}X_l}{\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}}\right\}^{\otimes 2}\right],$$

with $\mathrm{e}^{\beta_{0j}}$ set to $(\sum_{l=1}^{n}R_{jl}D_{jl})/\{\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}\}$. The matrix $\hat{G}_{\mathrm{b}}(\beta)$ is algebraically similar to the sample Hessian $\hat{B}(\gamma)$ in Sect. 3, with only the additional factor $1-D_{jl}$ in front of $\mathrm{e}^{X_l^{\mathrm{T}}\beta}$ in various places.

Similarly as in Sect. 3, we outline asymptotic theory conditionally on the risk sets and covariates and propose an improved model-based variance estimator. To accommodate small risk sets, asymptotic properties of $\hat{\beta}$ can be studied by exploiting the conditional unbiasedness of individual terms of the sample estimating function in (21) under model (17):

$$E\left\{\hat{\tau}_j(\beta^*)|R_{j,1:n},X_{1:n}\right\}=0,\quad j=1,\ldots,J, \tag{28}$$

where $R_{j,1:n}=(R_{j1},\ldots,R_{jn})$ and $X_{1:n}=(X_1,\ldots,X_n)$. This is a more elaborate property than unconditional unbiasedness (26). Under suitable regularity conditions similar as in fixed-design analysis of regression models, it can be shown that if model (17) is correctly specified, then $n^{1/2}(\hat{\beta}-\beta^*)$ converges in distribution to $\mathrm{N}(0,\Sigma_2)$ as $n\to\infty$, where $\Sigma_2=H_2(\beta^*)^{-1}G_2(\beta^*)H_2(\beta^*)^{-1}$, $H_2(\beta)=\mathrm{plim}_{n\to\infty}\hat{H}(\beta)$, $G_2(\beta)=\mathrm{plim}_{n\to\infty}\sum_{j=1}^{J}\sigma_j(\beta)$, and $\sigma_j(\beta)=n\,\mathrm{var}\{\hat{\tau}_j(\beta)|R_{j,1:n},X_{1:n}\}$, that is,

$$\frac{1}{n}\mathrm{var}\left\{\sum_{i=1}^{n}R_{ji}\frac{D_{ji}\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}-(1-D_{ji})\mathrm{e}^{X_i^{\mathrm{T}}\beta}\sum_{l=1}^{n}R_{jl}D_{jl}}{\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta}}X_i\middle|R_{j,1:n},X_{1:n}\right\}.$$

In the case where $P(T>t_J)$ is bounded away from 0 and all J risk sets are of sizes increasing to ∞, the asymptotic variance Σ_2 reduces to Σ in Proposition 3.

For the asymptotic variance Σ_2, our proposed estimator is $\hat{\Sigma}_{\mathrm{b}2}=\hat{H}^{-1}(\hat{\beta})\hat{G}_{\mathrm{b}2}(\hat{\beta})\hat{H}^{-1}(\hat{\beta})$, where $\hat{H}(\beta)$ is as in Proposition 3, $\hat{G}_{\mathrm{b}2}(\beta)=\sum_{j=1}^{J}\{\hat{\sigma}_j(\beta)+\hat{\sigma}_j^{\mathrm{T}}(\beta)\}/2$, and

$$\hat{\sigma}_j(\beta)=\frac{1}{n}\sum_{i=1}^{n}\left\{R_{ji}(1-D_{ji})\mathrm{e}^{X_i^{\mathrm{T}}\beta}\frac{\sum_{l=1}^{n}R_{jl}D_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta}(X_i-X_l)^{\otimes 2}}{(\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta})^2}\right.$$
$$\left.+R_{ji}\mathrm{e}^{X_i^{\mathrm{T}}\beta}\frac{\sum_{l=1}^{n}R_{jl}(1-D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta}(X_i-X_l)\sum_{k=1}^{n}R_{jk}D_{jk}(X_i-X_k)^{\mathrm{T}}}{(\sum_{l=1}^{n}R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta})^2}\right\}. \tag{29}$$

The matrix $\hat{\sigma}_j(\beta)$ is in general not symmetric, and $\{\hat{\sigma}_j(\beta)+\hat{\sigma}_j^{\mathrm{T}}(\beta)\}/2$ serves as a symmetrized version. The following properties can be established.

Proposition 4 *(i) Suppose that model (17) is correctly specified. For $j = 1, \ldots, J$, $\hat{\sigma}_j(\beta^*)$ is conditionally unbiased for $\sigma_j(\beta^*)$, that is,*

$$E\{\hat{\sigma}_j(\beta^*)|R_{j,1:n}, X_{1:n}\} = \sigma_j(\beta^*). \tag{30}$$

Hence $\hat{\Sigma}_{b2}$ can be a consistent estimator for Σ_2 even if some risk sets are of sizes which are bounded in probability as $J \to \infty$ and $n \to \infty$.
(ii) Suppose that at most one event is observed in each risk set $\{i : R_{ji} = 1\}$ for $j = 1, \ldots, J$. Then $\hat{\beta}$ is identical to the maximum partial likelihood estimator, and $\hat{H}(\beta)$ and $\hat{G}_{b2}(\beta)$ are both identical to $\hat{B}(\beta)$. Hence $\hat{\Sigma}_{b2}$ is identical to $\hat{B}^{-1}(\hat{\beta})$, the usual variance estimator for the maximum partial likelihood estimator.

