Analysis of odds, probability, and hazard ratios: From 2 by 2 tables to two-sample survival data

Zhiqiang Tan
Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, 08854, United States of America

ARTICLE INFO

Article history:
Received 15 March 2021
Received in revised form 9 February 2022
Accepted 8 May 2022
Available online 21 May 2022

Keywords:
Breslow–Peto estimator
Conditional likelihood
Mantel–Haenszel estimator
Model-robust variance estimation
Partial likelihood
Proportional hazards model

ABSTRACT

Analysis of 2 by 2 tables and two-sample survival data has been widely used. Exact calculation is computational intractable for conditional likelihood inference in odds ratio models with large marginals in 2 by 2 tables, or partial likelihood inference in Cox’s proportional hazards models with considerable tied event times. Approximate methods are often employed, but their statistical properties have not been formally studied while taking into account the approximation involved. We develop new methods and theory by constructing suitable estimating functions while leveraging knowledge from conditional or partial likelihood inference. We propose a weighted Mantel–Haenszel estimator in an odds ratio model such as Cox’s discrete-time proportional hazards model. Moreover, we consider a probability ratio model, and derive as a consistent estimator the Breslow–Peto estimator, which has been regarded as an approximation to partial likelihood estimation in the odds ratio model. We study both model-based and model-robust variance estimation. For the Breslow–Peto estimator, our new model-based variance estimator is no greater than the commonly reported variance estimator. We present numerical studies which support the theoretical findings.

1. Introduction

Analysis of 2 × 2 tables and two-sample survival data has been widely used. The subjects are covered in numerous articles and books (e.g., Andersen et al., 1993; Breslow and Day, 1980; Cox and Oaks, 1984; Kalbfleisch and Prentice, 1980; McCullagh and Nelder, 1989). The dominant approach is to use odds ratio models and conditional likelihood inference for handling 2 × 2 tables, and Cox’s (1972) proportional hazards models and partial likelihood inference for analyzing censored survival data. The two methods are closely related or, to some extent, equivalent to each other. In fact, conditional logistic regression can be implemented by calling a computer routine for fitting Cox’s regression model, as seen in the popular R package survival (Therneau, 2015).

There are, however, open problems in the existing theory and methods. For Cox’s proportional hazards model with survival data, the partial likelihood is powerful and analytically simple in the absence of tied event times. Large sample theory has been developed using counting processes, where the event time is commonly assumed to be absolutely continuous, thereby excluding the possibility of tied event times. There are two ways in which ties may arise in survival data. One way is that the underlying survival outcome is continuous, but the measurements are made in intervals which are not small enough. Second way is that the survival outcome of interest may be more meaningful if considered being
discrete instead of continuous. An example in the latter case can be found in Allison (1982), where the survival outcome is time to a change of employers among a sample of biochemists from the first year as assistant professors.

A possible approach for handling tied data is to use Cox’s (1972) discrete-time version of proportional hazards models, which amounts to modeling odds ratios of hazard probabilities. The associated partial likelihood is conceptually straightforward, but exact calculation is numerically difficult with a moderate or large number of ties. Alternatively, various ad hoc approximations to the exact partial likelihood have been proposed (Breslow, 1974; Efron, 1977; Peto, 1972). It is often said that such approximations could yield satisfactory results with a small number of ties, but this reasoning defeats the very purpose of using approximations to deal with a relatively large number of ties. There seems to be no formal theory to justify these approximate methods or study their operating characteristics. In fact, if the event time is truly discrete, then the estimators of Breslow (1974) and Efron (1977) would in general be inconsistent under Cox’s discrete-time proportional hazards model. For discrete-time survival analysis, it has also been proposed to use unconditional maximum likelihood estimation, 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). See Allison (1982) for a review. But such methods seem problematic in the presence of a large number of event times or intervals, which lead to the same number of nuisance parameters.

There are similar issues in the existing theory and methods for analyzing $2 \times 2$ tables under odds ratio models (Zelen, 1971; Breslow, 1976). Although various methods were proposed in the early literature including Mantel–Haenszel estimation (Cochran, 1954; Mantel and Haenszel, 1959), conditional maximum likelihood estimation (Breslow, 1981; McCullagh and Nelder, 1989) has been regarded as the “gold standard” for various reasons including optimal asymptotic properties (Lindsay, 1980) and superior empirical performance (Hauck, 1984), as remarked by Breslow and Cologne (1986). In particular, conditional likelihood inference is well-behaved with a fixed number of large tables or a large number of sparse tables. On the other hand, exact calculation for conditional likelihood estimation is numerically intractable for tables with large marginals, similarly to partial likelihood estimation with a large number of ties. Approximate methods have been proposed, with supportive numerical evidence (Breslow and Cologne, 1986; McCullagh and Nelder, 1989). But there seems to be no formal analysis of statistical properties of these methods, while taking into account the approximation involved.

To address the foregoing issues, we develop new methods and theory for analyzing $2 \times 2$ tables and two-sample survival data. See Tan (2020) for an extension to regression models for survival analysis. In contrast with previous methods, our general approach is to carefully construct estimating functions, while leveraging knowledge from conditional or partial likelihood inference in these problems.

- We propose a weighted Mantel–Hasenszel estimator in an odds ratio model, such that it is numerically tractable and expected to achieve similar performance as the conditional or partial likelihood estimator.
- We derive as a consistent estimator the Breslow–Peto estimator in a probability ratio model, even though the same estimator is known as an approximation to the partial likelihood estimator in the odds ratio model.
- We study both model-based and model-robust variance estimation, where model-robust variance estimation captures sampling variation of a point estimator if a posited model may be misspecified as an approximation to the truth.

In particular, for two-sample survival analysis, our work not only clarifies differences between probability and odds ratio models (although they both reduce to Cox’s hazard ratio model in the continuous-time limit), but also shows that the Breslow–Peto point estimator along with our new model-based and model-robust variance estimators are valid in the probability ratio model in both discrete and continuous time.

There are two distinctive features in our approach methodologically. First is use of estimating functions for point estimation while exploiting likelihood-based ideas, to achieve desirable numerical and statistical properties. Our point estimators are M-estimators based on concave objective functions and can be readily solved by standard optimizers. Second is incorporation of model-robust variance estimation, as an important complement to model-based inference in quantifying sampling variation with possible model misspecification. See Freedman (2006) and Buja et al. (2019) for related discussions on model-robust variance estimation. Both our model-based and model-robust variance estimators are of closed form given the point estimators.

In the following, we highlight main ingredients in the construction and properties of our point and variance estimators. First, in an odds ratio model for $2 \times 2$ tables, we derive simple estimating functions for the weighted Mantel–Haenszel estimator from two related angles. One is to achieve a close approximation to optimal estimating functions in minimizing asymptotic variances when positive responses are rare, which are often satisfied in applications. The other is to mimic conditional likelihood estimation in the extreme case where the total number of successes is 1 in each $2 \times 2$ table. The weighted Mantel–Haenszel estimator can be shown to be consistent and asymptotically normal in two asymptotic settings with large tables or many sparse tables, provided that the odds ratio model is valid. Moreover, to complement model-based inference, we derive a model-robust variance estimator, which are consistent in both asymptotic settings, while allowing for possible misspecification of the odds ratio model.

Second, in a probability ratio model for $2 \times 2$ tables, we construct simple estimating functions, which not only provides a reasonable approximation to optimal estimating functions in minimizing asymptotic variances, but also coincides with the Breslow–Peto approximation of the conditional likelihood in the odds ratio model. The resulting estimator, called the
Breslow–Peto estimator, can be shown to be consistent and asymptotically normal in two asymptotic settings with large tables or many sparse tables, provided that the probability ratio model is valid. Moreover, we derive a model-robust variance estimator and a model-based variance estimator. The new model-based variance estimator is shown to be no greater than the commonly reported model-based variance estimator for the Breslow–Peto estimator, although the two variance estimators are identical in the special case of a total of one success in each table. Finally, for two-sample survival analysis, we directly adopt the weighted Mantel–Haenszel estimator in an odds ratio model, or the Breslow–Peto estimator in a probability ratio model, with $2 \times 2$ tables constructed as usual from risk sets over time. The model-based variance estimators from $2 \times 2$ tables can be seen to remain valid due to a martingale argument. We then derive model-robust variance estimators for the weighted Mantel–Haenszel estimator and the Breslow–Peto estimator respectively. The latter variance estimator can be shown to coincide with an extension of Lin and Wei’s (1989) variance estimator, which is originally proposed and studied for the partial likelihood estimator in Cox’s proportional hazards model in continuous time. As a result, the Breslow–Peto estimator and our new model-based and model-robust variance estimators are valid in the probability ratio model in both continuous and discrete time.

For convenience, Table 1 lists the models and the point and variance estimators which are discussed in the remaining sections.

### 2. Analysis of $2 \times 2$ tables

Suppose that a series of $2 \times 2$ tables on a response and a factor are obtained independently from $J$ strata, as shown in Table 2. Denote $N_j = N_{1j} + N_{2j}$, and $N_\ast = \sum_{j=1}^{J} N_j$. For concreteness, the value 1 is called a success, and 2 a failure for the response. For each $j = 1, \ldots, J$, the counts $n_{1j}$ and $n_{2j}$ are assumed to be independent binomial, with fixed denominators $N_{1j}$ and $N_{2j}$ and unknown probabilities $p_{1j}$ and $p_{2j}$. Denote $p_{1j} = 1 - p_{1j}$ and $p_{2j} = 1 - p_{2j}$. The raw estimates of probabilities are defined as $\hat{p}_{1j} = n_{1j}/N_{1j}, \hat{p}_{2j} = n_{2j}/N_{2j}$, etc.

For asymptotic evaluation, it is of interest to examine two distinct settings, referred to as Settings I and II. In Setting I (large tables), the number of tables $K$ is fixed while individual cell counts increase to infinity. In Setting II (many sparse tables), the number of tables increases while the cell counts remain bounded.

#### 2.1. Odds ratio inference

Consider a model on the odds ratios, $\psi_j^\ast = p_{1j}p_{2j}/(p_{1j}p_{2j})$, as follows (Zelen, 1971; Breslow, 1976):

$$\log(\psi_j^\ast) = x_j^T \beta^\ast, \quad j = 1, \ldots, J,$$

where $x_j$ is a covariate vector associated with the $j$th table and $\beta^\ast$ is an unknown coefficient vector. Inference in such models has been extensively studied, particularly in the case of common odds ratios, $\psi_1^\ast = \cdots = \psi_J^\ast = \exp(\beta^\ast)$, corresponding to $x_1 = \cdots = x_J = 1$ (e.g., Cochran, 1954; Mantel and Haenszel, 1959). Our purpose for studying model (1) is two-fold. First, $2 \times 2$ tables stratified by covariates remains a simple but powerful design for case-control studies, where how the response is associated with the stratification covariates is allowed to be unspecified. For an unstratified case-control study, logistic regression can be employed, but depending on how regression terms are specified for the covariates. Second, analysis of $2 \times 2$ tables provides a basic setting to develop our methods and theory, before we tackle two-sample survival analysis with additional complication due to time dependency in Section 3.
For our method, we use the estimating function
\[ \tau_p(\beta) = \sum_{j=1}^{J} \rho_j(\beta)(\hat{p}_{1j}\hat{p}_{2j} - \psi_j(\beta)\hat{p}_{12j})x_j, \]  
(2)
where \( \psi_j(\beta) = \exp(x_j^T\beta) \) and \( \rho_j(\beta) \) is a scalar, non-random function of \( \beta \) for \( j = 1, \ldots, J \). This estimating function is apparently unbiased: \( E[\tau_p(\beta)] = 0 \) at \( \beta = \beta^* \). The associated estimator, denoted by \( \hat{\beta}_p \), is defined as a solution to \( \tau_p(\beta) = 0 \). A choice of \( \rho_j(\beta) \) independently of \( \beta \) is \( \rho_j^{(0)}(\beta) = N_{1j}N_{2j}/N_j \), which in the case of common odds ratios yields the original Mantel–Haenszel estimator
\[ \hat{\beta}^{(0)} = \log \left( \frac{\sum_{j=1}^{J} n_{1j}n_{2j}/N_j}{\sum_{j=1}^{J} n_{12j}n_{21j}/N_j} \right). \]

In general, a Mantel–Haenszel estimator, \( \hat{\beta}^{(0)} \), can be defined as a solution to (2) for \( \rho_j = \rho_j^{(0)} \), with possibly stratum-dependent covariates.

