Statistica Sinica Preprint No: SS-2024-0391	
Title	Robust Score Tests for Censored Outcomes and Incomplete Covariates Leveraging High-Dimensional Auxiliary Variables
Manuscript ID	SS-2024-0391
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0391
Complete List of Authors	Jiahui Feng and Kin Yau Wong
Corresponding Authors	Kin Yau Wong
E-mails	kin-yau.wong@polyu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

ROBUST SCORE TESTS FOR CENSORED OUTCOMES AND INCOMPLETE COVARIATES LEVERAGING HIGH-DIMENSIONAL AUXILIARY VARIABLES

Jiahui Feng[1] and Kin Yau Wong[1,2]

[1] *The Hong Kong Polytechnic University*

[2] *Hong Kong Polytechnic University Shenzhen Research Institute*

Abstract: In many cancer genomic studies, investigators are interested in testing the presence of association between a time-to-event outcome and covariates of interest. Such analyses are often complicated by missing data. When covariates of interest are missing for some subjects, it is desirable to leverage information from observed auxiliary variables, which are sometimes high-dimensional, to improve statistical power. In this paper, we consider a class of semiparametric transformation models for a potentially right-censored survival outcome and develop an association test between the outcome and a partially observed covariate. We impute the missing covariate values using high-dimensional auxiliary variables. To accommodate potential model misspecification, we combine results from multiple plausible models for the survival time to improve power. We establish the validity of the test under misspecification of the outcome model and an adaptively-selected model for the incomplete covariate. We demonstrate the validity of the proposed methods and the superiority over existing methods through extensive simulation studies and applications to major cancer genomic studies.

Key words and phrases: Imputation; Missing data; Post-selection inference; Survival analysis; Variable selection.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

1. Introduction

In cancer genomic studies, investigators are often interested in identifying genomic factors associated with the time to events of interest, such as the time to tumor progression or death since initial diagnosis. The recent advent of high-throughput technologies has provided unprecedented opportunities for researchers to discover such associations from massive collections of data. For example, multiple studies have been conducted for breast (Lánczky et al., 2016), lung (Välk et al., 2011), and ovarian cancers (Sieh et al., 2013), among others, to identify genomic factors that are associated with the time to cancer relapse or death. One common challenge in the analysis of event times is that they are often not exactly observed, that is, censored. Also, full parametric assumptions on the event time distribution are usually in doubt, and investigators typically adopt more flexible semiparametric models. Censoring and nonparametric components in the model complicate the likelihood and pose challenges on estimation and inference.

Another complication often encountered in cancer genomic studies is missing data, where genomic factors of interest are not always observed. For example, The Cancer Genome Atlas (TCGA) program (https://cancergenome.nih.gov/) collected multiple types of clinical, genomic, epigenomic, transcriptomic, and proteomic data, while the proteomic data are missing for many subjects. Missing data may also arise by design. Especially when some variables are difficult or expensive to measure, a two-phase design is commonly adopted, where the outcome and inexpensive covariates are observed for all subjects in the first phase and a sub-group of subjects are selected for measurements on expensive covariates in the second phase. For instance, in the National Heart, Lung,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

and Blood Institute Exome Sequencing Project (https://evs.gs.washington.edu/EVS/), all subjects were measured for genotyping array data, and only a sub-group of subjects with extreme values of the primary outcome was selected for whole-exome sequencing. Statistical methods for the analysis of all study subjects need to accommodate missingness in the covariates.

The simplest method to deal with missing data is the complete-case analysis, where observations with incomplete data are discarded before the analysis. When the missingness is completely at random (MCAR) (Rubin, 1976), i.e., the missing mechanism does not depend on any relevant data, the complete-case analysis is valid but statistically inefficient, as it discards information contained in the incomplete observations. By contrast, the complete-case analysis may be inconsistent under the missing at random (MAR) mechanism, where the missingness depends on the observed data. Another approach to handle missing data is imputation, where the missing values are imputed by plausible values obtained from the observed data, and conventional methods can then be adopted to analyze the completed data. However, single imputation methods that do not account for the variability in the imputation generally yield invalid inference. Although many methods have been developed for estimation under incomplete data, few methods focus on association tests. Most existing association testing methods that accommodate incomplete covariates are based on the score test, which is formulated based on imputed data or the full likelihood that includes the incomplete variable model (Hu et al., 2015; Derkach, Lawless and Sun, 2015; Bjørnland et al., 2018; Lawless, 2018; Wong, Zeng and Lin, 2019). These studies focus only on low-dimensional models.

In this paper, we focus on the association test between a right-censored survival out-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

come and an incomplete covariate, where potentially high-dimensional auxiliary variables are available to predict the missing values of the covariate of interest. The proposed method is distinct from existing methods in two major respects. First, we impute the missing covariate values from an adaptively-selected set of auxiliary variables to improve power. Most existing works focus on a prespecified set of low-dimensional auxiliary variables, which may not be available in practice. Second, we consider multiple outcome models and combine the testing results to yield higher power. We rigorously prove that the proposed method preserves the type I error under model misspecification and under general selection approaches for the auxiliary variables.