The variance estimator $\hat{\sigma}_j(\hat{\beta})$ and the resulting sandwich variance $\hat{\Sigma}_{b2}$ represent a new development beyond model-based variance estimation in Tan (2022) for the weighted Mantel–Haenszel estimator in an odds ratio model for analysis of 2×2 tables and two-sample survival analysis. The model-based variance estimator in Tan (2022) is adapted from that in Robins et al. (1986) for the Mantel–Haenszel estimator of a common odds ratio in 2×2 tables, such that the variance estimator is consistent in both asymptotic settings of large tables and many sparse tables. For two-sample analysis, the proposed estimator $\hat{\Sigma}_{b2}$ reduces to a variance estimator distinct from that in Robins et al. (1986) as well as in Flander (1985). See the Supplement for details.

For comparison, a suitable extension of model-based variance estimation from Robins et al. (1986) and Tan (2022) to regression models is $\hat{\Sigma}_{b3} = \hat{H}^{-1}(\hat{\beta})\hat{G}_{b3}(\hat{\beta}) \hat{H}^{-1}(\hat{\beta})$, where $\hat{H}(\beta)$ is as in Proposition 3, $\hat{G}_{b3}(\beta) = \sum_{j=1}^{J} \tilde{\sigma}_j(\beta)$, and

$$\tilde{\sigma}_j(\beta) = \frac{1}{n}\sum_{i=1}^{n} R_{ji}(1 - D_{ji})\mathrm{e}^{X_i^{\mathrm{T}}\beta} \times \\ \frac{\sum_{l=1}^{n} R_{jl}\{(1 - D_{jl})\mathrm{e}^{X_l^{\mathrm{T}}\beta} + D_{jl}\mathrm{e}^{X_i^{\mathrm{T}}\beta}\}(X_i - X_l)\sum_{k=1}^{n} R_{jk}D_{jk}(X_i - X_k)^{\mathrm{T}}}{(\sum_{l=1}^{n} R_{jl}\mathrm{e}^{X_l^{\mathrm{T}}\beta})^2}. \tag{31}$$

Although not apparent from the above definition, $\hat{\sigma}_j(\beta)$ can be equivalently expressed as a symmetric, nonnegative-definite matrix. Moreover, $\tilde{\sigma}_j(\beta^*)$ can be shown to be conditionally unbiased for $\sigma_j(\beta^*)$, i.e., $E\{\hat{\sigma}_j(\beta^*)|R_{j,1:n}, X_{1:n}\} = \sigma_j(\beta^*)$. See the Supplement for details. However, in contrast with Proposition 4(ii), the sandwich variance $\hat{\Sigma}_{b3}$ does not automatically reduce to $\hat{B}^{-1}(\hat{\beta})$, the usual variance estimator for the maximum partial likelihood estimator, in the special case of no tied events. A possible explanation is that $\hat{\sigma}_j(\beta)$ involves only two-way products of the event indicators D_{ji}, whereas $\tilde{\sigma}_j(\beta)$ involves three-way products of the event indicators.

4.4 Conditional inference given numbers of events

For odds ratio model (17), i.e., Cox's (1972) discrete-time propositional hazards model, a common approach for eliminating the nuisance parameters $(\beta_{01}, \ldots, \beta_{0J})$

is to perform likelihood inference successively conditionally on the numbers of events $(S_1, \ldots, S_J)$, in addition to the risk-set indicators and covariates, where $S_j = \sum_{i=1}^n R_{ji} D_{ji}$. This approach is theoretically desirable (e.g., Lindsay 1980, Lindsay 1983), but numerical implementation is intractable with a relatively large number of tied events. Remarkably, we show that, given both the numbers of events and the risk-set indicators and covariates, not only the individual terms, $\hat{\tau}_j(\beta)$, in the weighted Mantel–Haenszel estimating function are conditionally unbiased, but also the variance estimators $\hat{\sigma}_j(\beta)$ evaluated at β^* are conditionally unbiased.

Proposition 5 *Suppose that model (17) is correctly specified. For $j = 1, \ldots, J$, each individual term $\hat{\tau}_j(\beta)$ is conditionally unbiased given S_j:*

$$E\{\hat{\tau}_j(\beta^*)|S_j, R_{j,1:n}, X_{1:n}\} = 0, \tag{32}$$

where $S_j = \sum_{i=1}^n R_{ji} D_{ji}$, $R_{j,1:n} = (R_{j1}, \ldots, R_{jn})$, and $X_{1:n} = (X_1, \ldots, X_n)$. Moreover, $\hat{\sigma}_j(\beta^*)$ is conditionally unbiased for the conditional variance of $n^{1/2}\hat{\tau}_j(\beta^*)$:

$$E\{\hat{\sigma}_j(\beta^*)|S_j, R_{j,1:n}, X_{1:n}\} = n\,\mathrm{var}\{\hat{\tau}_j(\beta^*)|S_j, R_{j,1:n}, X_{1:n}\}. \tag{33}$$

There are two types of conditional unbiasedness, depending on whether the risk-set indicators and covariates are conditioned on or the number of events is further conditioned on. See (28) versus (32) for point estimation and (30) and (33) for variance estimation. Based on Proposition 5, we expect that under suitable regularity conditions, the point estimator $\hat{\beta}$ is consistent for β^*, and $n^{1/2}(\hat{\beta} - \beta^*)$ is asymptotically normal with mean 0 and a variance matrix consistently estimated by the sandwich variance estimator $\hat{\Sigma}_{b2}$, while conditioning on the number of events $(S_1, \ldots, S_J)$. Large sample theory along this direction can be studied in future work.