We develop our method in several steps. First, we find the optimal choice of \( \rho_j(\beta) \) in minimizing the asymptotic variance of \( \hat{\beta}_p \), in Setting I, provided that model (1) is valid. Then we derive a simple choice of \( \rho_j(\beta) \), defined as
\[ \rho_j^{(1)}(\beta) = \frac{N_{1j}N_{2j}}{N_{1j}\psi_j(\beta) + N_{2j}}, \]
such that it not only provides a reasonable approximation to the optimal choice but also leads to a desirable reduction of (2) to conditional likelihood estimation (Breslow, 1981; McCullagh and Nelder, 1989) when the total number of successes happens to be \( 1, n_{1j} + n_{2j} = 1 \). In general, conditional likelihood estimation is well-behaved in Setting II, and closely related to partial likelihood estimation (Cox, 1972) in survival analysis, which is discussed in Section 3. The resulting estimator, \( \hat{\beta}^{(1)} \), defined as a solution to \( \tau^{(1)}(\beta) = 0 \) is called the weighted Mantel–Haenszel estimator, where
\[ \tau^{(1)}(\beta) = \sum_{j=1}^{J} \frac{n_{1j}n_{2j} - \psi_j(\beta)n_{12j}n_{21j}}{N_{1j}\psi_j(\beta) + N_{2j}}x_j. \]

Incidentally, \( \hat{\beta}^{(1)} \) can be directly shown to be a maximizer of the concave function
\[ \sum_{j=1}^{J} \left[ N_{1j}\hat{A}_j - (N_{1j}\hat{A}_j + N_{2j}\hat{B}_j)\log(N_{1j}\psi_j(\beta) + N_{2j}) \right], \]
(5)
where \( \hat{A}_j = \hat{p}_{11j}\hat{p}_{22j} \) and \( \hat{B}_j = \hat{p}_{12j}\hat{p}_{21j} \). Finally, to complement model-based inference, we propose a model-robust estimator of the variance of \( \hat{\beta}^{(w)} \), which is consistent in both Settings I and II while allowing for possible misspecification of model (1).

**Remark 1.** The estimator \( \hat{\beta}^{(0)} \) is invariant to the exchange between the factor levels and between the response values with constant covariates. But such invariance in general fails with nonconstant covariates. By comparison, the weighted Mantel–Haenszel estimator \( \hat{\beta}^{(w)} \) is invariant to the exchange between the factor levels with possibly nonconstant covariates, but generally not to the exchange between the response values (except, for example, \( N_{1j} = N_{2j} \) for all \( j \)). In fact, it is preferable to apply the estimator \( \hat{\beta}^{(w)} \) with relatively small success probabilities \( (p_{11j}, p_{22j}) \), as discussed later in Section 2.1.1. For the empirical example in Section 4.1, a positive response corresponds to radiation exposure, which has a small success probability.

2.1.1. **Point estimation**

In this section, we discuss the derivation of the choice \( \rho_j^{(w)}(\beta) \) for the weighted Mantel–Haenszel estimator \( \hat{\beta}^{(w)} \). First, it can be shown that if model (1) is valid, then the asymptotic variance of \( \hat{\beta}_p \) is of the sandwich form \( N_{\bullet}^{-1}H_{\beta}^{-1}(\beta^*)G_{\beta}(\beta^*)H_{\beta}^{-1}(\beta^*) \) under standard regularity conditions in both Settings I and II (Davis, 1985), where
\[
G_{\beta}(\beta) = N_{\bullet}^{-1} \sum_{j=1}^{J} \rho_j^{(0)}(\beta) \text{var}((\hat{p}_{1j}\hat{p}_{2j} - \psi_j(\beta)\hat{p}_{12j})x_jx_j^T),
\]
\[
H_{\beta}(\beta) = N_{\bullet}^{-1} \sum_{j=1}^{J} \rho_j(\beta)\psi_j(\beta)p_{12j}p_{21j}x_jx_j^T.
\]
Then the optimal choice of \( \rho_j(\beta) \) can be obtained similarly as in theory of quasi-likelihood functions (McCullagh and Nelder, 1989). See the Supplement for a direct proof.
Proposition 1. Suppose that odds ratio model (1) is valid.

(i) The asymptotic variance of \( \hat{\beta}_j \) in both Settings I and II is minimized by the choice

\[
\rho_j^1(\beta) = \frac{\psi_j(\beta)p_{12j}p_{21j}}{\text{var}[\hat{p}_{1j}p_{2j2} - \psi_j(\beta)p_{12j}p_{21j}]} , \quad j = 1, \ldots, J.
\]

(ii) In Setting I as \( N_1 \to \infty \) and \( N_{1j}/N_j \) tending to a constant in \((0, 1)\) for each \( j \) with \( J \) fixed, the asymptotic variance of \( \hat{\beta}_j \) is also minimized by the choice

\[
\rho_j^1(\beta) = \frac{N_{1j}N_{2j}}{N_{1j}(p_{11j} + \psi_j(\beta)p_{12j}) + N_{2j}(\psi_j(\beta)p_{21j} + p_{22j})} , \quad j = 1, \ldots, J.
\]

The foregoing choice \( \rho_j^1(\beta) \) cannot be directly used, due to its dependency on the unknown quantities \((p_{11j}, p_{21j})\). In Setting I, this difficulty can in principle be overcome by replacing \((p_{11j}, p_{21j})\) with their consistent estimators \((\hat{p}_{11j}, \hat{p}_{21j})\). The resulting estimator of \( \beta \) can be shown to achieve the same asymptotic variance as the (infeasible) estimator \( \hat{\beta}_{\rho^1} \) with the optimal choice \( \rho_j^1(\beta) \). In Setting II, however, such data-dependent approximation of \( \rho_j^1(\beta) \) can lead to poor performance, because the variation of \((\hat{p}_{11j}, \hat{p}_{21j})\) is no longer negligible with bounded sizes \((N_{1j}, N_{2j})\).

To achieve good performance in both Settings I and II, we propose the simple choice \( \rho_j^{(w)}(\beta) \), defined in (3), as a data-independent (i.e., non-random) approximation to \( \rho_j^1(\beta) \). The relative error in this approximation for \( \beta = \beta^* \),

\[
\frac{\rho_j^{(w)}(\beta^*)}{\rho_j^1(\beta^*)} - 1 = \{1 - \psi_j(\beta^*)\} \frac{N_{1j}p_{11j} - N_{2j}p_{21j}}{N_{1j}\psi_j(\beta^*) + N_{2j}},
\]

is close to 0 whenever the odds ratio \( \psi_j(\beta^*) \) is close to 1 or the success probabilities \((p_{11j}, p_{21j})\) are close to 0. The resulting estimator \( \hat{\beta}^{(w)} \) is expected to perform similarly to the (infeasible) optimal estimator \( \hat{\beta}_{\rho^1} \) in Setting I, especially when differences between the two groups are small or positive responses are rare.

The appropriateness of the proposed choice \( \rho_j^{(w)}(\beta) \) in Setting II (as well as Setting I) can also be seen from the following connection to conditional likelihood estimation, which is known for its superior performance in both Settings I and II (Breslow, 1981; McCullagh and Nelder, 1989). In fact, the condition score function is

\[
s(\beta) = \sum_{j=1}^J \{n_{11j} - \mu_{11j}(\beta)\} x_j,
\]

where \( \mu_{11j}(\beta) \) is the conditional expectation of \( n_{11j} \) given the marginals \((n_{11j} + n_{21j}, N_{1j}, N_{2j})\) in the jth table with odds ratio \( \psi_j(\beta) \). In the case of \( n_{11j} + n_{21j} = 1 \) (i.e., a total of one success in the two groups), it can be directly shown that

\[
n_{11j} - \mu_{11j}(\beta) = n_{11j} - \left(\frac{(n_{11j} + n_{21j})N_{1j}\psi_j(\beta)}{N_{1j}\psi_j(\beta) + N_{2j}}\right)
\]

\[
= \frac{n_{11j}n_{22j} - \psi_j(\beta)n_{12j}n_{21j} + (1 - \psi_j(\beta))n_{11j}n_{12j}}{N_{1j}\psi_j(\beta) + N_{2j}},
\]

where the last step holds because \((n_{11j}, n_{21j}) = (0, 1) \) or \((1, 0)\) and hence \(n_{11j}n_{22j} = 0\). Therefore, the jth contribution to the proposed estimating function \( r^{(w)}(\beta) \) in (4) coincides with that to the conditional score function \( s(\beta) \) when the total number of successes is 1 in the jth table. The coincidence in such an extreme case suggests that the proposed estimator \( \hat{\beta}^{(w)} \) tends to perform similarly to the maximum conditional likelihood estimator \( \hat{\beta}^{(c)} \), defined as a solution to \( s(\beta) = 0 \).

There is another implication from the discussion above on the approximation of \( \hat{\beta}^{(w)} \) to \( \hat{\beta}_{\rho^1} \) with small \((p_{11j}, p_{21j})\) in Setting II and the reduction to \( \hat{\beta}^{(c)} \) in the case of a total of one success. It is more desirable to apply \( \hat{\beta}^{(w)} \) with relatively small success probabilities for responses. Otherwise, the labeling of response values should be exchanged.

In the Appendix, we provide additional remarks to compare the proposed estimator \( \hat{\beta}^{(w)} \) with the conditional likelihood estimator \( \hat{\beta}^{(c)} \) and the estimator of Davis (1985).