In general, when a model is selected based on the observed data, the distribution of a statistic constructed from the model differs from that when the model is prespecified. Drawing inference based on a selected model, i.e., post-selection inference, is highly challenging. Some investigators considered conditional inference for the model parameters given a subset of selected covariates; see Fithian, Sun and Taylor (2017), Lee et al. (2016) and Tibshirani et al. (2016). Others focused on constructing uniformly valid confidence intervals regardless of the preceding model selection procedure; see Berk et al. (2013), Bachoc, Leeb and Pötscher (2019) and Bachoc, Preinerstorfer and Steinberger (2020). This line of work, nevertheless, is not directly applicable to the current problem, as in the current setting, the selected model is not the model of interest, and modifications of the inferential procedures due to model selection are in fact not necessary. Recently, Wong and Feng (2023) developed a score test under a generalized linear regression setting, where missing covariates are imputed from a set of adaptively-selected auxiliary variables. Nevertheless, the model is restricted to be fully parametric.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

To model a survival time, we adopt a general class of transformation models (Dabrowska and Doksum, 1988), which includes the proportional hazards (PH) model (Cox, 1972) and the proportional odds (PO) model (Bennett, 1983; Pettitt, 1984) as special cases. This class of models has been studied by Cheng, Wei and Ying (1995, 1997), Chen, Jin and Ying (2002), and Zeng and Lin (2006), among others. In the transformation models, there is a transformation parameter that is typically prespecified or chosen by some information criterion in practice. In this paper, instead of fixing the transformation parameter, we perform separate tests under different choices of the transformation parameter values and combine the results by taking the largest test statistic. We demonstrate that the proposed test tends to be more powerful than assuming a single (incorrect) model, so the proposed procedure is particularly useful when the true outcome model is unclear. Our theoretical development needs to account for model misspecification and thus is more challenging than the existing works on transformation models.

Although related, the current work represents substantial advances over our previous work in Wong and Feng (2023). First, Wong and Feng (2023) focused on fully parametric models, whereas we consider semiparametric transformation models. While parametric models may be suitable in specific cases, the majority of studies of event times prefer the semiparametric Cox model due to its flexibility and the availability of a simple partial likelihood for estimation and inference. Second, we develop an approach to accommodate multiple outcome models within a single framework, whereas Wong and Feng (2023) only considered a single outcome model.

The rest of this paper is structured as follows. In Section 2, we formulate the model, the hypothesis, and the proposed score test. In Section 3, we establish the asymptotic

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

properties of the proposed test. In Section 4, we report the results from simulation studies. In Section 5, we provide applications to real bladder urothelial carcinoma and breast cancer datasets. Finally, we conclude the paper with a few remarks. Technical details and additional numerical results are provided in the Appendix and Supplementary Material.

2. Methods

2.1 Model, hypothesis, and the imputation score test

Let T denote a time-to-event outcome, S denote a covariate of interest, $\boldsymbol{X}$ denote a vector of other covariates, and $\boldsymbol{A}$ denote a potentially high-dimensional vector of auxiliary variables. Assume that $S = \boldsymbol{\gamma}_X^{\mathrm{T}} \boldsymbol{X} + \epsilon_S$ for some regression parameter vector $\boldsymbol{\gamma}_X$, and ϵ_S is a random variable independent of $\boldsymbol{X}$. The null hypothesis of interest is

$$H_0 : \epsilon_S \text{ is independent of } T.$$

It states that besides the component explained by $\boldsymbol{X}$, S is independent of the event time.

In this paper, we focus on testing H_0 under a (set of) semiparametric transformation model(s). In particular, the model assumes that the cumulative hazard function of T conditional on $(\boldsymbol{X}, S)$ is

$$\Lambda(t \mid \boldsymbol{X}, S) = G\{\Lambda(t) \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \beta S)\}, \tag{2.1}$$

where Λ is an unknown increasing function in $[0, \tau]$ with τ being the end-of-study time and $\Lambda(0) = 0$, G is a pre-specified transformation function that is strictly increasing with $G(0) = 0$, and $\boldsymbol{\alpha}$ and β are regression parameters. For example, we may consider the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

class of Box–Cox transformations

$$G(x) = \begin{cases} \{(1+x)^{\rho} - 1\}/\rho & \text{for } \rho > 0, \\ \log(1+x) & \text{for } \rho = 0, \end{cases} \tag{2.2}$$

where ρ is a pre-specified transformation parameter. In this family, $\rho = 1$ corresponds to the PH model, and $\rho = 0$ corresponds to the PO model. Alternatively, we may consider the class of logarithmic transformations

$$G(x) = \begin{cases} r^{-1}\log(1+rx) & \text{for } r > 0, \\ x & \text{for } r = 0, \end{cases} \tag{2.3}$$

where r is a pre-specified transformation parameter. The choices of $r = 0$ and $r = 1$ correspond to the PH and PO models, respectively. Note that model (2.1) can be written as a linear transformation model

$$\log \Lambda(T) = -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} - \beta S + \epsilon_T,$$

where ϵ_T is an error term with $P(\epsilon_T < x) = 1 - \exp[-G\{\exp(x)\}]$. Particularly, the choices of the extreme value distribution and standard logistic error distribution for ϵ_T yield the PH and PO models, respectively. In this formulation, β can be interpreted as the linear effect of the covariate S on a transformation of T.

Suppose that T is possibly right-censored at C, which is assumed to be independent of $(T, S, \boldsymbol{A})$ given $\boldsymbol{X}$. Let $Y = \min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Also, suppose that S may be missing, and let R be the indicator of whether S is observed, i.e., $R = 1$ if S is observed, and $R = 0$ if otherwise. We assume that R is conditionally independent of $(S, \boldsymbol{A})$ given $(Y, \Delta, \boldsymbol{X})$. The observed data from a random sample of n subjects consist of $(Y_i, \Delta_i, \boldsymbol{X}_i, \boldsymbol{A}_i, R_i S_i, R_i)$ for $i = 1, \ldots, n$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Under the transformation model (2.1), we test the null hypothesis $H_0' : \beta = 0$, that the covariate of interest S does not have an effect on the (transformed) hazard of T given $\boldsymbol{X}$. Note that under (2.1), $\Lambda(t \mid \boldsymbol{X}, S) = G[\Lambda(t)\exp\{(\boldsymbol{\alpha} + \beta\boldsymbol{\gamma}_X)^{\mathrm{T}}\boldsymbol{X} + \beta\epsilon_S\}]$, so H_0' is equivalent to H_0. To construct a test, we first fit a working model of S against $(\boldsymbol{X}, \boldsymbol{A})$ and use this model to impute the missing values. Because the auxiliary variables $\boldsymbol{A}$ may be high-dimensional, we propose to select a low-dimensional subset of the components of $\boldsymbol{A}$ to construct the model of S. Let $\mathcal{K}$ denote the indices of the selected components of $\boldsymbol{A}$, where $\mathcal{K} \subset \{1, \ldots, p\}$ with p being the dimension of $\boldsymbol{A}$. We can select the components of $\boldsymbol{A}$ using existing variable selection approaches, such as lasso (Tibshirani, 1996) or feature screening (Fan and Lv, 2008; Fan and Song, 2010); a formal formulation of the selection procedure is given in Section 3. Let $\boldsymbol{W}_{\mathcal{K}}$ denote the vector that consists of $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ indexed by $\mathcal{K}$. We fit a working model of $S = \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}} + \delta$, where δ is a mean-zero error term, and $\boldsymbol{\gamma}_{\mathcal{K}}$ is a vector of regression parameters. We estimate $\boldsymbol{\gamma}_{\mathcal{K}}$ by solving $\sum_{i=1}^{n} R_i(S_i - \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\boldsymbol{W}_{\mathcal{K},i} = \boldsymbol{0}$, and let $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ denote the estimator. Note that this working model of S is introduced only to enhance the power of the test and does not alter the formulation of the outcome model or the null hypothesis.