Conditional unbiasedness given numbers of events, similar to (32), is known to be satisfied by the Mantel–Haenszel estimating function for a common odds ratio in 2×2 tables (Breslow 1981). In that setting, conditional unbiasedness similar to (33) is also established for the variance estimator in Robins et al. (1986). In fact, similarly to $\hat{\sigma}_j(\beta)$, the variance estimator $\tilde{\sigma}_j(\beta)$ in (31) as an extension of Robins et al. (1986) can also be shown to be conditionally unbiased, that is, $E\{\tilde{\sigma}_j(\beta^*)|S_j, R_{j,1:n}, X_{1:n}\} = n\,\mathrm{var}\{\hat{\tau}_j(\beta^*)|S_j, R_{j,1:n}, X_{1:n}\}$. Nevertheless, the variance estimator $\hat{\sigma}_j(\beta)$ enjoys an exact reduction in the case of no tied events: if $S_j = 1$, then $\hat{\sigma}_j(\beta^*) = n\,\mathrm{var}\{\hat{\tau}_j(\beta^*)|S_j = 1, R_{j,1:n}, X_{1:n}\}$, not just in expectation, by Proposition 4(ii) and the fact that the sample Hessian $\hat{B}(\beta^*)$ is equal to $n\,\mathrm{var}\{\hat{\tau}_j(\beta^*)|S_j = 1, R_{j,1:n}, X_{1:n}\}$.

4.5 Estimation of survival probabilities

Similarly as in Sect. 3, we discuss estimation of survival probabilities for individuals with fixed covariates x_0. For simplicity, assume that $x_0 = 0$ in model (2). Then the hazard probability $p_j(x_0)$ is identified as $\mathrm{expit}(\beta_{0j})$, and can be estimated from (20) as

$$\hat{q}_j(x_0) = \text{expit}(\hat{\beta}_0 j) = \frac{\sum_{i=1}^n R_{ji} D_{ji}}{\sum_{i=1}^n R_{ji} D_{ji} + R_{ji}(1 - D_{ji}) \mathrm{e}^{X_i^{\mathrm{T}} \hat{\beta}}}$$

The kth survival probability, $P_k(x_0) = P(T > t_k | X = x_0)$, can be estimated as

$$\hat{Q}_k(x_0) = \prod_{j=1}^{k} \{1 - \hat{q}_j(x_0)\}, \quad k = 1, \ldots, J. \tag{34}$$

The estimators $\hat{q}_j(x_0)$ and $\hat{Q}_k(x_0)$ for $p_j(x_0)$ and $P_k(x_0)$ are automatically restricted to between 0 and 1, in contrast with $\hat{p}_j(x_0)$ and $\hat{P}_k(x_0)$ in Sect. 3. The cumulative hazard probability, $\sum_{j=1}^k p_j(x_0)$, can be estimated as $\sum_{j=1}^k \hat{q}_k(x_0)$.

The standard errors for $\hat{Q}_k(x_0)$ can be obtained using Taylor expansions (or the delta method) and either model-robust or model-based variance estimator for $\hat{\beta}$. See the Supplement for detailed derivation and formulas, assuming that all J risk sets are sufficiently large. Further investigation is needed, including formal justification of our formulas in the setting of large J and small risk sets.

5 Comparison and extension

5.1 Pooled logistic regression

For odds ratio model (17), i.e., Cox's (1972) discrete-time propositional hazard model, conditional likelihood inference given numbers of events is usually considered statistically superior while exact solution can be numerically challenging. For completeness, it is helpful to discuss another existing approach which directly uses maximum likelihood estimation over the main parameter β and nuisance parameters $(\beta_{01}, \ldots, \beta_{0J})$ in model (17) (e.g., Allison 1982). The estimators, $\tilde{\beta}$ and $(\tilde{\beta}_{01}, \ldots, \tilde{\beta}_{0J})$, are defined jointly as a maximizer to the log likelihood function

$$\sum_{j=1}^{J} \sum_{i=1}^{n} R_{ji} \left\{ D_{ji}(\beta_{0j} + X_i^{\mathrm{T}} \beta) - \log\left(1 + \mathrm{e}^{\beta_{0j} + X_i^{\mathrm{T}} \beta}\right) \right\}.$$

Equivalently, $\tilde{\beta}$ and $(\tilde{\beta}_{01}, \ldots, \tilde{\beta}_{0J})$ are determined jointly as a solution to

$$\sum_{i=1}^{n} R_{ji} \left\{ D_{ji} - \text{expit}(\beta_{0j} + X_i^{\mathrm{T}} \beta) \right\} = 0, \quad j = 1, \ldots, J, \tag{35}$$

$$\sum_{j=1}^{J} \sum_{i=1}^{n} R_{ji} \left\{ D_{ji} - \text{expit}(\beta_{0j} + X_i^{\mathrm{T}} \beta) \right\} X_i = 0. \tag{36}$$

This approach can be called pooled logistic regression, formally the same as fitting J logistic regression models with a common coefficient vector β across individual datasets. On one hand, the estimating Eqs. (35)–(36) are seemingly similar to estimating Eqs. (18)–(19) for weighted Mantel–Haenszel estimation in model (17), as well as (4)–(5) for Breslow–Peto estimation in model (3). On the other hand, there are fundamental differences between these methods which we explain as follows.