2.1.2. Variance estimation

In this section, we propose a consistent estimator of the asymptotic variance of the weighted Mantel–Haenszel estimator \( \hat{\beta}^{(w)} \) in both Settings I and II, while allowing for possible misspecification of model (1). Such a variance estimator for a point estimator is referred to as model-robust with respect to the associated model. In contrast, model-based variance estimators are constructed such that they are inconsistent unless the associated model is correctly specified. There are various model-based variance estimators for the Mantel–Haenszel estimator \( \hat{\beta}^{(o)} \) under the assumption of common odds ratios. See Kuritz et al. (1988) for a review and the Supplement for further discussion about relationships between existing variance estimators. To describe the asymptotic behavior of \( \hat{\beta}^{(w)} \), we adopt the standard theory of estimation in misspecified models (e.g., White, 1982; Manski, 1988). First, it can be shown in both Settings I and II, under regularity conditions similar to those in Davis (1985), that \( \hat{\beta}^{(w)} \) converges in probability to a target value \( \hat{\beta}^{(w)} \), defined as a solution
to
\[ 0 = \sum_{j=1}^{J} \rho_j^{(w)}(\beta) \{ p_{1j} p_{2j} - \psi_j(\beta) p_{1j} p_{2j} \} x_j, \]  
(8)

or equivalently as a unique maximizer of the concave function
\[ \sum_{j=1}^{J} [N_j A_j x_j^T \beta - (N_j A_j + N_2 B_j) \log(N_j \psi_j(\beta) + N_2)], \]  
(9)

where \(A_j = p_{1j} p_{2j}\) and \(B_j = p_{1j} p_{2j}\). Eq. (8) and function (9) are the population versions of (4) and (5) respectively. If model (1) is valid, then \(\hat{\beta}^{(w)} = \beta^*\) such that \(\psi_j(\beta^*) = \psi_j^*\) for \(j = 1, \ldots, J\). Otherwise, \(\psi_j(\beta^{(w)})\) may differ from \(\psi_j^*\). Moreover, it can be shown that \(N_*^{-1/2}(\hat{\beta}^{(w)} - \bar{\beta}^{(w)})\) converges in distribution to \(N(0, \Sigma^{(w)})\), where \(\Sigma^{(w)} = H^{(w)}(\beta)G^{(w)}(\beta)H^{(w)}(\beta)^{-1}(\beta)|_{\beta = \bar{\beta}^{(w)}}\) with
\[ G^{(w)}(\beta) = N_*^{-1} \sum_{j=1}^{J} \rho_j^{(w)}(\beta) \text{var} \{ p_{1j} \hat{p}_{2j} - \psi_j(\beta) \hat{p}_{1j} \hat{p}_{2j} \} x_j x_j^T, \]
\[ H^{(w)}(\beta) = N_*^{-1} \sum_{j=1}^{J} (N_j A_j + N_2 B_j) \frac{N_j N_2 \psi_j(\beta)}{[N_j \psi_j(\beta) + N_2]^2} x_j x_j^T. \]
The matrix \(H^{(w)}(\beta)\) is obtained from the negative Hessian of function (9). The asymptotic variance of \(\hat{\beta}^{(w)}\) is \(\hat{\Sigma}^{(w)} = N_*^{-1} \Sigma^{(w)} = N_*^{-1} H^{(w)}(\beta) G^{(w)}(\beta) H^{(w)}(\beta)^{-1}(\beta)|_{\beta = \bar{\beta}^{(w)}}\), which is invariant if scaling by \(N_*^{-1}\) is dropped from the right hand side and from \(G^{(w)}\) and \(H^{(w)}\).

For the variance matrix \(\Sigma^{(w)}\), our proposed estimator is
\[ \hat{\Sigma}^{(w)} = \hat{H}^{(w)}(\beta) \hat{G}^{(w)}(\beta) \hat{H}^{(w)}(\beta)^{-1}(\beta)|_{\beta = \bar{\beta}^{(w)}}, \]
where \(\hat{H}^{(w)}(\beta)\) is defined as \(H^{(w)}(\beta)\) with \((A_j, B_j)\) replaced by \((\hat{A}_j, \hat{B}_j)\), but \(\hat{G}^{(w)}(\beta) = \sum_{j=1}^{J} \rho_j^{(w)}(\beta) \hat{\delta}_j(\beta) x_j x_j^T\) with
\[ \hat{\delta}_j(\beta) = \frac{\hat{p}_{1j} \hat{p}_{12j}}{N_j - 1} \left( \hat{p}_{22j} + \psi_j(\beta) \hat{p}_{21j} \right)^2 + \frac{\hat{p}_{2j} \hat{p}_{22j}}{N_j - 1} \left( \hat{p}_{1j} + \psi_j(\beta) \hat{p}_{12j} \right)^2 \]
\[- \left( \psi_j(\beta) - 1 \right)^2 \frac{\hat{p}_{1j} \hat{p}_{12j}}{N_j - 1} \left( \hat{p}_{22j} + \psi_j(\beta) \hat{p}_{21j} \right) + \frac{\hat{p}_{2j} \hat{p}_{22j}}{N_j - 1} \left( \hat{p}_{1j} + \psi_j(\beta) \hat{p}_{12j} \right)^2. \]
Then the following properties can be established.

**Proposition 2.** Let \(\sigma_j(\beta) = \text{var} \{ \hat{p}_{1j} \hat{p}_{2j} - \psi_j(\beta) \hat{p}_{1j} \hat{p}_{2j} \}. \) Assume that \(N_{1j} \geq 2\) and \(N_{2j} \geq 2\) for each \(j = 1, \ldots, J\).
(i) \(\hat{\sigma}_j(\beta) \geq 0\) for any fixed \(\beta\) and \(j = 1, \ldots, J\).
(ii) \(\hat{\sigma}_j(\beta)\) is an unbiased estimator of \(\sigma_j(\beta)\) for any fixed \(\beta\) and \(j = 1, \ldots, J\).
(iii) \(\hat{\Sigma}^{(w)}\) is a consistent estimator of \(\Sigma^{(w)}\) in both Settings I and II, with possible misspecification of model (1).

The estimator \(\hat{\delta}_j(\beta)\) serves as a finite-sample correction to the simpler version
\[ \hat{\delta}_j(\beta) = \frac{\hat{p}_{1j} \hat{p}_{12j}}{N_j} \left( \hat{p}_{22j} + \psi_j(\beta) \hat{p}_{21j} \right)^2 + \frac{\hat{p}_{2j} \hat{p}_{22j}}{N_j} \left( \hat{p}_{1j} + \psi_j(\beta) \hat{p}_{12j} \right)^2. \]
(10)

The estimator \(\hat{\Sigma}^{(w)}\) with \(\hat{\delta}_j(\beta)\) replaced by \(\hat{\delta}_j(\beta)\) remains a consistent estimator of \(\Sigma^{(w)}\) in Setting I, but in general becomes inconsistent in Setting II with bounded \((N_{1j}, N_{2j})\). Similar formulas to \(\hat{\sigma}_j(\beta)\) above can be found in Guilbaud (1983).

It is instructive to compare the variance estimator \(\hat{\Sigma}^{(w)}\) with model-based variance estimators. For the Mantel-Haenszel estimator \(\hat{\beta}^{(0)}\), Robins et al. (1986) proposed a variance estimator which is consistent in both Settings I and II under the assumption of common odds ratios. Their estimator of \(\sigma_j(\beta)\) is defined as
\[ \hat{\sigma}_j^{(b)}(\beta) = \frac{\psi_j(\beta) \hat{p}_{2j} \hat{p}_{2j}}{N_j} \left( \hat{p}_{22j} + \psi_j(\beta) \hat{p}_{21j} \right) + \frac{\hat{p}_{2j} \hat{p}_{22j}}{N_j} \left( \hat{p}_{1j} + \psi_j(\beta) \hat{p}_{12j} \right). \]

With possibly nonconstant covariates, it can be shown that if model (1) is valid, then a consistent estimator of \(\Sigma^{(w)}\) in both Settings I and II is
\[ \hat{\Sigma}^{(wb)} = \hat{H}^{(w)}(\beta) \hat{G}^{(wb)}(\beta) \hat{H}^{(w)}(\beta)^{-1}(\beta)|_{\beta = \bar{\beta}^{(w)}}, \]
where \(\hat{H}^{(w)}(\beta)\) is as before, but \(\hat{G}^{(wb)}(\beta) = N_*^{-1} \sum_{j=1}^{J} \rho_j^{(w)}(\beta) \hat{\delta}_j^{(b)}(\beta) x_j x_j^T\). Hence \(\hat{\Sigma}^{(wb)}\) would be identical to \(\hat{\Sigma}^{(wb)}\), except that \(\hat{\delta}_j(\beta)\) is in place of \(\hat{\delta}_j^{(b)}(\beta)\). In fact, \(\hat{\delta}_j(\beta)\) is unbiased for \(\sigma_j(\beta)\) with any fixed \(\beta\) as shown in Proposition 2, whereas
Moreover, \( \hat{\sigma}^{(b)}(\beta) \) is unbiased for \( \sigma(\beta) \) only with \( \beta = \beta^* \) such that \( \psi_j(\beta^*) = \psi_j^* \). Algebraically, \( \hat{\sigma}(\beta) \) is a bivariate polynomial of \((\hat{p}_{11j}, \hat{p}_{21j}) \) including a term \( \hat{p}_{11j}^2 \hat{p}_{21j}^2 \) of total degrees 4, whereas \( \hat{\sigma}^{(b)}(\beta) \) involves terms only up to 3 total degrees such as \( \hat{p}_{11j}^3 \hat{p}_{21j}^3 \).

**Remark 2.** Both the variance estimators \( \hat{\Sigma}^{(w)} \) and \( \hat{\Sigma}^{(wb)} \) are invariant to the exchange between the factor levels, but generally not to that between the response values. This is similar to the invariance properties of the point estimator \( \hat{\beta}^{(w)} \), discussed in Remark 1.

### 2.2. Probability ratio inference

Consider a model on the probability ratios, \( \phi_j^* = p_{11j}/p_{21j} \), as follows:

\[
\log(\phi_j^*) = x_j^T \gamma^*, \quad j = 1, \ldots, J.
\]

where \( x_j \) is a covariate vector associated with the \( j \)-th table and \( \gamma^* \) is an unknown coefficient vector. Compared with odds ratio model (1), such probability ratio models have been directly studied to a lesser extent for various reasons. First, odds ratios models are popular, especially in retrospective studies, due to the invariance of odds ratios to prospective or retrospective sampling. Second, the availability of conditional likelihood inference given table marginals can be appealing in model (1), regarding elimination of nuisance parameters. Third, odds ratios are often considered an approximation to probability ratios when success probabilities are small, corresponding to rare diseases in biomedical applications. Nevertheless, odds ratios are persistently biased estimates of probability ratios in being further away from 1, unless the probabilities are identical between the two factor levels. Moreover, as discussed below, model-robust inference in model (11) can be carried with carefully constructed estimating functions, in a parallel manner to that in model (1). There is also a remarkable connection to Breslow–Peto modification of partial likelihood estimation with tied event times.

#### 2.2.1. Point estimation

We use the estimating function

\[
\zeta_j(\gamma) = \sum_{j=1}^J q_j(\gamma)(\hat{p}_{11j} - \phi_j(\gamma)\hat{p}_{21j})x_j,
\]

where \( \phi_j(\gamma) = \exp(x_j^T \gamma) \) and \( q_j(\gamma) \) is a scalar, non-random function of \( \gamma \) for \( j = 1, \ldots, J \). This estimating function is unbiased: \( E[\zeta_j(\gamma)] = 0 \) at \( \gamma = \gamma^* \). The associated estimator, denoted by \( \hat{\gamma}_j \), is defined as a solution to \( \zeta_j(\gamma) = 0 \). A choice of \( \phi_j(\gamma) \) similar to \( \rho_j^{(0)} \) is \( \phi_j^{(0)}(\gamma) = N_1N_2j/N_j \), independently of \( \gamma \), which in the case of common probability ratios yields \( \hat{\gamma}_j^{(0)} = \log((\sum_{j=1}^J n_{11j}N_2j/N_j)/(\sum_{j=1}^J n_{21j}N_1j/N_j)) \).