We then perform a score test based on the outcome model (2.1) and the imputed values of S. First, we estimate $\boldsymbol{\alpha}$ and Λ under the null hypothesis. Because the model involves the nonparametric component Λ, we adopt the nonparametric maximum likelihood estimation (NPMLE) approach of Zeng and Lin (2007). In particular, we treat Λ as a step function with jumps only at the observed survival times. Let $t_1 < \cdots < t_m$ denote the set of observed survival times, with m being the number of unique observed survival times, and λ_k denote the jump size of Λ at t_k for $k = 1, \ldots, m$. The log-likelihood

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

function pertaining to $(\boldsymbol{\alpha}, \Lambda)$ is

$$\sum_{i=1}^{n} \Delta_i \Big[\log G' \Big\{ \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k \Big\} + \log \lambda_{k(i)} + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i \Big] - G \Big\{ \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k \Big\}, \tag{2.4}$$

where $\lambda_{k(i)}$ is the jump size of Λ at time Y_i, and G' is the first derivative of G.

Let $\widehat{\boldsymbol{\zeta}} \equiv (\widehat{\boldsymbol{\alpha}}, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_m)$ be the maximizer of (2.4) and $\boldsymbol{\zeta} \equiv (\boldsymbol{\alpha}, \lambda_1, \ldots, \lambda_m)$ be the corresponding generic vector of parameters. Also, let $\xi_i(\boldsymbol{\zeta}) = \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k$, $G_i'(\boldsymbol{\zeta}) = G'\{\xi_i(\boldsymbol{\zeta})\}$, $G_i''(\boldsymbol{\zeta}) = G''\{\xi_i(\boldsymbol{\zeta})\}$, and G'' denote the second derivative of G. The (scaled) score statistic for β evaluated at the NPMLE $\widehat{\boldsymbol{\zeta}}$ and $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ is

$$U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = n^{-1/2} \sum_{i=1}^{n} \Big\{ \Delta_i + \Delta_i \frac{G_i''(\widehat{\boldsymbol{\zeta}})}{G_i'(\widehat{\boldsymbol{\zeta}})} \xi_i(\widehat{\boldsymbol{\zeta}}) - G_i'(\widehat{\boldsymbol{\zeta}}) \xi_i(\widehat{\boldsymbol{\zeta}}) \Big\} \{ R_i S_i + (1 - R_i) \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \}.$$

Note that this statistic coincides with the score statistic derived based on the full likelihood, with the error term δ in the model of S following a mean-zero normal distribution.

To derive the null distribution of the score statistic, we first calculate the (asymptotic) variance of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ using a linear expansion around the "true" parameter values. To define the true values under a possibly misspecified model, let

$$f(y, \delta \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda, G) = \exp\big[- G\{\Lambda(y) e^{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \beta S}\} \big] \big[G'\{\Lambda(y) e^{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \beta S}\} \lambda(y) e^{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \beta S} \big]^{\delta}$$

for $y > 0$ and $\delta = 0, 1$, where $\lambda(t) = \mathrm{d}\Lambda(t)/\mathrm{d}t$. We define $\boldsymbol{\alpha}_0$, β_0 and Λ_0 to be the values that solve the following equations simultaneously:

$$\mathrm{E}\Big\{ \frac{\partial \log f(Y, \Delta \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda, G)}{\partial(\boldsymbol{\alpha}, \beta)} \Big\} = \mathbf{0}, \tag{2.5}$$

$$\mathrm{E}\Big[\frac{\partial \log f\{Y, \Delta \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda + \epsilon \int h(s)\, \mathrm{d}\Lambda(s), G\}}{\partial \epsilon} \Big|_{\epsilon = 0} \Big] = 0 \text{ for all } \|h\|_\infty \le 1. \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Clearly, if the transformation model is correctly specified, then $(\boldsymbol{\alpha}_0, \beta_0, \Lambda_0)$ are just the true parameter values.