An easy difference is that closed-form solutions for $(\gamma_{01}, \ldots, \gamma_{0J})$ from (4) with fixed γ or for $(\beta_{01}, \ldots, \beta_{0J})$ from (18) with fixed β can be derived, whereas such a closed-form solution is not available from Eq. (35). A deeper difference is that, to borrow the terminology of profile likelihood, the profile estimating Eq. (7) in γ is conditionally unbiased according to (15), and the profile estimating Eq. (21) in β is conditionally unbiased according to (28), both given the risk sets and covariates. A profile estimating equation in β can also be defined from Eqs. (35)–(36), in spite of no closed-form solution for $(\beta_{01}, \ldots, \beta_{0J})$. But this estimating equation in β does not satisfy conditional unbiasedness in a similar manner as (15) or (28). Finally, the profile estimating Eq. (21) in β is also conditionally unbiased according to (32), given the numbers of events in addition to the risk sets and covariates. This unbiasedness is shared by the conditional score equation in the approach of conditional likelihood inference. For these reasons, weighted Mantel–Haeszel estimation is expected to achieve superior finite-sample performance, similarly as conditional likelihood estimation, over pooled logistic regression, in particular with a large number of time points J.

The preceding discussion also explains that pooled logistic regression can be problematic in fitting model (17) with finely discretized data in finite samples, which is in agreement with the understanding that maximum likelihood estimation with a large number of nuisance parameters may not generally be desirable.

5.2 Time-varying coefficient and time-dependent covariates

Our theory and methods are so far developed in the context of models (1) and (2), with time-independent regression coefficients and time-independent covariates. Nevertheless, the development can be readily extended to handle time-varying coefficients and time-dependent covariates, similarly as in Cox's continuous-time proportional hazards models. First, consider an extension of models (1) and (2), where $p_j(x)$ is redefined as

$$p_j(x) = P(Y = t_j, \delta = 1 | Y \geq t_j, X(t_j) = x),$$

where $X(t_j)$ is the covariate vector at time t_j. Then estimating Eqs. (7) for $\hat{\gamma}$ and (21) for $\hat{\beta}$ can be extended by replacing X_i with $X_i(t_j)$ within the jth risk set. Similar modification can be applied to the model-based and model-robust variance estimators. Next, time-varying coefficients can be accommodated by a reformulation using time-dependent covariates. For example, consider model (1) extended with a time-varying coefficient for a scalar covariate $x^{(1)}$:

$$p_j(x) = p_j(x_0) \exp\left\{x^{(1)} b(t_j; \gamma_{11}^*, \gamma_{12}^*) + x^{(2)\mathrm{T}} \gamma_2^*\right\}, \tag{37}$$

where $x_0 = 0$, $x = (x^{(1)}, x^{(2)\mathrm{T}})^{\mathrm{T}}$, $b(t; \gamma_{11}, \gamma_{12})$ is a function of time, defined as $\gamma_1 + u^{\mathrm{T}}(t)\gamma_{12}$ using a basic vector $u(t)$, and $\gamma^* = (\gamma_{11}^*, \gamma_{12}^{*\mathrm{T}}, \gamma_2^{*\mathrm{T}})^{\mathrm{T}}$ are unknown coefficients. Model (37) can be put in the form of (1), where x is replaced by the time-dependent covariate vector $(x^{(1)}, x^{(1)}u^{\mathrm{T}}(t_j), x^{(2)\mathrm{T}})^{\mathrm{T}}$ at time t_j associated with the coefficient vector γ^*.

6 Numerical studies

6.1 Analysis of veteran's lung cancer data

We compare different methods in analysis of the data on a Veteran's Administration lung cancer trial used in Kalbflwisch & Prentice (1980). The trial included 137 male patients with advanced lung cancer. The outcome of interest is time to death in days, and there are six covariates measured at randomization: treatment (test or standard), age in years, Karnofsky score (ranged 10 to 99), time in months from diagnosis to the start of treatment, cell type (a nominal factor of 4 levels), and prior therapy (yes or no). The corresponding regression terms are denoted as `treat`, `age`, `Karn`, `diagt`, `cell2`, `cell3`, `cell4` (for the contrasts between levels 2–4 versus 1), and `prior`. As the dataset involves a (small) nonempty set of tied events, Cox's proportional hazards model is applied with some approximations, for example, indicated by the option `ties` in R package `survival`.