Our proposed choice of \( q_j(\gamma) \) is similar to \( \rho_j^{(w)} \):

\[
q_j^{(w)}(\gamma) = \frac{N_1N_2j}{N_1j\phi_j(\gamma) + N_2j}.
\]

The resulting estimator, \( \gamma^{(w)} \), is defined as a solution to \( \zeta^{(w)}(\gamma) = 0 \), where

\[
\zeta^{(w)}(\gamma) = \sum_{j=1}^J \frac{n_{11j}N_2j - \phi_j(\gamma)n_{21j}N_1j}{N_1j\phi_j(\gamma) + N_2j}x_j,
\]

which can be equivalently rewritten as

\[
\zeta^{(w)}(\gamma) = \sum_{j=1}^J \left\{ n_{11j} - \frac{(n_{11j} + n_{21j})N_1j\phi_j(\gamma)}{N_1j\phi_j(\gamma) + N_2j} \right\} x_j,
\]

Moreover, \( \hat{\gamma}_j^{(w)} \) can be directly shown to be a maximizer of the concave function

\[
\sum_{j=1}^J \left[ n_{11j}x_j^T \gamma - (n_{11j} + n_{21j})\log(N_1j\phi_j(\gamma) + N_2j) \right],
\]

which is the log-likelihood of a pseudo-model that the \( n_{11j} + n_{21j} \) successes are independent and identically distributed Bernoulli, each with probability \( N_1j\phi_j(\gamma)/(N_1j\phi_j(\gamma) + N_2j) \) from factor level 1 and the remaining probability from factor level 2.

The derivation of our choice \( q_j^{(w)}(\gamma) \) can be seen from two angles, similarly as in Section 2.1.1. One is based on an approximation to the optimal choice of \( q_j(w) \) in Setting I with model (11) correctly specified. In fact, it can be shown...
that if model (11) is valid, then the asymptotic variance of \( \hat{\gamma}_q \) is of the sandwich form \( N^{-1}_q D_q^{-1}(\gamma^*)C_q(\gamma^*)D_q^{-1}(\gamma^*) \) under standard regularity conditions in both Settings I and II, where

\[
C_q(\gamma) = N^{-1}_q \sum_{j=1}^{J} q_j^2(\gamma) \text{var}(\hat{p}_{11j} - \phi_j(\gamma)\hat{p}_{21j})x_j^T.
\]

\[
D_q(\gamma) = N^{-1}_q \sum_{j=1}^{J} q_j(\gamma)\phi_j(\gamma)p_{21j}x_j^T.
\]

The optimal choice of \( q_j(\gamma) \) can be obtained as follows, similarly as in Proposition 3. There is however a subtle difference: \( q_j^*(\gamma) \) is optimal in both Settings I and II with \( q_j^*(\gamma^*) = q_j^*(\gamma^*) \) exactly, whereas \( \rho_j^*(\beta) \) is optimal in Setting I but not Setting II.

**Proposition 3.** Suppose that odds ratio model (1) is valid. The asymptotic variance of \( \hat{\gamma}_q \) in both Settings I and II is minimized by the choice

\[
q_j^*(\gamma) = \frac{\phi_j(\gamma)p_{21j}}{\text{var}(\hat{p}_{11j} - \phi_j(\gamma)\hat{p}_{21j})}, \quad j = 1, \ldots, J,
\]

or equivalently by the choice

\[
q_j^*(\gamma) = \frac{N_{1j}N_{2j}}{N_{1j}\phi_j(\gamma)p_{21j} + N_{2j}\hat{p}_{11j}}, \quad j = 1, \ldots, J.
\]

By Proposition 3, the relative error of the data-independent choice \( q_j(\gamma) \) as an approximation to the optimal choice \( q_j^*(\gamma) \) for \( \gamma = \gamma^* \) is

\[
\frac{q_j^*(\gamma^*)}{q_j(\gamma^*)} - 1 = -\frac{N_{1j}\phi_j(\gamma^*)p_{21j} + N_{2j}\hat{p}_{11j}}{N_{1j}\phi_j(\gamma^*) + N_{2j}},
\]

which is close to 0 whenever the success probabilities \( (p_{11j}, p_{21j}) \) are close to 0. Hence the proposed estimator \( \hat{\gamma}^{(w)} \) is expected to perform close to being optimal in both Settings I and II, especially when positive responses are rare.

The second motivation for our choice \( q_j(\gamma) \) is that the resulting estimator \( \hat{\gamma}^{(w)} \) coincides with the maximum partial likelihood estimator with Breslow’s (1974) and Peto’s (1972) modification for tied death times in two-sample survival analysis. The \( J \) table can be constructed from the death and survival counts by the factor levels among the risk set at the \( j \)th death time. The estimating function (15) or the criterion function (16) is identical to the score function or the log-likelihood function in the Breslow–Peto modification of partial likelihood estimation, as shown in Proposition 5. Henceforth, the estimator \( \hat{\gamma}^{(w)} \) can be referred to as the Breslow–Peto estimator.

The preceding coincidence needs to be carefully understood. In the case of a total of one success in \( J \)th table, the \( J \)th contribution of \( \tau^{(w)}(\beta) \) and that of \( \zeta^{(w)}(\gamma) \) are both identical to that of the conditional score function \( s(\beta) \). In general, with total numbers of successes greater than 1, the three functions \( s(\beta) \), \( \tau^{(w)}(\beta) \), and \( \zeta^{(w)}(\gamma) \) differ from each other. The first two \( s(\beta) \) and \( \tau^{(w)}(\beta) \) lead to consistent estimation of \( \beta^* \) under odds ratio model (1). In contrast, the estimating function \( \zeta^{(w)}(\gamma) \) leads to consistent estimation of \( \gamma^* \) under probability ratio model (11), even though it was considered to approximate the conditional score \( s(\beta) \) in the context of Cox’s (1972) discrete-time proportional hazards model. See Appendix Remark 7 and Section 3 for further discussion.

### 2.2.2. Variance estimation

We present two estimators of the asymptotic variance of the Breslow–Peto estimator \( \hat{\gamma}^{(w)} \) in both Settings I and II. One is a model-based variance estimator, consistent provided that model (11) is valid. The other is a model-robust variance estimator, consistent in the presence of possible model misspecification.

First, we describe the asymptotic behavior of \( \hat{\gamma}^{(w)} \) while allowing for possible misspecification of model (11). By standard theory of estimation with model misspecification (e.g., White, 1982; Manski, 1988), it can be shown in both Settings I and II that \( \hat{\gamma}^{(w)} \) converges in probability to a target value \( \gamma^{(w)} \), defined as a solution to

\[
0 = \sum_{j=1}^{J} q_j^{(w)}(\gamma)|p_{11j} - \phi_j(\gamma)p_{21j}|x_j,
\]

or equivalently as a unique maximizer of the concave function

\[
\sum_{j=1}^{J} [N_{1j}p_{11j}x_j^T - (N_{1j}p_{11j} + N_{2j}p_{21j})\log(N_{1j}\phi_j(\gamma) + N_{2j})].
\]
Our model-based variance estimator is such that

\[ \text{Proposition 4.} \]

The model-robust estimator is \( \hat{V}(w) = \hat{D}(w)^{-1}(y) \hat{C}(w)(y) \hat{D}(w)^{-1}(y) \big| y = \hat{\gamma}(w) \).

The matrix \( D(w)(y) \) is obtained from the negative Hessian of function (18). The asymptotic variance of \( \hat{\gamma}(w) \) is then \( N^{-1} V(w) = N^{-1} D(w)^{-1}(y) C(w)(y) D(w)^{-1}(y) \big| y = \hat{\gamma}(w) \).

For the variance matrix \( V(w) \), our model-robust estimator is \( \hat{V}(w) = \hat{D}(w)^{-1}(y) \hat{C}(w)(y) \times \hat{D}(w)^{-1}(y) \big| y = \hat{\gamma}(w) \), where \( \hat{D}(w)(y) \) is defined as \( D(w)(y) \) with \((p_{11j}, p_{21j})\) replaced by \((\hat{p}_{11j}, \hat{p}_{21j})\), but \( \hat{C}(w)(y) = N^{-1} \sum_{j=1}^{J} q_{j}^{2}(y) \hat{v}_{j}(y) x_{j} x_{j}^{T} \)

with \( \hat{v}_{j}(y) = \frac{\hat{p}_{11j} \hat{p}_{21j}}{N_{ij} - 1} + \phi_{j}^{2}(y) \frac{\hat{p}_{21j} \hat{p}_{22j}}{N_{2j} - 1}, \ j = 1, \ldots, J. \)

Our model-based variance estimator is \( \hat{V}(w) = \hat{D}(w)^{-1}(y) \hat{C}(w)(y) \hat{D}(w)^{-1}(y) \big| y = \hat{\gamma}(w) \), where \( \hat{C}(w)(y) = N^{-1} \sum_{j=1}^{J} q_{j}^{2}(y) \hat{v}_{j}(y) \).

The following properties can be established.

**Proposition 4.** Let \( v_{j}(y) = \text{var}[\hat{p}_{11j} - \phi_{j}(y) \hat{p}_{21j}] \).

(i) \( \hat{v}_{j}(y^{*}) \) is an unbiased estimator of \( v_{j}(y^{*}) \) for \( j = 1, \ldots, J \), and \( \hat{V}(w) \) is a consistent estimator of \( V(w) \) in both Settings I and II, provided that model (11) is valid.

(ii) Assume that \( N_{ij} \geq 2 \) and \( N_{2j} \geq 2 \) for each \( j = 1, \ldots, J \). Then \( \hat{v}_{j}(y) \) is an unbiased estimator of \( v_{j}(y) \) for any fixed \( y \) and \( j = 1, \ldots, J \). Moreover, \( \hat{V}(w) \) is a consistent estimator of \( V(w) \) in both Settings I and II, with possible misspecification of model (11).

The estimator \( \hat{v}_{j}(y) \) is based on the usual variance estimator for sample proportions with binary data. Algebraically, \( \hat{v}_{j}(y) \) and \( \hat{v}_{j}(y) \) are bivariate polynomials of \((\hat{p}_{11j}, \hat{p}_{21j})\), but \( \hat{v}_{j}(y) \) involves only the cross-product \( \hat{p}_{11j} \hat{p}_{21j} \), not \( \hat{p}_{21j} \hat{p}_{22j} \).

There is, in general, no definite direction in which \( \hat{v}_{j}(y) \) and \( \hat{v}_{j}(y) \) are greater, and hence comparison of magnitudes between \( \hat{V}(w) \) and \( \hat{V}(w) \) is problem-dependent. However, there is a definite comparison between the model-based variance estimator \( N^{-1} \hat{V}(w) \) and the commonly reported variance estimator for the Breslow–Peto estimator \( \hat{\gamma}(w) \).

**Corollary 1.** The model-based variance estimator \( N^{-1} \hat{V}(w) \) is, in the order on variance matrices, always no greater than \( N^{-1} \hat{D}(w)^{-1}(\hat{\gamma}(w)) \), which is the inverse negative Hessian of the criterion function (16), commonly reported as the model-based variance estimator for the Breslow–Peto estimator \( \hat{\gamma}(w) \).