Let $\boldsymbol{\zeta}_0 = (\boldsymbol{\alpha}_0, \lambda_{0,1}, \ldots, \lambda_{0,m})$, where $\lambda_{0,k} = \Lambda_0(t_k) - \Lambda_0(t_{k-1})$ for $k = 1, \ldots, m$ with $t_0 = 0$. For a given $\mathcal{K}$, define $\boldsymbol{\gamma}_{0\mathcal{K}} \equiv \arg\min_\gamma \mathrm{E}\{R(S - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\}$ as the true value of $\boldsymbol{\gamma}_{\mathcal{K}}$. Also, define $\psi_i(\boldsymbol{\zeta}) = G_i''(\boldsymbol{\zeta})/G_i'(\boldsymbol{\zeta})$. The Taylor series expansion of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $(\boldsymbol{\zeta}_0, \boldsymbol{\gamma}_{0\mathcal{K}})$ yields

$$U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = n^{-1/2} \sum_{i=1}^{n} \Big[\big\{ \Delta_i + \Delta_i \psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) \big\} \big\{ R_i S_i + (1 - R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \big\} - \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) - \widehat{\boldsymbol{I}}_{\beta\zeta}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\zeta\zeta}^{-1} \boldsymbol{U}_{\zeta,i} \Big] + o_p(1) \tag{2.7}$$

under some regularity conditions, where $\boldsymbol{U}_{\zeta,i} = (\boldsymbol{U}_{\alpha,i}^{\mathrm{T}}, U_{\lambda_1,i}, \ldots, U_{\lambda_m,i})^{\mathrm{T}}$,

$$\boldsymbol{U}_{\alpha,i} = \big\{ \Delta_i + \Delta_i \psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) \big\} \boldsymbol{X}_i,$$

and $U_{\lambda_k,i}$ $(k = 1, \ldots, m)$ is the derivative of the ith term of log-likelihood function with respect to λ_k:

$$U_{\lambda_k,i} = \frac{\Delta_{k(i)}}{\lambda_{0,k}} + I(Y_i \geq t_k) \big\{ \Delta_i \psi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0) \big\} \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i),$$

and $\widehat{\boldsymbol{I}}_{\zeta\zeta}$, $\widehat{\boldsymbol{I}}_{\beta\zeta}$, $\widehat{\boldsymbol{I}}_{\gamma\gamma}$ and $\widehat{\boldsymbol{I}}_{\beta\gamma}$ correspond to second derivatives of the nonparametric log-likelihood function; detailed formulations of the second derivative terms are given in the Appendix. The first term in the summation in (2.7) corresponds to the score statistic for β. The second term arises from expanding $U_\beta(\boldsymbol{\zeta}_0, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $\boldsymbol{\gamma}_{0\mathcal{K}}$ and expressing $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} - \boldsymbol{\gamma}_{0\mathcal{K}}$ as its linear approximation. Similarly, the third term results from expanding $U_\beta(\widehat{\boldsymbol{\zeta}}, \boldsymbol{\gamma}_{0\mathcal{K}})$ at $\boldsymbol{\zeta}_0$ and expressing $\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0$ as its linear approximation. These linear approximations are derived from viewing $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ and $\widehat{\boldsymbol{\zeta}}$ as solutions to estimating equations obtained from the derivatives of the squared error loss and the log-likelihood.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Based on this expansion, we can estimate the asymptotic variance of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by $\widehat{\sigma}(\mathcal{K})^2 = n^{-1}\sum_{i=1}^n \{\widehat{\sigma}_i(\mathcal{K}) - \overline{\sigma}(\mathcal{K})\}^2$, where

$$\widehat{\sigma}_i(\mathcal{K}) = \big\{\Delta_i + \Delta_i\psi_i(\widehat{\boldsymbol{\zeta}})\xi_i(\widehat{\boldsymbol{\zeta}}) - G_i'(\widehat{\boldsymbol{\zeta}})\xi_i(\widehat{\boldsymbol{\zeta}})\big\}\big\{R_iS_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}$$
$$- \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}) - \widehat{\boldsymbol{I}}_{\beta\zeta}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\zeta\zeta}^{-1}\widehat{\boldsymbol{U}}_{\zeta,i},$$

$\overline{\sigma}(\mathcal{K}) = n^{-1}\sum_{i=1}^n \widehat{\sigma}_i(\mathcal{K})$, and $\widehat{\boldsymbol{U}}_{\zeta,i}$ is $\boldsymbol{U}_{\zeta,i}$ with true parameter values replaced by estimators. Note that in the definition of $\widehat{\sigma}_i(\mathcal{K})$, the true parameter values in $\widehat{\boldsymbol{I}}_{\zeta\zeta}$, $\widehat{\boldsymbol{I}}_{\beta\zeta}$, $\widehat{\boldsymbol{I}}_{\gamma\gamma}$ and $\widehat{\boldsymbol{I}}_{\beta\gamma}$ are replaced by estimators. For an asymptotic size α test, we reject H_0 if $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})^2/\widehat{\sigma}(\mathcal{K})^2 \geq \chi^2_{1,\alpha}$. When H_0 is rejected, we suggest to use the sign of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as an estimate of the direction of the effect of S on the hazard of T.

Although the test is derived under the transformation model, it remains valid under the general hypothesis H_0 regardless of whether this model is correctly specified. We make two remarks about this robustness property. First, under H_0, the score statistic for β under the transformation model is mean zero at $(\boldsymbol{\alpha}, \Lambda) = (\boldsymbol{\alpha}_0, \Lambda_0)$ and $\beta = 0$, so the score test for H_0' is a test for H_0. Note that the contribution of a generic data point $(Y, \Delta, \boldsymbol{X}, S)$ to the log-likelihood is

$$\Delta \log\left[G'\{\Lambda(Y)e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}+\beta S}\}\lambda(Y)e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}+\beta S}\right] - G\{\Lambda(Y)e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}+\beta S}\} \equiv g(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S; Y, \Delta, \Lambda).$$

Under H_0, we have

$$\mathrm{E}\left\{\frac{\partial}{\partial\beta}g(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S; Y, \Delta, \Lambda)\Big|_{\beta=0}\right\}$$
$$= \mathrm{E}\left\{\frac{\partial}{\partial\mu}g(\mu; Y, \Delta, \Lambda)\Big|_{\mu=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}}\boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X}\right\} + \mathrm{E}\left\{\frac{\partial}{\partial\mu}g(\mu; Y, \Delta, \Lambda)\Big|_{\mu=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}}\right\}\mathrm{E}(\epsilon_S).$$

At $(\boldsymbol{\alpha}_0, \Lambda_0)$, the first term and the first expectation of the second term on the right-hand side above are 0. Therefore, under H_0, the score statistic of β is mean zero even if

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

the transformation model is misspecified. By contrast, when H_0 does not hold, ϵ_S and $\partial \log g/\partial \mu|_{\mu=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}}$ are generally correlated.