Kaplan–Meier survival curves suggest non-proportional hazards over time in the two treatment groups, while ignoring other covariates (Tan 2022, Supplement). Hence we fit hazard probability and odds models by allowing time-varying coefficients with the treatment variable. As discussed in Sect. 5, such models can be stated using time-dependent covariates (or regression terms), defined as functions of the time and treatment variables, `time` and `treat`. For simplicity, we include two time-dependent regression terms, `treat2` and `treat3`, defined as `treat` $*\, 1\{$`time` $> 100\}$ and `treat` $*\, 1\{$`time` $> 200\}$. The coefficients for these two terms represent changes after day 100 or 200 in the association of the test treatment with hazard probabilities or odds.

To study discrete-time inference, we also apply various methods to further discretized data, obtained by grouping the original times in intervals of 20 days. For concreteness, the censored-late option is used as mentioned in Sect. 2. An uncensored time in $(t_{j-1}, t_j]$ is labeled t_j, whereas a censored time in $[t_{j-1}, t_j)$ is labeled t_j. The censoring indicator is kept unchanged. See Tan (2022, Supplement) for more details.

Tables 1–2 present the results on the original and discretized data. For the original data with a small number of tied deaths, the estimates of BP, Efron, CML, and wMH are similar to each other in various degrees, although the BP point estimates associated with probability ratios are consistently closer to 0 than those of CML and wMH associated with odds ratios, except for the coefficient of `diagt` which is the least accurately estimated as measured by the t-statistic. The Plogit point estimates show noticeable differences (or biases) from those of CML and wMH.

For the discretized data with more tied deaths, the BP point estimates are more substantially closer to 0 than those of CML and wMH, which remain similar to each

Table 1 Analysis of veteran's lung cancer data (original)

	BP	Efron	CML	wMH	Plogit		BP	Efron	CML	wMH	Plogit
	Point estimate						Point estimate				
treat	.379	.383	.384	.383	.392	diagt	−.064	−.047	−.060	−.038	−.080
treat2	−.493	−.494	−.498	−.494	−.511	cell2	.830	.835	.836	.830	.865
treat3	.472	.476	.483	.475	.437	cell3	1.152	1.161	1.167	1.167	1.196
age	−.813	−.829	−.813	−.838	−.804	cell4	.372	.374	.376	.376	.385
Karn	−.320	−.322	−.324	−.323	−.334	prior	.083	.083	.084	.087	.082

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	CML	wMH	Plogit
	Model-based SE							Model-robust SE			
treat	.245	.243	.245	.246	.247	.248	.221	.223	–	.224	.227
treat2	.516	.515	.516	.518	.515	.524	.481	.484	–	.482	.496
treat3	.645	.645	.646	.647	.644	.670	.622	.624	–	.622	.662
age	.931	.927	.930	.937	.930	.954	1.029	1.036	–	1.035	1.082
Karn	.056	.056	.056	.057	.056	.058	.053	.054	–	.054	.057
diagt	.918	.897	.919	.930	.947	.945	.790	.790	–	.800	.833
cell2	.283	.282	.283	.284	.284	.288	.306	.309	–	.310	.321
cell3	.313	.311	.313	.315	.315	.319	.273	.275	–	.277	.284
cell4	.292	.291	.292	.293	.292	.297	.247	.248	–	.248	.258
prior	.232	.231	.232	.233	.234	.238	.217	.219	–	.220	.226

Note: BP, wMH, or Plogit denotes Breslow–Peto estimator $\hat{\gamma}$, weighted Mantel–Haenszel estimator $\hat{\beta}$, or pooled logistic estimator $\tilde{\beta}$, implemented by the R package `dSurvival` (Tan 2020b). oldBP, Efron, or CML denotes results from Cox's regression `coxph` with `ties="breslow"`, `"efron"`, or `"exact"` in the R package `survival` (Therneau 2015). oldBP and BP are identical to each other in point estimates and model-robust SEs. The point estimates and SEs for `age`, `Karn`, and `diagt` are reported after multiplied by 10^2, 10, and 10^2 respectively

Table 2 Analysis of veteran's lung cancer data (discretized)

	BP	Efron	CML	wMH	Plogit		BP	Efron	CML	wMH	Plogit
	Point estimate						Point estimate				
treat	.307	.346	.415	.420	.422	diagt	−.007	−.129	.048	.040	.034
treat2	−.476	−.463	−.567	−.484	−.581	cell2	.778	.859	.926	.916	.955
treat3	.419	.437	.546	.406	.507	cell3	1.047	1.159	1.365	1.382	1.393
age	−.459	−.744	−.362	−.754	−.343	cell4	.366	.408	.464	.517	.473
Karn	−.267	−.310	−.358	−.337	−.368	prior	.053	.103	.061	.079	.056