This can be shown by noting that \( \hat{D}(w)(y) \) is identical to \( \hat{C}(w)(y) \) with \( \hat{v}_{j}(y) \) replaced by \( \phi_{j}(y) \hat{p}_{21j}/N_{ij} + \hat{p}_{11j}/N_{2j} \), which is at least as large as \( \hat{v}_{j}(y) \). On the other hand, in the special case where \( n_{1j} = 0 \) or \( n_{2j} = 0 \) for \( j = 1, \ldots, J \) (for example, a total of one success in each table), \( \hat{D}(w)(y) \) is identical to \( \hat{C}(w)(y) \) and hence the variance estimate \( N^{-1} \hat{V}(w) \) numerically agrees with the variance estimate \( N^{-1} \hat{D}(w)^{-1}(\hat{\gamma}(w)) \).

In the Appendix, we provide additional remarks on the variance estimators \( N^{-1} \hat{V}(w) \) for \( \hat{p}(w) \) and \( N^{-1} \hat{V}(w) \) for \( \hat{\gamma}(w) \), on conditional inference in the case of a total of one success per table, and on Wald and score-type tests, related to the log-rank test.

### 3. Two-sample survival analysis

Suppose that survival data and covariates are obtained as \( \{(Y_{i}, \delta_{i}, Z_{i}) : i = 1, \ldots, N\} \) 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 \( Z_{i} \) is a covariate. All individuals are assumed to be event-free at entry, \( T_{i} > 0 \) for all \( i \), and hence it is possible that \( Y_{i} = 0 \) if \( \delta_{i} = 0 \), but not if \( \delta_{i} = 1 \). For two-sample analysis, each covariate \( Z_{i} \) is two-level, being 1 or 2 if individual is from group 1 or 2. Assume that \( \{(T_{i}, C_{i}, Z_{i}) : i = 1, \ldots, N\} \) are independent and identically distributed copies of \( (T, C, Z) \), and hence \( \{(Y_{i}, \delta_{i}, Z_{i}) : i = 1, \ldots, N\} \) are independent and identically distributed copies of \( (Y, \delta, Z) \) 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 $Z$.

In practice, time is usually recorded in discrete units such as days or weeks. Assume that there is a discrete time scale, $0 = t_0 < t_1 < \cdots < t_j < t_{j+1}$, such that $(Y, \delta)$ 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. There are subtle issues when the survival data are collected by grouping continuous or fine-scaled measurements, but such detailed data are not available. See Kaplan and Meier (1958) and Thompson (1977) for early treatment and the Supplement for further discussion.

For survival analysis, it is commonly of interest to estimate hazard functions. In discrete time, the hazard at time $t_j$ given covariate $Z = z$ is defined as the probability $\pi_{zj} = P(T = t_j| T \geq t_j, Z = z)$ for $z = 1, 2$. Under conditionally independent censoring, $\pi_{zj}$ can be identified from observed data as $\pi_{zj} = p_{zj}$, where $p_{zj}$ is the event probability at $t_j$ calculated within the population risk set $\{Y \geq t_j\}$.

$$
p_{zj} = P(Y = t_j, \delta = 1|Y \geq t_j, Z = z). 
$$

The population risk set $\{Y \geq t_j\}$ represents individuals who are event-free (or alive) just prior to time $t_j$. For ease of interpretation, we treat $(p_{11j}, p_{21j})$ interchangeably with the hazard probabilities $(\pi_{11j}, \pi_{21j})$ whenever possible. We also study inference about odds and probability ratios directly in terms of $(p_{11j}, p_{21j})$, which coincide with $(\pi_{11j}, \pi_{21j})$ if conditionally independent censoring holds, but otherwise remain empirically identifiable. Our results are then applicable even if conditionally independent censoring is violated; see Tan (2006), Section 4 for a similar approach.

With the concept of risk sets, two-sample survival analysis can be easily related to analysis of $2 \times 2$ tables in Section 2 (e.g., Mantel, 1966). For $z = 1, 2$ and $j = 1, \ldots, J$, let $N_{2j}$ be the size of $\{1 \leq i \leq N : Y_i = t_j, Z_i = z\}$, the sample risk set associated with time $t_j$ given covariate $Z = z$, and let $N_{1j}$ be the number of events (or deaths) at $t_j$ within the risk set, i.e., the size of $\{1 \leq i \leq N : Y_i = t_j, \delta_i = 1, Z_i = z\}$. Then the survival data $(Y_i, \delta_i) : i = 1, \ldots, N$ can be transformed into a series of $J$ tables as shown in Table 2. By this connection, it is helpful to exploit similar methods from analysis of $2 \times 2$ tables for two-sample survival analysis. Unless otherwise stated, we use the same notation $(p_{11j}, p_{21j})$ and $(n_{11j}, n_{21j})$ etc, although there are important differences. In particular, the $J$ tables are not independent and the sizes $(N_{1j}, N_{2j})$ are random.

Consider two types of models for the probabilities $(p_{11j}, p_{21j})$. The first is model (1), $\log(\psi^*) = x^T \beta^*$, on the odds ratios $\psi^*_j = p_{11j}/p_{21j}$, is that,

$$
\frac{p_{11j}}{1 - p_{11j}} = \frac{\exp(x_j^T \beta^*)}{1 - \exp(x_j^T \beta^*)}, \quad j = 1, \ldots, J.
$$

(20)

where $x_j = x(t_j)$ and each component of $x(\cdot)$ is a function of time, for example, a piecewise-constant function. The second is model (11), $\log(\phi^*) = x^T \gamma^*$, on the probability ratios $\phi^*_j = p_{11j}/p_{21j}$, is that,

$$
p_{11j} = \exp(x_j^T \gamma^*)p_{21j}, \quad j = 1, \ldots, J.
$$

(21)

The first model (20) is known as the discrete-time version of Cox’s (1972) proportional hazards model, typically used when handling tied death times. For this model, partial likelihood inference can be performed given the total numbers of deaths in the $J$ tables, but is computationally costly. In practice, the Breslow–Peto approximation is widely employed, although it can be less accurate than other options (Efron, 1977). By comparison, the second model (21) has received limited attention, but seems more suitable to be called a proportional hazards model because it is directly concerned with the hazard ratios $p_{11j}/p_{21j}$. Nevertheless, a remarkable finding which motivates our interest in model (21) is that the Breslow–Peto approximation to the exact partial likelihood estimator of $\beta^*$ in model (20) yields a consistent estimator of $\gamma^*$ in model (21).

3.1. Point estimation

For models (20) and (21), point estimators of $\beta^*$ and $\gamma^*$ can be directly adopted from Section 2 by the transformation of the survival data into $2 \times 2$ tables. Such connections are in fact exploited in the construction of estimators in Section 2.

For model (20), the weighted Mantel–Haenszel estimator $\hat{\beta}(w)$ is a solution to $\tau(w)(\beta) = 0$ with $\tau(w)(\beta)$ defined in (4).

Computation of $\hat{\beta}(w)$ is straightforward due to the simplicity of the estimating function $\tau(w)(\beta)$. By comparison, the exact partial likelihood estimator $\hat{\beta}(c)$ is a solution to $s(\beta) = 0$, with $s(\beta)$ defined in (6). Computation of $\hat{\beta}(c)$ is burdensome because the conditional mean $\mu_{11j}(\beta)$ is intractable when the total number of deaths, $n_{11j} + n_{21j}$, is large. The widely used Breslow–Peto approximation to the partial likelihood leads to the estimating function

$$
\sum_{j=1}^{J} \sum_{1 \leq i \leq N : Y_i = t_j, \delta_i = 1} \left[ 1|Z_i = 1\right] \frac{N_{1j}\psi_j(\beta)}{N_{1j}\psi_j(\beta) + N_{2j}} \right] x_j.
$$

By the definitions of $(n_{11j}, n_{21j})$, this function can be calculated as

$$
\sum_{j=1}^{J} \left\{ n_{11j} - \frac{(n_{11j} + n_{21j})N_{1j}\psi_j(\beta)}{N_{1j}\psi_j(\beta) + N_{2j}} \right\} x_j.
$$

(257)
which coincides with \( \zeta^{(w)}(\gamma) \) in (14) and (15), except for the change between \( \beta \) and \( \gamma \), with \( \psi(\beta) = \exp(\chi_1^T \beta) \) and \( \phi(\gamma) = \exp(\chi_1^T \gamma) \) as defined in Section 2.

**Proposition 5.** The estimator defined by the Breslow–Peto approximation to the partial likelihood estimator for \( \beta^* \) in odds ratio model (20) is algebraically identical to the estimator \( \hat{\gamma}^{(w)} \) for \( \gamma^* \) in probability ratio model (21).

The probability ratios \( p_{1ij}/p_{2ij} \) are always closer to 1 than the odds ratios \( p_{1ij}p_{23j}/(p_{12j}p_{21j}) \), except when \( p_{11j} = p_{21j} \). This result agrees with the previous finding that the Breslow–Peto approximation produces a conservative bias in estimating regression coefficients too close to 0 in proportional hazards models (Cox and Oakes, 1984).

### 3.2. Variance estimation

We derive model-based and model-robust estimators of the asymptotic variance of the weighted Mantel–Haenszel estimator \( \hat{\gamma}^{(w)} \) associated with model (20) or respectively the Breslow–Peto estimator \( \hat{\gamma}^{(w)} \) associated with model (21). As mentioned earlier, a complication is that the stratum sizes \( N_{1j}, N_{2j} \) are random and the \( 2 \times 2 \) tables constructed from risk sets over time are correlated over time. However, this difficulty can be handled by deriving appropriate influence functions, which take into account the contributions over time from each individual in the sample.

First, we describe the asymptotic distribution of \( \hat{\gamma}^{(w)} \) and a model-robust variance estimator, while allowing for possible misspecification of model (21). See the Appendix Remark 9 for a comparison of our result with Lin and Wei (1989). The following notation is used. For \( z = 1, 2 \), let \( P_{zj} = P(Y \geq t_j, Z = z) \) and \( P_{2zj} = P(Y = t_j, \delta = 1, Z = z) \), which are the unconditional probabilities of an individual being included in the risk set associated with time \( t_j \) or respectively observed to experience an event at time \( t_j \). The hazard is then identified by \( p_{2zj} = P_{2zj}/P_{zj} \) in (19) under conditionally independent censoring. Denote the corresponding indicators as \( I_{zj}(Y, Z) = 1\{Y \geq t_j, Z = z\} \) and \( I_{2zj}(Y, \delta, Z) = 1\{Y = t_j, \delta = 1, Z = z\} \).

Define

\[
h_j(Y, \delta, Z; \gamma) = \frac{P_{2j}}{P_{1j}\phi(\gamma) + P_{2j}(P_{11j} - P_{11j}I_{1j})} - \frac{P_{1j}\phi(\gamma)}{P_{1j}\phi(\gamma) + P_{2j}(I_{2j} - P_{22j}I_{2j})} + \frac{P_{11j} - \phi(\gamma)p_{21j}}{|P_{1j}\phi(\gamma) + P_{2j}|^2(P_{22j}I_{1j} + \phi(\gamma)p_{21j}I_{2j})}.
\]

For a column vector \( b \), denote \( b^{0:2} = bb^T \). The matrix \( N\hat{\gamma}^{(w)} \) is identical to \( N\hat{\gamma}^{(w)} \) in Section 2.2.2, whereas \( \hat{\gamma}^{(w)} \) and \( \hat{\gamma}^{(w)} \) differ from \( \hat{\gamma}^{(w)} \) and \( \hat{\gamma}^{(w)} \) even after rescaling.