Second, the variance estimator $\widehat{\sigma}(\mathcal{K})^2$ is robust against misspecification of the transformation model, because it is based on a sum-of-squares expression, rather than the Hessian of the (nonparametric) log-likelihood. The sum-of-squares estimator is based on a linear expansion of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$, which does not rely on the correct specification of the transformation model. By contrast, the standard variance estimator based on the Hessian of the log-likelihood is generally inconsistent if the transformation model is misspecified.

2.2 Supremum test

The above score test is based on a single transformation function G, and the misspecification of G would result in power loss. To improve power, we propose a supremum test that combines the results from multiple choices of G. Suppose that we have q plausible choices of monotonically increasing transformation functions, denoted by $\{G^{(j)} : j = 1, \ldots, q\}$ with $G^{(j)}(0) = 0$. Let $\alpha^{(j)}$, $\beta^{(j)}$, and $\Lambda^{(j)}$ be the parameters under the jth transformation. For each j, we construct the proposed imputation score test statistic developed in Section 2.1. In particular, let $\widehat{\boldsymbol{\zeta}}^{(j)}$ denote the NPMLE under transformation function $G^{(j)}$ and $\beta^{(j)} = 0$. Let $U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ and $\widehat{\sigma}^{(j)}(\mathcal{K})$ denote the corresponding score statistic and estimated standard deviation, respectively. Define

$$Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \max_{1 \le j \le q} |Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})|, \tag{2.8}$$

where $Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})/\widehat{\sigma}^{(j)}(\mathcal{K})$. For simplicity of presentation, we write $Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ and $U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as $\widehat{Z}^{(j)}$ and $\widehat{U}_\beta^{(j)}$, respectively. The supremum test effec-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

tively takes the transformation model that yields the strongest evidence against the null hypothesis.

We approximate the distribution of $\left(\widehat{Z}^{(1)}, \ldots, \widehat{Z}^{(q)}\right)$ by a multivariate normal with mean $\mathbf{0}$ and variance $\widehat{\boldsymbol{V}}(\mathcal{K})$ and then approximate the distribution of $Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by the maximum of the absolute multivariate normal random vector. The variance matrix $\widehat{\boldsymbol{V}}(\mathcal{K})$ is a $(q \times q)$-matrix with diagonal elements 1 and the (j,k)th element being

$$\widehat{V}_{jk}(\mathcal{K}) = \frac{1}{n\widehat{\sigma}^{(j)}(\mathcal{K})\widehat{\sigma}^{(k)}(\mathcal{K})} \sum_{i=1}^{n} \left(\widehat{U}_{\beta,i}^{(j)} - \frac{1}{n}\sum_{i'=1}^{n} \widehat{U}_{\beta,i'}^{(j)}\right)\left(\widehat{U}_{\beta,i}^{(k)} - \frac{1}{n}\sum_{i'=1}^{n} \widehat{U}_{\beta,i'}^{(k)}\right), \quad j,k = 1, \ldots, q,$$

where $\widehat{U}_{\beta,i}^{(j)}$ is the ith term in the summation of $\widehat{U}_{\beta}^{(j)}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, q$.

To obtain an asymptotic size α test, we use the Monte-Carlo method to obtain the critical value of the test. The algorithm is as follows:

Algorithm 1 Supremum test

1: Generate M i.i.d. random samples $(z_m^{(1)}, \ldots, z_m^{(q)})$ $(m = 1, \ldots, M)$ from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\widehat{\boldsymbol{V}}(\mathcal{K})$;

2: Compute the test statistic T_m from the mth sample: $T_m = \max_{1 \le j \le q} |z_m^{(j)}|$;

3: Reject H_0 if $Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ is larger than the $(1-\alpha)$th quantile of $(T_1, \ldots, T_M)$.

3. Theoretical Properties

In this section, we present the asymptotic joint distribution of the score statistics $(\widehat{U}_{\beta}^{(1)}, \ldots, \widehat{U}_{\beta}^{(q)})$, from which we can obtain the validity of the proposed (supremum) score test. In general, we use the superscript (j) to denote a statistic or quantity evaluated under the jth transformation model. Let $\widehat{\boldsymbol{U}}_{\beta}(\mathcal{K}) = \left(\widehat{U}_{\beta}^{(1)}, \ldots, \widehat{U}_{\beta}^{(q)}\right)^{\mathrm{T}}$, where we suppressed the dependence on $\mathcal{K}$ on the right-hand side.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

We prove that under some regularity conditions, $\boldsymbol{\Sigma}(\mathcal{K})^{-1/2}\widehat{\boldsymbol{U}}_\beta(\mathcal{K})$ converges to a multivariate normal distribution under H_0 even when $\mathcal{K}$ is chosen randomly, for some $\boldsymbol{\Sigma}(\mathcal{K})$ defined in the proof of Theorem 1. To precisely state the theoretical result, let $\mathcal{K}^*$ be a general model selection operator, such that for an m-vector of outcome variables $\mathcal{Y}$ and an $(m \times p)$-matrix of covariates $\mathcal{Z}$, $\mathcal{K}^*(\mathcal{Y}, \mathcal{Z}) : \mathbb{R}^m \times \mathbb{R}^{m\times p} \to \mathcal{C}_p$, where $\mathcal{C}_p$ is the collection of subsets of $\{1, \ldots, p\}$. We assume that the model for S is selected based on the residual $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{A}$, where $\widehat{\boldsymbol{\gamma}}_X \equiv (\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{X}_i S_i$ is the least-squares estimator of S on $\boldsymbol{X}$ using the subjects with $R = 1$. The selected components of $\boldsymbol{A}$ are $\mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$, where $\mathcal{S}$ is a vector that consists of $\{S_i : R_i = 1\}$, and $\boldsymbol{X}$ and $\boldsymbol{A}$ are matrices that consist of rows of $\{\boldsymbol{X}_i : R_i = 1\}$ and $\{\boldsymbol{A}_i : R_i = 1\}$, respectively. For simplicity of presentation, we write $\mathcal{K}^* = \mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$. Therefore, $\mathcal{K}$ can be viewed as the observed value of $\mathcal{K}^*$.