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	CML	wMH	Plogit
	Model-based SE							Model-robust SE			
treat	.241	.204	.244	.273	.305	.275	.191	.219	–	.264	.244
treat2	.514	.473	.516	.554	.570	.561	.452	.488	–	.528	.533
treat3	.645	.611	.645	.684	.694	.707	.600	.628	–	.669	.709
age	.920	.794	.923	1.055	1.087	1.072	.924	1.039	–	1.216	1.250
Karn	.054	.047	.055	.066	.063	.067	.046	.053	–	.060	.068
diagt	.925	.746	.931	1.192	1.173	1.205	.704	.786	–	.925	1.139
cell2	.279	.250	.281	.306	.327	.310	.270	.300	–	.348	.342
cell3	.309	.269	.312	.351	.375	.355	.236	.265	–	.302	.291
cell4	.291	.270	.292	.311	.324	.315	.224	.243	–	.261	.264
prior	.232	.205	.234	.257	.272	.262	.196	.223	–	.247	.248

other at least for coefficients with relatively large t-statistics. This difference can be properly explained by the fact that BP estimates are associated with odds ratios, whereas the CML and wMH estimates are associated with probability ratios. In addition, in a more pronounced manner than in Table 1, the commonly reported variance estimates in the column oldBP are inflated compared with the proposed variance estimates in the column BP, as expected by Corollary 1. For example, for the coefficient of `treat`, the BP point estimate is smaller than CML by $1 - .307/.415 = 26.0\%$, and the oldBP variance estimate is larger than the proposed BP variance estimate by $(.241/.204)^2 - 1 = 39.6\%$. The Efron estimates tend to fall between BP and CML estimates. The Plogit point estimates still show various differences from those of CML and wMH.

For illustration, Fig. 2 shows the estimated survival probabilities using the BP and wMH methods with the discretized data, for individuals in the test or standard treatment group and with certain fixed covariate values. As allowed by the specified models, the test treatment compared with the standard treatment is associated with increasingly lower survival probabilities over time before day 100 or after day 200, while the trend is reversed between day 100 and 200. The BP estimate of the last survival probability is negative, a possibility mentioned in Sect. 3. This also reflects the fact that such estimates in the right tail are usually inaccurate.

6.2 Simulation study

To further compare different methods, we also conduct simulation studies. The first study, reported below, involves simulated data satisfying proportional hazards in continuous time, whereas the second study, reported in the Supplement, involves simulated data where proportional hazards are violated even in continuous time. Additional results are also provided in the Supplement, based on simulation settings with increased censoring fractions. Similar conclusions can be drawn as discussed below.

For each simulation, a sample of size $n = 200$ is generated as follows, mimicking a randomized trial. The treatment variable `Tr` is generated as 1 (test) or 2 (standard) with probabilities .5 each, and four covariates `X1`–`X4` are generated, independently of `Tr`, as multivariate normal with means 0 and covariances $2^{-|j-k|}$ between jth and kth covariates for $1 \le j, k \le 4$. The event time $\tilde{T}$ is generated as Exponential with scale parameter $\exp(-X^{\mathrm{T}}\beta^*)$, where X consists of `Tr` and `X1`–`X4` and $\beta^* = (-.4, .6, -.4, .3, .1)^{\mathrm{T}}$. The censoring variable $\tilde{C}$ is generated as Uniform between 0 and $4\exp(-X^{\mathrm{T}}\beta^*)$. To study discrete-time inference, we assume that two sets of observed data (Y, δ) are obtained from continuous-time variables $(\tilde{T}, \tilde{C})$ as follows: $\delta = 1\{\tilde{T} \le \tilde{C}\}$ and Y is defined by discretizing $\tilde{Y} = \min(\tilde{T}, \tilde{C})$ in intervals of length .01 or .2, using the censor-late option as in Tan (2022, Supplement). The probability of $\delta = 0$ (censored event) is about 24.7%. Both probability model (1) and odds model (2) are fit with the regression terms `Tr` and `X1`–`X4`. These models are only mildly misspecified, and even less misspecified for the finely discretized data than for the coarsely discretized data. Nevertheless, inference can be performed by treating these models as approximations.