**Proposition 6.** Assume that \( P(T > t_j) \geq p_0 \) for a constant \( p_0 > 0 \).

(i) \( \hat{\gamma}^{(w)} \) converges in probability to a target value \( \gamma \) which solves the equation

\[
0 = \sum_{j=1}^J \frac{P_{1j}p_{2j}}{P_{1j}\phi(\gamma) + P_{2j}} (P_{11j} - \phi(\gamma)p_{21j})x_j.
\]

Moreover, \( N^{1/2}(\hat{\gamma}^{(w)} - \gamma) \) converges in distribution to \( N(0, V^{(w)}) \), with \( V^{(w)} = D^{(w)T}(\gamma)C^{(w)}(\gamma)D^{(w)}(\gamma) \), where

\[
C^{(w)}(\gamma) = \text{var}\left\{ \sum_{j=1}^J h_j(Y, \delta, Z; \gamma)x_j \right\},
\]

\[
D^{(w)}(\gamma) = \sum_{j=1}^J (P_{1j}p_{1j} + P_{2j}p_{2j}) \frac{P_{1j}p_{2j}\phi(\gamma)}{|P_{1j}\phi(\gamma) + P_{2j}|^2} x_jx_j^T.
\]

(ii) A consistent estimator of \( V^{(w)} \) is \( \hat{V}^{(w)} = \hat{D}^{(w)T}(\gamma)\hat{C}^{(w)}(\gamma)\hat{D}^{(w)}(\gamma) \), where \( \hat{C}^{(w)}(\gamma) = N^{-1} \sum_{j=1}^J \hat{h}_j(Y, \delta, Z; \gamma)x_jx_j^T \).

As shown in the proof, \( h_j(Y, \delta, Z; \gamma) \) is obtained by linearizing the function \( n_{1j}N_{2j} - \phi(\gamma)n_{21j}N_{1j} / \{N_{1j}\phi(\gamma) + N_{2j}\} \) as a sample average \( \sum_{i=1}^N n_{1ij}n_{2ij}Y_{1ij} / \{n_{1ij}\phi(\gamma) + N_{2ij}\} \), with \( \gamma \) evaluated at \( \hat{\gamma}^{(w)} \), which captures the contribution of all individuals to the \( j \)th term of \( \zeta^{(w)}(\gamma) \) in (14), calculated from the risk set associated with time \( t_j \). If model (21) is correctly specified, then \( \hat{\gamma}^{(w)} = \gamma^* \) and hence \( \hat{\gamma}^{(w)} \) is a consistent estimator of \( \gamma^* \). Moreover, the third term of \( h_j(Y, \delta, Z; \gamma) \) reduces to 0 with \( \gamma = \gamma^* \), because \( p_{11j} - \phi(\gamma)p_{21j} = 0 \) in this case. This simplification can also be seen from the fact that \( n_{1j}N_{2j} - \phi(\gamma)n_{21j}N_{1j} \) changes sign as \( \gamma \) changes from \( \gamma^* \) to 0, forms a martingale difference when successively conditioning on the \( j \)th risk set, and hence the effect of the variation in the stratum sizes \( N_{1j}, N_{2j} \) is negligible, of the order \( o_p(N^{-1/2}) \). If model (21) is misspecified, then \( p_{11j} - \phi(\gamma)p_{21j} \) is in general nonzero with \( \gamma = \gamma^{(w)} \) and hence the third term of \( h_j(Y, \delta, Z; \gamma) \) is needed.
The preceding discussion shows that if model (21) is correctly specified, then a consistent estimator of $V^{(w_2)}$ is $\hat{V}^{(w_2)}$, defined as $\hat{V}^{(w_2)}$ except with the third term of $h_i(Y, \delta, Z; \gamma)$ removed. The variance estimator $N^{-1}\hat{V}^{(w_2)}$ is then identical to $N^{-1}\hat{V}^{(w_2)}$ in Section 2.2.2 except for $(N_{ij} - 1, N_{2j} - 1)$ used in $\hat{h}_i(\gamma)$. Such a small-sample adjustment is technically not needed here because $P(T > t_j)$ is assumed to be bounded away from 0. This boundedness condition is standard in large sample theory for survival analysis (e.g., Andersen et al., 1993, Section IV.3.2), although further investigation can be of interest. Alternatively, the preceding discussion also shows that another valid model-based variance estimator for $\hat{\gamma}^{(w)}$ is $N^{-1}\hat{V}^{(w_b)}$, which, by Corollary 1, can be directly compared with the existing model-based variance estimator for $\hat{\gamma}^{(w)}$.

**Corollary 1.** If model (21) is correctly specified, then $\hat{\gamma}^{(w)}$ is a consistent estimator of $\gamma^*$ with asymptotic variance $N^{-1}\hat{V}^{(w_2)}$, where the third term of $h_i(Y, \delta, Z; \gamma)$ can be removed. Moreover, $\hat{V}^{(w_2b)} = (N/N_*)\hat{V}^{(w_b)}$, with $\hat{V}^{(w_b)}$ defined in Section 2.2.2, is a consistent estimator of $\hat{V}^{(w_2)}$.

(ii) The variance estimator $N^{-1}\hat{V}^{(w_2b)}$ is, in the order on variance matrices, no greater than $N^{-1}\hat{D}^{(w_2)} - 1(\hat{\gamma}^{(w)})$, which is the commonly reported model-based variance estimator for the Breslow–Peto estimator $\hat{\gamma}^{(w)}$ when used as an approximation to the exact partial likelihood estimator in model (20).

Next, we describe the asymptotic distribution of $\hat{\beta}^{(w)}$ and a model-robust variance estimator, while allowing for possible misspecification of model (20). The same notation is used as above. Define

$$g_i(Y, \delta, Z; \beta) = \frac{P_{ij} p_{2j} + \psi_j(\beta) p_{12j}}{P_{ij} \psi_j(\beta) + P_{2j}} (l_{1ij} - p_{1ij} l_{1j}) - \frac{P_{ij} p_{1j} + \psi_j(\beta) p_{12j}}{P_{ij} \psi_j(\beta) + P_{2j}} (l_{2ij} - p_{2ij} l_{2j}) + \frac{P_{ij} p_{2j} - \psi_j(\beta) p_{12j}}{P_{ij} \psi_j(\beta) + P_{2j}} (p_{2j} l_{1j} + \psi_j(\beta) p_{1j} l_{2j}).$$

The matrix $N\hat{H}^{(w_2)}$ is identical to $N\hat{H}^{(w)}$ in Section 2.1.2, whereas $\hat{\Sigma}^{(w_2)}$ and $\hat{\Sigma}^{(w)}$ differ from $\hat{\Sigma}^{(w)}$ and $\hat{\Sigma}^{(w)}$ even after rescaling.

**Proposition 7.** Assume that $P(T > t_j) \geq p_0$ for a constant $p_0 > 0$.

(i) $\hat{\beta}^{(w)}$ converges in probability to a target value $\beta^{(w)}$ which solves the equation

$$0 = \sum_{j=1}^{J} \frac{P_{ij} p_{2j}}{P_{ij} \psi_j(\beta) + P_{2j}} (p_{1ij} p_{2j} - \psi_j(\beta) p_{12j}) x_j.$$

Moreover, $N^{1/2}(\hat{\beta}^{(w)} - \beta^{(w)})$ converges in distribution to $N(0, \Sigma^{(w,2)})$, with $\Sigma^{(w,2)} = H^{(w,2)} - 1(\beta)G^{(w,2)}(\beta)H^{(w,2)} - 1(\beta)_{\beta=\hat{\beta}^{(w)}},$ where

- $G^{(w,2)}(\beta) = \text{var} \left\{ \sum_{j=1}^{J} g_i(Y, \delta, Z; \beta) x_j \right\}$,
- $H^{(w,2)}(\beta) = \frac{1}{N} \sum_{j=1}^{J} \left( P_{ij} p_{1ij} p_{2j} + P_{2j} p_{12j} p_{2j} \right) \frac{P_{ij} p_{2j} \psi_j(\beta)}{P_{ij} \psi_j(\beta) + P_{2j}} x_j^T x_j.$

(ii) A consistent estimator of $\Sigma^{(w,2)}$ is $\hat{\Sigma}^{(w,2)} = \hat{H}^{(w,2)} - 1(\beta)G^{(w,2)}(\beta)\hat{H}^{(w,2)} - 1(\beta)_{\beta=\hat{\beta}^{(w)}},$ where $\hat{G}^{(w,2)}(\beta) \equiv N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{I} g_i(Y, \delta_i, Z_i; \beta) x_i$, $\hat{g}_i(Y, \delta, Z; \beta)$ is defined as $g_i(Y, \delta, Z; \beta)$, and $\hat{H}^{(w,2)}(\beta)$ defined as $H^{(w,2)}(\beta)$, with $(P_{ij}, P_{2j})$ replaced by $(\hat{P}_{ij}, \hat{P}_{2j}) = (N_{ij}, N_{2j}) / N$ and $(p_{1ij}, p_{12j}, p_{2j})$ replaced by $(\hat{p}_{1ij}, \hat{p}_{12j}, \hat{p}_{2j}).$

The preceding results exhibit a remarkable similarity to Proposition 6. If model (20) is correctly specified, then $\hat{\beta}^{(w)} = \beta^* \approx \beta^{(w)}$ and hence $\hat{\beta}^{(w)}$ is a consistent estimator of $\beta^*$. Moreover, the third term of $g_i(Y, \delta, Z; \beta)$ reduces to 0 with $\beta = \beta^*$. Hence $\hat{\Sigma}^{(w,2)}$ can be simplified as $\hat{\Sigma}^{(w,2b)}$, defined as $\hat{\Sigma}^{(w,2)}$ with the third term of $g_i(Y, \delta, Z; \gamma)$ removed. The variance estimator $N^{-1}\hat{\Sigma}^{(w,2b)}$ for $\hat{\beta}^{(w)}$ is then identical to $N^{-1}\hat{\Sigma}^{(w)}$ in Section 2.1.2 except with $\hat{g}_i(\gamma)$ modified such that its third term is removed and $(N_{ij}, N_{2j})$ are used instead of $(N_{ij} - 1, N_{2j} - 1)$. Alternatively, another valid model-based variance estimator for $\hat{\beta}^{(w)}$ is $N^{-1}\hat{\Sigma}^{(w,b)}$ derived from Robins et al. (1986) in Section 2.1.2.

**Corollary 3.** If model (20) is correctly specified, then $\hat{\beta}^{(w)}$ is a consistent estimator of $\beta^*$ with asymptotic variance $N^{-1}\hat{\Sigma}^{(w,2)}$, where the third term of $g_i(Y, \delta, Z; \gamma)$ can be removed. Moreover, $\hat{\Sigma}^{(w,2b)} = (N/N_*)\hat{\Sigma}^{(w,b)}$, with $\hat{\Sigma}^{(w,b)}$ defined in Section 2.1.2, is a consistent estimator of $\Sigma^{(w,2)}$.