Let $\|\cdot\|_{\psi_\xi}$ be an Orlicz norm, such that $\|X\|_{\psi_\xi} = \inf\{\eta > 0 : \mathrm{E}\{\exp(|X|^\xi/\eta^\xi)\} \le 2\}$, and $\|\cdot\|$ be the Euclidean norm. We establish the asymptotic property of $\widehat{\boldsymbol{U}}_\beta(\mathcal{K}^*)$ under the following conditions. Some conditions involve a generic positive constant M.

(C1) For some $\xi \in (0, 2]$, $\|S\|_{\psi_\xi} + \max_j \|A_j\|_{\psi_\xi} < M$. The covariate $\boldsymbol{X}$ is bounded, so that $P(\|\boldsymbol{X}\| < M) = 1$.

(C2) There exists a sequence of collections of models Ω_n, such that $P(\mathcal{K}^* \in \Omega_n) \to 1$, $\sup_{\mathcal{K}\in\Omega_n} |\mathcal{K}| = O(n^\nu)$, and $\log|\Omega_n| = O(n^\kappa)$, where ν and κ are constants that satisfy $\nu < 4\xi/(5\xi + 12)$, $5\nu/4 + 3\kappa/\xi < 1$, and $4\nu/3 + 8\kappa/(3\xi) < 1$, and $|\mathcal{C}|$ denotes the cardinality of the set $\mathcal{C}$. Also, $\inf_{\mathcal{K}\in\Omega_n} \lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_\mathcal{K}\boldsymbol{W}_\mathcal{K}^{\mathrm{T}})\} > M^{-1}$, $\sup_{\mathcal{K}\in\Omega_n} \mathrm{E}\{(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_\mathcal{K})^4\} < M$, where $\lambda_{\min}(\boldsymbol{C})$ denotes the minimum eigenvalue of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

the matrix $\boldsymbol{C}$. In addition, $\inf_{\mathcal{K}\in\Omega_n} \lambda_{\min}\{\boldsymbol{\Sigma}(\mathcal{K})\} > M^{-1}$.

(C3) The probability $P(R = 1 \mid Y, \Delta, \boldsymbol{X}) > M^{-1}$ almost surely.

(C4) Under H_0, $\boldsymbol{A}$ is independent of $(Y, \Delta, \boldsymbol{X})$.

(C5) The models selected based on the estimated residuals $(S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ and the actual residuals $(S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ are such that

$$P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) \neq \mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})\Big\} = o(1)$$

and

$$\sup_{\mathcal{K}\in\Omega_n} \frac{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) = \mathcal{K}\Big\}}{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A}) = \mathcal{K}\Big\}} < M,$$

where $\boldsymbol{\gamma}_{0X}$ is the true value of $\boldsymbol{\gamma}_X$.

(C6) For a random sample of size m, let $\widetilde{\mathcal{S}} = (S_1, \ldots, S_m)^{\mathrm{T}}$, $\widetilde{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^{\mathrm{T}}$, and $\widetilde{\mathcal{A}} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)^{\mathrm{T}}$. The random variable

$$\sup_{\mathcal{K}\in\Omega_m} \left| \frac{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K} \mid \widetilde{\mathcal{A}}\}}{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K}\}} - 1 \right|$$

converges to 0 in mean as $m \to \infty$.

(C7) For $j = 1, \ldots, q$, the transformation function $G^{(j)}$ is continuously differentiable up to the fourth order. Also, equations (2.5) and (2.6), with $G = G^{(j)}$ for $j = 1, \ldots, q$, have unique solutions $(\boldsymbol{\alpha}_0^{(j)}, \beta_0^{(j)}, \Lambda_0^{(j)})$, where $\boldsymbol{\alpha}_0^{(j)}$ lies in the interior of a known compact set, and $\Lambda_0^{(j)}$ is continuously differentiable with positive derivative in $[0, \tau]$. In addition, with probability one, $P(C \geq \tau \mid \boldsymbol{X}, S) = P(C = \tau \mid \boldsymbol{X}, S) > M^{-1}$ and $P(T \geq \tau \mid \boldsymbol{X}, S) > M^{-1}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

(C8) The operator $(\boldsymbol{W}_{\alpha}^{(j)}, W_{\Lambda}^{(j)})$ defined in the Supplementary Material is continuously invertible for $j = 1, \ldots, q$.

Remark 1. Conditions (C1)–(C6) are adopted from Wong and Feng (2023). Condition (C4) can be similarly relaxed to Condition (C4') of Wong and Feng (2023). Essentially, Condition (C5) requires that the models selected based on the true and estimated residuals of S on $\boldsymbol{X}$ are (asymptotically) the same, and Condition (C6) requires that the model selection probabilities are the same whether conditional on the covariates or not. See the Supplementary Material of Wong and Feng (2023) for the verification of Conditions (C5) and (C6) under marginal screening. Condition (C7) is standard for semiparametric transformation models. In particular, it requires an end-of-study time τ, such that no events beyond τ could be observed, and there exists a nonvanishing proportion of subjects whose censoring time is equal to τ. This ensures that the NPMLE of Λ is uniformly consistent over the entire closed interval $[0, \tau]$. Condition (C8) essentially consists of assumptions for the Z-estimator master theorem (Theorem 3.3.1 of van der Vaart and Wellner (1996)), which guarantees the asymptotic normality of $(\widehat{\boldsymbol{\alpha}}^{(j)}, \widehat{\Lambda}^{(j)})$ for $j = 1, \ldots, q$. In this paper, we directly assume the required conditions instead of proving them based on properties of the true model, because we do not assume the form of the true model. By contrast, if we specify the true model, then we may establish Condition (C8) based on properties of the true model along the lines of, for example, Zeng, Lin and Lin (2008).