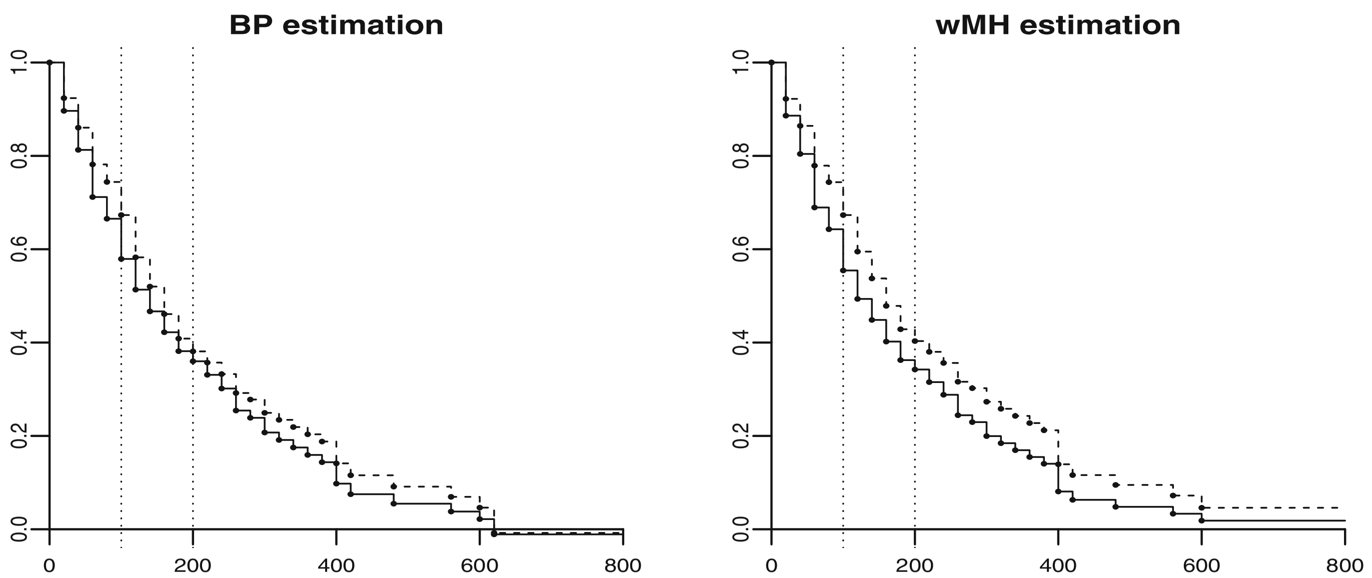

Fig. 2 Survival probabilities for individuals in the test (solid) or standard (dashed) treatment group, with covariates `age`= 60, `Karn`= 60, `diagt`= 9, `cell`= 1, and `prior`= 0. Two vertical lines are placed at days 100 and 200

Table 3 Comparison from simulated data (finely discretized, nearly prop hazards)

	BP	Efron	CML	wMH	Plogit
	Target value				
Tr	−.398	−.400	–	−.402	−.402
X1	.597	.600	–	.603	.603
X2	−.399	−.401	–	−.403	−.403
X3	.298	.299	–	.301	.301
X4	.100	.100	–	.101	.101

	BP	Efron	CML	wMH	Plogit	BP	Efron	CML	wMH	Plogit
	Point mean					Point SD				
Tr	−.404	−.406	−.408	−.408	−.415	.175	.176	.177	.177	.181
X1	.606	.609	.612	.612	.623	.109	.110	.110	.110	.113
X2	−.405	−.408	−.410	−.410	−.417	.115	.116	.116	.116	.119
X3	.303	.305	.306	.307	.312	.114	.114	.115	.115	.118
X4	.104	.105	.105	.105	.107	.100	.101	.101	.101	.103

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	wMH	Plogit
	Model-based SE						Model-robust SE			
Tr	.171	.170	.171	.172	.172	.174	.168	.169	.170	.174
X1	.109	.108	.109	.110	.110	.111	.106	.107	.108	.110
X2	.114	.113	.114	.115	.115	.116	.112	.112	.113	.116
X3	.112	.112	.112	.113	.113	.114	.110	.110	.111	.113
X4	.099	.098	.099	.099	.099	.100	.096	.097	.098	.100

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	wMH	Plogit
	Cov90 (model-based SE)						Cov90 (model-robust SE)			
Tr	.887	.886	.885	–	.885	.880	.882	.882	.882	.882
X1	.905	.901	.902	–	.902	.896	.894	.895	.894	.890
X2	.897	.896	.896	–	.896	.893	.888	.888	.891	.888
X3	.903	.901	.902	–	.901	.898	.888	.888	.887	.888
X4	.894	.892	.892	–	.892	.888	.886	.886	.884	.883

Note: See the footnote for Table 1. Point mean and SD are the Monte Carlo mean and standard deviation of the point estimates, and model-based and model-robust SEs are the square roots of the Monte Carlo mean of the model-based and model-robust variance estimates. Cov90 is the coverage proportion of the 90% Wald confidence intervals, depending on whether model-based SE or model-robust SE is used.

Tables 3–4 present the results from 2000 repeated simulations. The target values, $\bar{\gamma}$ or $\bar{\beta}$, are calculated as the Monte Carlo means of estimates from 200 repeated simulations each of size 10^4, and, due to model misspecification on the discrete data, differ from the values in β^* used to generate the underlying continuous-time data. Calculation of the target values for CML would be computationally prohibitive using the R package `Survival` with this large sample size n and hence is not pursued.

Table 4 Comparison from simulated data (coarsely discretized, nearly prop hazards)

	BP	Efron	CML	wMH	Plogit
	Target value				
Tr	−.347	−.380	–	−.420	−.420
X1	.519	.569	–	.629	.629
X2	−.347	−.380	–	−.420	−.420
X3	.259	.284	–	.314	.314
X4	.087	.095	–	.105	.105

	BP	Efron	CML	wMH	Plogit	BP	Efron	CML	wMH	Plogit
	Point mean					Point SD				
Tr	−.352	−.386	−.428	−.430	−.435	.160	.176	.194	.197	.198
X1	.525	.578	.640	.645	.650	.097	.108	.122	.126	.124
X2	−.352	−.387	−.428	−.432	−.436	.103	.115	.127	.131	.130
X3	.263	.289	.320	.323	.326	.103	.114	.126	.129	.128
X4	.090	.099	.110	.111	.112	.090	.100	.110	.113	.113