So far, our development assumes that the survival data are recorded in pre-specified intervals, as commonly found in practice. From a theoretical perspective, consider the setting where the survival time $T$ is absolutely continuous with hazard functions $\lambda(t|Z)$. Both models (20) and (21) can be seen to reduce to a Cox proportional hazards model

$$\lambda(t|Z = 1) = \lambda(t|Z = 0) \exp[\alpha^T X(t)],$$

(24)
where \( \alpha^* \) is an unknown coefficient vector. A natural question is how the proposed point and variance estimators behave, with time intervals chosen to be sufficiently small such that there is a total of at most one death at \( t_j \) in each interval \( (t_{j-1}, t_j] \). By construction, both \( \hat{\beta}^{(w)} \) and \( \hat{\gamma}^{(w)} \) coincide with the conditional likelihood estimator \( \hat{\beta}^{(c)} \), i.e., the partial likelihood estimator of \( \alpha^* \) without tied death times. Moreover, as mentioned after Corollary 1, the model-based variance estimator \( N^{-1} \hat{V}^{(w,2b)} \) for \( \hat{\beta}^{(w)} \) is identical to the usual model-based variance estimator \( N^{-1} \hat{D}^{(w,2)}(\hat{\beta}^{(w)}) \) for \( \hat{\beta}^{(c)} \). By Appendix Remark 7 and Robins et al. (1986, Section 5), the model-based variance estimator \( N^{-1} \hat{\Sigma}^{(w,2)}(\hat{\beta}^{(c)}) \) is also a consistent variance estimator for \( \hat{\beta}^{(c)} \) if model (24) is correctly specified. Finally, by Appendix Remark 9, the model-robust variance estimator \( N^{-1} \hat{\psi}^{(w,2)}(\hat{\beta}^{(w)}) \) coincides with Lin and Wei’s (1989) variance estimator. Such a relationship does not hold for the model-robust variance estimator \( N^{-1} \hat{\Sigma}^{(w,2)}(\hat{\beta}^{(w)}) \) from these remarks, we obtain the following result.

**Corollary 4.** If model (24) is correctly specified, then \( \hat{\beta}^{(w)} = \hat{\gamma}^{(w)} \) is a consistent estimator of \( \alpha^* \) with an asymptotic variance matrix which can be consistently estimated by \( N^{-1} \hat{\Sigma}^{(w,2b)}(\hat{\beta}^{(w)}) \) and \( N^{-1} \hat{\psi}^{(w,2b)}(\hat{\beta}^{(c)}) \). Moreover, the asymptotic variance matrix of \( \hat{\gamma}^{(w)} \) can be consistently estimated by \( N^{-1} \hat{V}^{(w,2b)} \), with possible misspecification of model (24).

Although both models (20) and (21) reduce to model (24) in the continuous-time limit, the estimator \( \hat{\gamma}^{(w)} \) and its model-based and model-robust variance estimators \( N^{-1} \hat{V}^{(w,2b)}(\hat{\beta}^{(c)}) \) and \( N^{-1} \hat{V}^{(w,2b)}(\hat{\beta}^{(c)}) \) associated with model (21) extend directly from discrete time to partial likelihood inference in the continuous-time limit. For model (20), model-robust variance estimation seems not directly extended for the estimator \( \hat{\beta}^{(w)} \) in the continuous-time limit, and also remains to be studied for the conditional likelihood estimator \( \hat{\beta}^{(c)} \). Further investigation is needed to fully address questions.

**4. Numerical studies**

**4.1. Analysis of 2 × 2 tables**

First, we reanalyze the data from the Oxford Childhood Cancer Survey (Breslow and Day, 1980). The survey is a retrospective study in which cases (children who died of cancer) and controls (children who were alive and well) were identified, and their exposure to in utero radiation were ascertained. Hence the factor is dying of cancer or not, and the response is radiation exposure or not. A total of 120 strata were used by age and year of birth. Following previous analyses, we fit three odds ratio models: (i) \( \log(\psi_j) = \beta_0 \), (ii) \( \log(\psi_j) = \beta_0 + \beta_1 x_j \), and (iii) \( \log(\psi_j) = \beta_0 + \beta_1 x_j + \beta_2 (x_j^2 - 22) \), where \( x_j \) indexes year of birth.

Table 3 presents the point estimates and associated standard errors. The weighted Mantel–Haenszel estimates and standard errors agree well with those from conditional likelihood inference, which is statistically the “gold standard” but computationally costly to implement. In fact, the wMH results agree with conditional likelihood inference more closely than those in Davis (1985). The (unweighted) Mantel–Haenszel estimates differ noticeably from conditional likelihood estimates, sometimes with inflated standard errors. For \( \beta_1 \) in model (ii), MH yields a point estimate different from CML by \( 0.0428/0.0385 - 1 = 11.2\% \) in absolute values, and a variance estimate greater than CML by \( (0.0162/0.144)^2 - 1 = 26.6\% \). The model-based and model-robust standard errors appear to be aligned with each other here.

To further compare estimators, we conduct simulations under various settings similar to previous studies (e.g., Hauck et al., 1982; Robins et al., 1986). In particular, Table 4 presents results from 2000 repeated simulations in the following two settings, where the success probabilities \( p_{11j}, p_{21j} \) are close to 0. See the Supplement for additional results. For the first setting, \( J = 40 \) tables are simulated with log odds ratios \( \psi_j = \log(2) \), probabilities \( p_{21j} = .03 + .001j \) between .031 and .07 for \( j = 1, \ldots, 40 \), and binomial sizes \( (N_{1j}, N_{2j}) = (16, 4) \) for \( j = 1, \ldots, 20 \) and \( (4, 16) \) for \( j = 21, \ldots, 40 \). The second setting is the same as the first, except that \( p_{11j} \) and \( p_{21j} \) are related with log probability ratios \( \phi_j = \log(2) \)
for \( j = 1, \ldots, 40 \). For each setting, models (1) and (11) are fit with constant covariates, corresponding to common odds ratios or probability ratios, which are valid only in the first or second setting respectively.

The following observations can be obtained from Table 4. In the first setting, the weighted Mantel–Haenszel and conditional maximum likelihood estimators perform similarly to each other, with smaller biases and variances than the unweighted Mantel–Haenszel estimator. The Breslow–Peto estimator, by our calculation or equivalently R package \texttt{survival}, is downward biased by \( 1 - 0.6376/0.6931 = 8.0\% \), as BP can be seen to be centered around a target log odds ratio which is smaller than the common log odds ratio. The model-based variance estimator commonly reported for BP is biased upward by \( (.3349/.3156)^2 - 1 \), whereas the proposed model-robust variance estimator (as well as the new model-based variance estimator, although not guaranteed by theory) reasonably matches the Monte Carlo variance.

In the second setting, BP is centered around the true log probability ratio with a negligible bias as expected. The commonly used variance estimator for BP is biased upward by \( (.3310/.326)^2 - 1 \), whereas both the proposed model-based and model-robust variance estimators agree properly with the Monte Carlo variance. The estimators, MH, wMH, and CML, are upward biased compared with the true log probability ratio log(2), as these estimators can be considered to be centered about a target log odds ratio which is greater than the common log probability ratio.

### 4.2. Two-sample survival analysis

First, we perform two-sample analysis of the data on a Veteran’s Administration lung cancer trial used in Kalbfleisch and Prentice (1980). The data consist of 137 observations with right-censored survival time (apparently in days) and several covariates. For two-sample analysis, the two groups are defined by a treatment variable \( Z \), labeled as 1 or 2 if test or standard. Kaplan–Meier survival curves suggest non-proportional hazards over time in the two groups (see the Supplement). Hence we fit odds ratio model (20) and probability ratio model (21), with the time functions \( x(t) = (1, x_1(t), x_2(t))^T \), where \( x_1(t) = 1(100 < t \leq 200) \) and \( x_2(t) = 1(t > 200) \). In addition, to study discrete-time inference, we also apply various estimators to further discretized data, obtained by grouping the original times in intervals of 20 days. For concreteness, the censored-late option is used as discussed in the Supplement, i.e., 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 \( t_{j-1} \).

Table 5 presents the results on the original and discretized data. For the original data with a small number of tied deaths, all the estimates obtained are similar to each other in various degrees, although the BP point estimates are slightly closer to 0 than the estimates from MH, wMH, and CML, and the model-based variance estimates on the row oldBP are larger than those on the row BP as expected by Corollary 2.

For the discretized data with more tied deaths, the BP point estimates are more substantially closer to 0 than the estimates from MH, wMH, and CML, which remain relatively similar to each other. This difference can be properly explained by the fact that BP estimates are associated with odds ratios, whereas the other estimates are associated with probability ratios, rather than poor approximation of BP to CML as would often be claimed. The commonly reported model-based variance estimates on the row oldBP are also more inflated compared with the proposed variance estimates on the row BP. For example, for estimation of \( \beta_2 \) with discretized data, the BP point estimate is smaller than CML in absolute values by \( 1 - .8881/1.0054 = 11.7\% \), and the oldBP variance estimate is larger than the proposed BP variance estimate by \( (.5279/.4953)^2 - 1 = 13.6\% \).

For both the original and discretized data, the Efron point and variance estimates fall between those from CML and oldBP. The interpretation of these estimates can be ambiguous. Efron’s method is derived as an improved approximation over Breslow–Peto to CML for odds ratio estimation. But a precise meaning for the population limits of Efron’s estimators remains to be identified in general with discrete data.

To further compare estimators, we also conduct a simulation study. For each simulation, a sample of size \( n = 200 \) is generated as follows. The group variable \( Z \) is generated as 1 or 2 with probability .5 each. The event time \( T \) is generated as Weibull with shape and scale parameters 2 and 1 or respectively 1 and 1 in the first or second group. The censoring time \( C \) is generated as 4 times Beta(2, 2) in the first group, and uniform(0, 4) in the second group. Two sets of observed
Extensions can be pursued for structural modeling in causal inference, regression settings for studying associations with multiple discrete or continuous covariates in Tan (2020), and further the latter approximation, the resulting estimates can be difficult to interpret or misleading. Our approach is extended to terms of well-defined population quantities. This is in contrast with previous methods, where two types of approximations are confounded, first in specifying continuous-time models and then in deriving numerically tractable estimation. Due to the latter approximation, the resulting estimates can be difficult to interpret or misleading. Our approach is extended to regression settings for studying associations with multiple discrete or continuous covariates in Tan (2020), and further extensions can be pursued for structural modeling in causal inference.

5. Conclusion

We develop new methods and theory using estimating functions along with model-based and model-robust variance estimation for analysis of $2 \times 2$ tables and two-sample survival data. Both odds and probability ratio models are formulated directly in a discrete scale as parsimonious approximations to the unknown underlying, possibly continuous-time, process. Our point and variance estimators are defined in a numerically exact manner, and can be interpreted in terms of well-defined population quantities. This is in contrast with previous methods, where two types of approximations are confounded, first in specifying continuous-time models and then in deriving numerically tractable estimation. Due to the latter approximation, the resulting estimates can be difficult to interpret or misleading. Our approach is extended to regression settings for studying associations with multiple discrete or continuous covariates in Tan (2020), and further extensions can be pursued for structural modeling in causal inference.
Appendix

In the following remarks, we compare the proposed estimator \( \hat{\beta}^{(w)} \) with the conditional likelihood estimator \( \hat{\beta}^{(c)} \) and the estimator of Davis (1985) in analysis of 2 \( \times \) 2 tables.