We have the following results.

Theorem 1. *Under Conditions (C1)–(C8) and* H_0, $\boldsymbol{\Sigma}(\mathcal{K}^*)^{-1/2}\widehat{\boldsymbol{U}}_{\beta}(\mathcal{K}^*)$ *converges weakly to the standard multivariate normal distribution.*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Note that $\mathcal{K}^*$ is a random object, reflecting the data-dependent nature of the selected model of S. In the special case where $\mathcal{K}^*$ is prespecified, the theorem simplifies to the standard result of asymptotic normality of the score statistic.

Theorem 2. *Under Conditions (C1)–(C8) and H_0,*

$$\sup_{\mathcal{K}\in\Omega_n} \|\widehat{\boldsymbol{\Sigma}}(\mathcal{K}) - \boldsymbol{\Sigma}(\mathcal{K})\|_F \xrightarrow{p} 0,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The proofs of Theorems 1 and 2 are given in the Supplementary Material. From these two theorems, we have the following result that implies the validity of the proposed test.

Theorem 3. *Under Conditions (C1)–(C8) and H_0, for any $t \in \mathbb{R}$,*

$$P\Big\{Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*}) < t\Big\} - P\Big\{\|\widehat{\boldsymbol{V}}(\mathcal{K}^*)^{1/2}\boldsymbol{Z}\|_\infty < t\Big\} \to 0,$$

where $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_q)$ and is independent of the observed data.

4. Simulation studies

4.1 Study 1 — Single model tests

Let $\boldsymbol{X} = (X_1, \ldots, X_5)^{\mathrm{T}}$, where (X_1, X_2, X_3) are mean-zero normal variables with $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ $(j, k = 1, 2, 3)$, $X_4 \sim$ Bernoulli(0.25), $X_5 \sim$ Bernoulli(0.35), and X_4 and X_5 are independent of each other and of (X_1, X_2, X_3). Let $\boldsymbol{A}$ be a p-vector of independent standard normal random variables. We set $S = \boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}}\boldsymbol{A} + \boldsymbol{\gamma}_{A,2}^{\mathrm{T}}\boldsymbol{A}^2 + \delta$, where $\boldsymbol{A}^2$ is a p-vector of the squared components of $\boldsymbol{A}$, δ is standard normal, and $\boldsymbol{\gamma}_X = (0.1, \ldots, 0.1)^{\mathrm{T}}$. We set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components and 0 elsewhere, and set $\boldsymbol{\gamma}_{A,2}$ to be 0.1 at the first 5 components and 0 elsewhere.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

We considered three failure time models:

Model 1: $\Lambda(t \mid \boldsymbol{X}, S) = \Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)$;

Model 2: $\Lambda(t \mid \boldsymbol{X}, S) = \log\{1 + \Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}$;

Model 3: $T = \exp(-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} - \beta S) + \epsilon$, where $\epsilon \sim \mathrm{Exp}(1)$.

In all models, we set $\boldsymbol{\alpha} = (0.2, -0.2, 0.2, -0.2, 0.2)^{\mathrm{T}}$ and $\Lambda(t) = 0.01t$. Model 1 and Model 2 are the PH and PO models, respectively, whereas Model 3 does not take the form of a transformation model. We generated the censoring time C from an exponential distribution with mean $\lambda_0 + \lambda_1 X_4$, where λ_0 and λ_1 were chosen to yield a censoring rate of approximately 50–60%, and the means of the exponential distributions differ by 25% or more. We considered two missing-data mechanisms. The first mechanism is MCAR. The second mechanism is MAR, where we first randomly selected 40% of the subjects with $X_5 = 1$ and 10% of the subjects with $X_5 = 0$ into a subcohort, and the selected subjects would have observed S. For subjects outside the subcohort, we selected a fraction of subjects with censored event time to have missing S to attain the desired missing proportion. If the missing proportion was not attained by setting all censored subjects to have missing S, then a subset of subjects with observed event time would also be selected. We set the number of Monte-Carlo replicates for approximating the null distribution of the test statistics to be $M = 500{,}000$. We considered sample sizes of $n = 500$ and 1000, and numbers of auxiliary variables of $p = 200$, 500, 1000 and 1500. For the alternative hypothesis, we set $\beta = 3n^{-1/2}$ for Models 1 and 2, and $\beta = 1.5n^{-1/2}$ for Model 3. For each setting, we simulated 50,000 and 10,000 replicates for $\beta = 0$ and $\beta \neq 0$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

In this subsection, we consider the performance of the proposed test under a given transformation model and compare it with existing methods. We compare the performance of six tests: (1) the score test using complete data only; (2) the score test with missing values imputed under a working linear model of S on $\boldsymbol{X}$ and components of $\boldsymbol{A}$ selected using marginal screening, where a component of $\boldsymbol{A}$ is selected if its absolute empirical correlation with $S - \widehat{\gamma}_X^{\mathrm{T}} \boldsymbol{X}$ among the subjects with complete data is larger than a certain threshold; (3) the score test based on the full likelihood with a working linear model of S against $\boldsymbol{X}$ only, which is similar to the method proposed by Lawless (2018); (4) the score test based on the full likelihood with the same model of S as (2); (5) the proposed test, where the working model of S is selected in the same way as (2); and (6) the score test based on the full likelihood with a linear model of S against $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ that are associated with S. We refer to methods (1)–(6) as the complete-case analysis, the simple imputation method, the covariate-only method, the full likelihood method, the proposed method, and the true model method. For methods (2), (4), and (5), the threshold for screening is selected using BIC. For the true model method, the variance of the score statistic is estimated using the proposed empirical variance estimator. For all methods, we fit the correct failure time model under Models 1 and 2, and under Model 3, we fit both the PH and PO models.