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	wMH	Plogit
	Model-based SE						Model-robust SE			
Tr	.171	.155	.171	.188	.191	.190	.153	.168	.187	.190
X1	.107	.095	.108	.121	.126	.122	.095	.105	.122	.122
X2	.113	.102	.114	.126	.129	.127	.101	.111	.126	.127
X3	.111	.101	.112	.124	.127	.125	.099	.109	.123	.124
X4	.098	.089	.098	.108	.110	.109	.088	.096	.108	.109

	oldBP	BP	Efron	CML	wMH	Plogit	BP	Efron	wMH	Plogit
	Cov90 (model-based SE)						Cov90 (model-robust SE)			
Tr	.921	.888	.889	–	.888	.887	.884	.885	.881	.884
X1	.934	.898	.902	–	.905	.902	.896	.894	.889	.898
X2	.924	.896	.896	–	.895	.895	.890	.888	.884	.894
X3	.930	.897	.899	–	.898	.900	.888	.887	.886	.892
X4	.927	.889	.894	–	.895	.890	.888	.888	.887	.888

The point estimates show relatively small biases from the target values, with Plogit the most biased. For coarsely discretized data, the BP estimates are attenuated from β^* toward 0, whereas the CML and wMH are amplified away from 0, by the nature of how the target values of these estimators are associated with probability or odds ratios.

All the model-based and model-robust variance estimates, except oldBP, appear to reasonably match the Monte Carlo variances, with only small under-estimation. Such agreement between model-based and model-robust variance estimation in the current setting (but not the second setting below) can be explained by the fact the the probability and odds models are mildly misspecified here. The commonly reported model-based variance estimates in the column oldBP are upward biased, for example, by $(.107/.097)^2 = 21.7\%$ for the coefficient of $\texttt{X1}$ with coarsely discretized data.

The results from the second study are presented in Tables S1–S2 in the Supplement, where probability model (1) and odds model (2) are more severely misspecified than in the first study. Comparison with Tables 3–4 from the first study shows that the point estimates are more noticeably biased from the target values, and all the variance estimates, except oldBP, exhibit more substantial under-estimation, so that the coverage proportions deviate more from the 90% nominal level. This shows that point and interval estimation accuracy is affected by the degree of model misspecification. Nevertheless, compared with the model-based variance estimates, the model-robust ones tend to match the Monte Carlo variances more closely, and the associated Wald confidence intervals overall achieve coverage propositions closer to the nominal level.

7 Conclusion

For discrete-time survival analysis, we develop new methods and theory using numerically simple and conditionally unbiased estimating functions, along with model-based and model-robust variance estimation, in hazard probability and odds models. The latter is known as Cox's discrete-time proportional hazards model. Due to conditional unbiasedness, our methods are expected to perform satisfactorily in a broad range of settings, with small or large numbers of tied events corresponding to a large or small number of time intervals. In fact, the Breslow–Peto and the weighted Mantel–Haenszel estimators and the associated model-based variances estimators reduce to the partial likelihood estimator and the associated variance estimator in the extreme case of only one event per risk set as would be observed in the continuous-time setting. In this sense, our work provides unified methods for both discrete- and continuous-time survival analysis. Similar ideas can be pursued to address other related problems.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10985-022-09585-1.

References

Allison PD (1982) Discrete-time methods for the analysis of event histories. Sociol Methodol 13:61–98
Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical Models Based on Counting Processes. Springer, New York
Breslow NE (1974) Covariance analysis of censored survival data. Biometrics 30:89–100
Breslow NE (1981) Odds ratio estimators when the data are sparse. Biometrika 68:73–84
Buja A, Berk R, Brown L, George E, Pitkin E, Traskin M, Zhao L, Zhang K (2019) Models as approximations I: Consequences illustrated with linear regression. Stat Sci 34:523–544
Cochran WG (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10:417–451
Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc Ser B 34:187–220
Cox DR, Oaks DO (1984) Analysis of survival data. Chapman & Hall, London
Efron B (1977) The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc 72:557–565
Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley, New York
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Lin DY, Wei LJ (1989) The robust inference for the Cox proportional hazards model. J Am Stat Assoc 84:1074–1079
Lindsay BG (1980) Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. Philos Trans R Soc Ser A 296:639–665
Lindsay BG (1983) Efficiency of the conditional score in a mixture setting. Ann Stat 11:486–197
Manski CF (1988) Analog estimation methods in econometrics. Chapman & Hall, New York
Mantel N, Haenszel WM (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748
Peto R (1972) Contribution to the discussion of Cox (1972): regression models and life tables. J R Stat Soc Ser B 34:205–207
Prentice RL, Gloeckler LA (1978) Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34:57–67
Robins JM, Breslow NE, Greenland S (1986) Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models. Biometrics 42:311–324
Tan Z (2020) Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. Biometrika 107:137–158
Tan Z (2020b) dSurvival: discrete-time survival analysis, R package version 1.0. http://www.stat.rutgers.edu/~ztan
Tan Z (2022) Analysis of odds, probability, and hazard ratios: from 2 by 2 ables to two-sample survival data. J Stat Plan Inference 221:248–265
Therneau TM (2015) A Package for Survival Analysis, version 2.38
Therneau TM, Grambsch PM, Fleming TR (1990) Martingale based residuals for survival models. Biometrika 77:147–160
Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Springer, New York
Thompson WA Jr (1977) On the treatment of grouped observations in life studies. Biometrics 33:463–470
Tsiatis AA (1981) A large sample study of Cox's regression model. Ann Stat 9:93–108
White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50:1–25
Willett JB, Singer JD (2004) Discrete-time survival analysis. In: Kaplan D (ed) SAGE handbook of quantitative methodology for the social sciences, pp 200–213