**Remark 3.** There are several reasons why the proposed estimator \( \hat{\beta}^{(w)} \) can be worthwhile compared with the conditional likelihood estimator \( \hat{\beta}^{(c)} \), even though \( \hat{\beta}^{(c)} \) seems attractive on various grounds (e.g., Lindsay, 1980; Hauck, 1984). First, it is straightforward to compute \( \hat{\beta}^{(w)} \) with estimating function \( \tau^{(w)}(\beta) \) in a closed form, whereas computation of \( \hat{\beta}^{(c)} \) is difficult for tables with large marginals due to the complexity in numerical evaluation of \( \mu_{11}(\beta) \). See Breslow and Cologne (1986) and McCullagh and Nelder (1989) for approximate methods for \( \hat{\beta}^{(c)} \). Second and more importantly, \( \tau^{(w)}(\beta) \) is an unbiased estimating function under model (1), and hence consistency can be directly established by standard large sample theory. The approximate methods for \( \hat{\beta}^{(c)} \) are numerically motivated, but their statistical properties remain to be studied. Finally, analytical simplicity of \( \tau^{(w)}(\beta) \) also facilitates model-robust variance estimation with possible misspecification of model (1) as discussed in Section 2.1.2.

**Remark 4.** Davis (1985) considered a class of estimating functions,

\[
\sum_{j=1}^{l} \left[ g_{ij}(\beta) \tilde{P}_{1j} \tilde{P}_{2j} - g_{2j}(\beta) \psi_{j}(\beta) \tilde{P}_{12j} \tilde{P}_{21j} \right] x_{j},
\]

where \( g_{ij}(\beta) = g(\beta; n_{1i1}, n_{1i2}, n_{2i1}, n_{2i2}) \) and \( g_{2j}(\beta) = g(\beta; n_{11j} + 1, n_{12j} - 1, n_{21j} - 1, n_{22j} + 1) \) for some scalar, possibly data-dependent function \( g \). It is shown that (25) is an unbiased estimating function, conditionally on the marginals \((n_{11j} + 1, n_{12j}, n_{21j}, n_{22j})\), \( j = 1, \ldots, J \), although such conditional unbiasedness seems not further pursued. In fact, an optimal choice of \( g \) restricted to be non-random and hence \( g_{ij} = g_{2j} \) is stated, without proof, first in the same form as \( \rho_{1j}(\beta) \) in Proposition 1, and then in an expression which appears to disagree with our calculation in the proof of Proposition 2(ii).

Nevertheless, a concrete choice of \( g \) was then proposed, but differently from our choice \( \rho_{1j}(\beta) \). In numerical examples of Breslow and Cologne (1986), the estimator of Davis (1985) was found to sometimes differ noticeably from the conditional likelihood estimator \( \hat{\beta}^{(c)} \).

In the following remark, we discuss variance estimation for the (unweighted) Mantel–Haenszel estimator \( \hat{\beta}^{(0)} \), with possibly non-constant covariates.

**Remark 5.** For the (unweighted) Mantel–Haenszel estimator \( \hat{\beta}^{(0)} \), similar results can be obtained as in Section 2.1.2 for the weighted Mantel–Haenszel estimator \( \hat{\beta}_{j}^{(w)} \). In fact, a model-based estimator for the asymptotic variance of \( \hat{\beta}^{(0)} \) is \( \hat{\Sigma}(\hat{\beta}^{(0)}) = \hat{H}^{(0)}(\beta) \times \hat{\Sigma}^{(0)}(\beta) \), where \( \hat{H}^{(0)}(\beta) = \sum_{j=1}^{J} \rho_{j}^{(0)}(\beta) \tilde{P}_{12j} \tilde{P}_{21j} x_{j} x_{j}^{T} \), \( \hat{\Sigma}^{(0)}(\beta) = \sum_{j=1}^{J} \rho_{j}^{(0)}(\beta) \tilde{\sigma}_{j}^{2}(\beta) x_{j} x_{j}^{T} \) and \( \hat{\Sigma}^{(0)}(\beta) = \sum_{j=1}^{J} \rho_{j}^{(0)}(\beta) \tilde{\sigma}_{j}^{2}(\beta) x_{j} x_{j}^{T} \). With nonconstant covariates, these variance estimators are generally not invariant to either the exchange between the factor levels or between the response values, similarly as the point estimator \( \hat{\beta}^{(0)} \) discussed in Remark 1. With constant covariates, however, \( \hat{\Sigma}(\hat{\beta}^{(0)}) \) becomes invariant to the exchange between the factor levels and between the response values, similarly as the point estimator \( \hat{\beta}^{(0)} \). The estimator \( \hat{\Sigma}(\beta(\hat{\beta}^{(0)})) \), which reduces to a variance estimator proposed in Robins et al. (1986), remains invariant to the exchange between the factor levels, but not to that between the response values. Hence a symmetrized version of \( \hat{\Sigma}(\beta(\hat{\beta}^{(0)})) \) is also proposed in Robins et al. (1986) to achieve two-way invariance.

In the following remarks, we discuss the variance estimators \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) for \( \hat{\beta}^{(w)} \) and \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) for \( \hat{\gamma}^{(w)} \), conditional inference in the case of a total of one success per table, and Wald and score-type tests in analysis of 2 \( \times \) 2 tables.

**Remark 6.** Consider the special case where \( n_{11j} = 0 \) or \( n_{21j} = 0 \) for \( j = 1, \ldots, J \) (including but not restricted to a total of one success in each table). The two estimates \( \hat{\beta}^{(w)} \) and \( \hat{\gamma}^{(w)} \) are identical to each other, as seen from the second equality in (7). The matrix \( \hat{H}^{(w)}(\beta) \) can also be shown to be identical to \( \hat{D}^{(w)}(\gamma) \) and \( \hat{C}^{(w)}(\gamma) \), but \( \hat{C}^{(w)}(\beta) \) does not seem to be related to \( \hat{C}^{(w)}(\gamma) \) in a simple way, with both \( \beta \) and \( \gamma \) evaluated at \( \hat{\beta}^{(w)} = \hat{\gamma}^{(w)} \). Hence the model-based variance estimate \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) for \( \hat{\beta}^{(w)} \) may differ from \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) and \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) for \( \hat{\gamma}^{(w)} \), even though \( \hat{\beta}^{(w)} = \hat{\gamma}^{(w)} \).

**Remark 7.** We mainly study unconditional inference, with possible model misspecification. Conditional inference has been well established given four marginals in each table under odds ratio model (1) (Breslow, 1981; McCullagh and Nelder, 1989), although how to perform conditional inference under probability ratio model (11) is not clear. In our approach, both \( \hat{\beta}^{(w)} \) and \( \hat{\gamma}^{(w)} \) are constructed to coincide with the conditional likelihood estimator \( \hat{\beta}^{(c)} \) given a total of one success in each table under model (1). Moreover, if model (1) is valid, then given a total of one success in each table, \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) is a consistent variance estimator for \( \hat{\beta}^{(w)} = \hat{\gamma}^{(w)} \) by Robins et al. (1986), Section 5, and so is the usual variance estimator \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) or \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \). An interesting difference, however, is that \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \) and \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \), but not \( N_{-1}^{-1} \hat{\Sigma}^{(wb)} \), depend on the binary data \( n_{11j} \) only through the point estimate \( \hat{\beta}^{(w)} = \hat{\gamma}^{(w)} \).
Remark 8. Based on point and variance estimation, Wald and score-type tests can be derived by standard arguments (e.g., Lin and Wei, 1989). Of particular interest is to test the null hypothesis that $p_{ij1} = p_{ij2}$ for $j = 1, \ldots, J$, which can be expressed as $\beta^* = 0$ in model (1) or equivalently $\gamma^* = 0$ in model (11), with constant covariates. For testing $\gamma^* = 0$, a score-type test statistic based on $\xi^*(\gamma)$ is $\xi^*(\gamma)/\sqrt{\sum_{j=1}^{J} \hat{q}_j^w(0) \hat{v}_j^*(0)}/\sqrt{\sum_{j=1}^{J} N_j}$. Under the null hypothesis, the variance estimator $\hat{\sigma}^*(\gamma)$ can be replaced by another unbiased estimator of $\text{var}(\hat{\beta}_{1j} - \hat{\beta}_{2j})$ defined as

$$
\hat{\sigma}^*(\gamma) = \hat{\sigma}^*(\gamma) = \left[ \sum_{j=1}^{J} N_j \hat{v}_j^*(0) \right]^{-1/2} \left[ \sum_{j=1}^{J} N_j \hat{v}_j^*(0) \right]^{1/2}
$$

which is identical to the log-rank test statistic. Similarly, for testing $\beta^* = 0$, the score-type test statistic based on $\tau^*(\beta)$ is $\tau^*(\beta)/\sqrt{\sum_{j=1}^{J} \hat{q}_j^w(0) \hat{v}_j^*(0)}/\sqrt{\sum_{j=1}^{J} N_j}$, which can also be shown to yield the log-rank test statistic if $\hat{v}_j^*(0)$ is replaced by the pooled-sample variance estimator $\hat{v}_j^*(0)$. These connections are reassuring, especially for two-sample survival analysis in Section 3, where the log-rank test is known to be applicable to survival data in both continuous and discrete time.

Finally, we compare model-robust variance estimation in Proposition 6 with Lin and Wei (1989) for two-sample survival analysis.

Remark 9. Proposition 6 can be deduced from Lin and Wei (1989), who derived a model-robust variance estimator for the partial likelihood estimator in Cox’s proportional hazards model. In the presence of tied death times, Lin & Wei’s result remains valid if the Breslow–Peto modification is used. Their variance estimator for $\hat{\gamma}(w)$ in the two-sample discrete-time setting can be expressed as $N^{-1} \hat{V}_w(2)$, with $\hat{V}_w(2)$ defined in Proposition 6 but $\hat{C}_w(2)(\gamma) = N^{-1} \sum_{j=1}^{J} \hat{h}_j^w(Y, \Delta, Z; \gamma) x_j^2$, where

$$
\hat{h}_j^w(Y, \Delta, Z; \gamma) = 1\{Y = t\} \delta \left( \frac{1}{(1 - \hat{N}_j \hat{\phi}_j(\gamma) + \hat{N}_j)} \right) - 1\{Y \geq t\} \left( \frac{\hat{N}_j \hat{\phi}_j(\gamma) \right) \frac{1}{(1 - \hat{N}_j \hat{\phi}_j(\gamma) + \hat{N}_j)}
$$

We show that $\hat{h}_j(Y, \Delta, Z; \gamma)$ algebraically coincides with $\hat{h}_j^w(Y, \Delta, Z; \gamma)$ in the proof of Proposition 6, and hence our model-robust variance estimator for $\hat{\gamma}(w)$ agrees with Lin and Wei’s (1989). On one hand, this agreement is expected because there is a unique model-robust influence function for $\hat{\gamma}(w)$, whether model (20) or (21) is valid or not. On the other hand, as discussed above, our representation with $\hat{h}_j(Y, \Delta, Z; \gamma)$ is explicit in separating what variation persists and what vanishes if model (21) is valid. Accounting for the former component leads to a new model-based variance estimator for $\hat{\gamma}(w)$, which is no greater than the usually used model-based variance estimator for $\hat{\gamma}(w)$.

Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2022.05.002.

References

Breslow, N.E., 1981. Odds ratio estimators when the data are sparse. Biometrics 37, 483–492.