The results under a missing proportion of 60% are plotted in Figures 1 and 2, and the results under a missing proportion of 30% are presented in Figures S1 and S2 in the Supplementary Material; for methods that inflate the type I error, their performance under the alternative hypothesis is not presented. The significance level is set to be 0.05. Under Models 1 and 2 with sample size 1000, all methods appear to preserve the type I

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

error. Theoretically, the complete-case analysis and the simple imputation method would inflate the type I error under MAR, but the empirical results do not clearly exhibit such a pattern under the current setting. Under Model 3, the complete-case analysis, the simple imputation method, the covariate-only analysis, and the full likelihood generally inflate the type I error, because these methods estimate the variance based on the Hessian of the log-likelihood, which is misspecified in this setting. As expected, the proposed method utilizes information about missing data contained in the auxiliary variables and tends to yield higher power than the complete-case analysis and covariate-only method. Under Models 1 and 2, the full likelihood method tends to be slightly more powerful than the proposed method, but the corresponding type I errors for the full likelihood method are also slightly inflated. In fact, both methods have the same numerator in the score statistic, so the difference is due to small-sample differences in the variance estimators. We also evaluate the sign of the score statistic $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ when H_0 is rejected. In all models under the alternative hypotheses, whenever H_0 is rejected, $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ is of the same sign as the true value of β.

4.2 Study 2 — Supremum tests

Having established the validity of the proposed method under a prespecified transformation model, in this subsection, we consider the supremum test under the same Models 1–3 and covariate distributions. The supremum test is performed with 6 working outcome models ($q = 6$), including transformation models with the Box–Cox transformation (2.2) at $\rho = 0$, 0.5, 1, and 1.5 and the logarithm transformation (2.3) at $r = 0$ and 1.5. Note that $\rho = 0$ and $\rho = 1$ correspond to the PO and PH models, respectively. For comparison,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

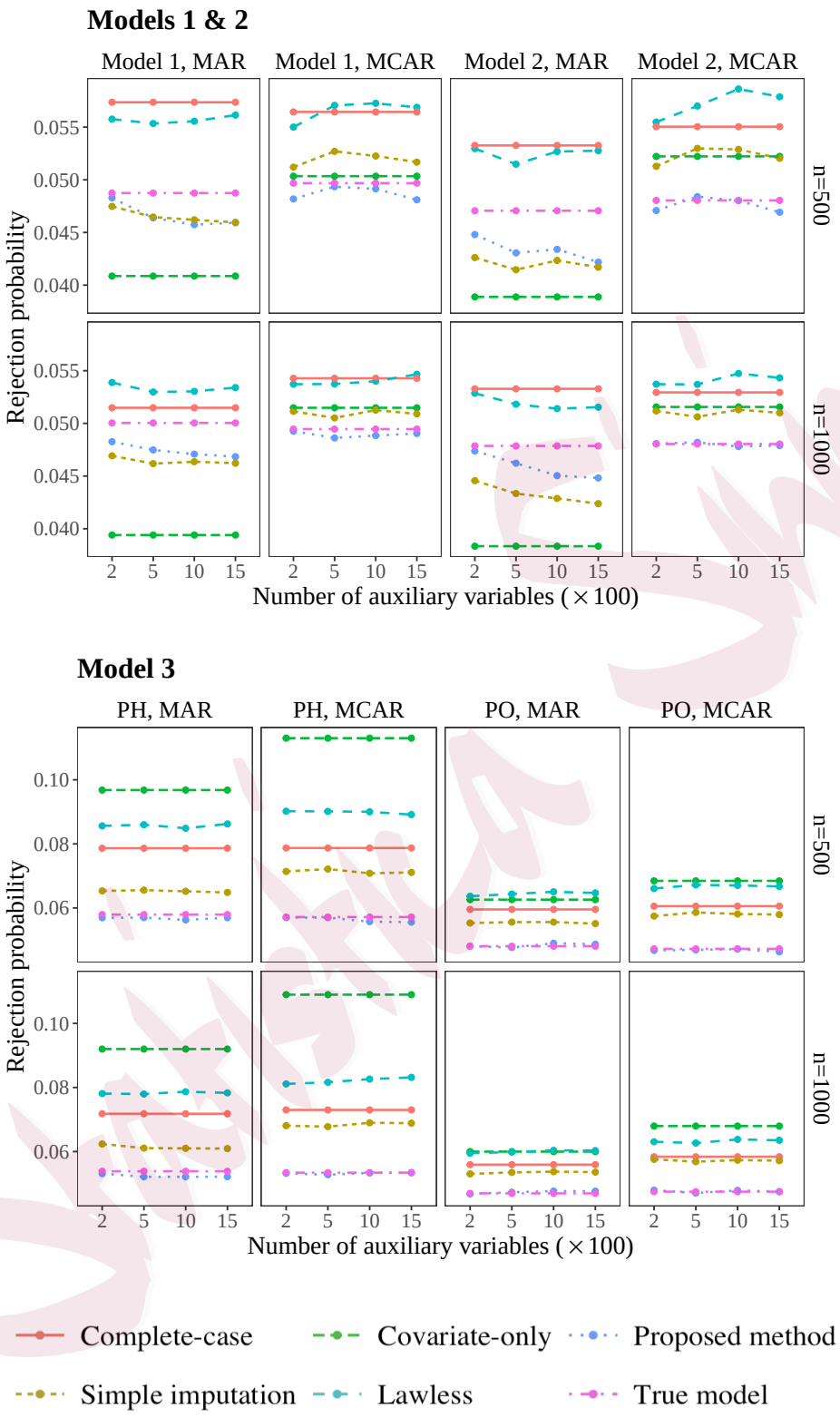

Figure 1: Study 1 — Rejection probabilities under a missing proportion of 60% and the null hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

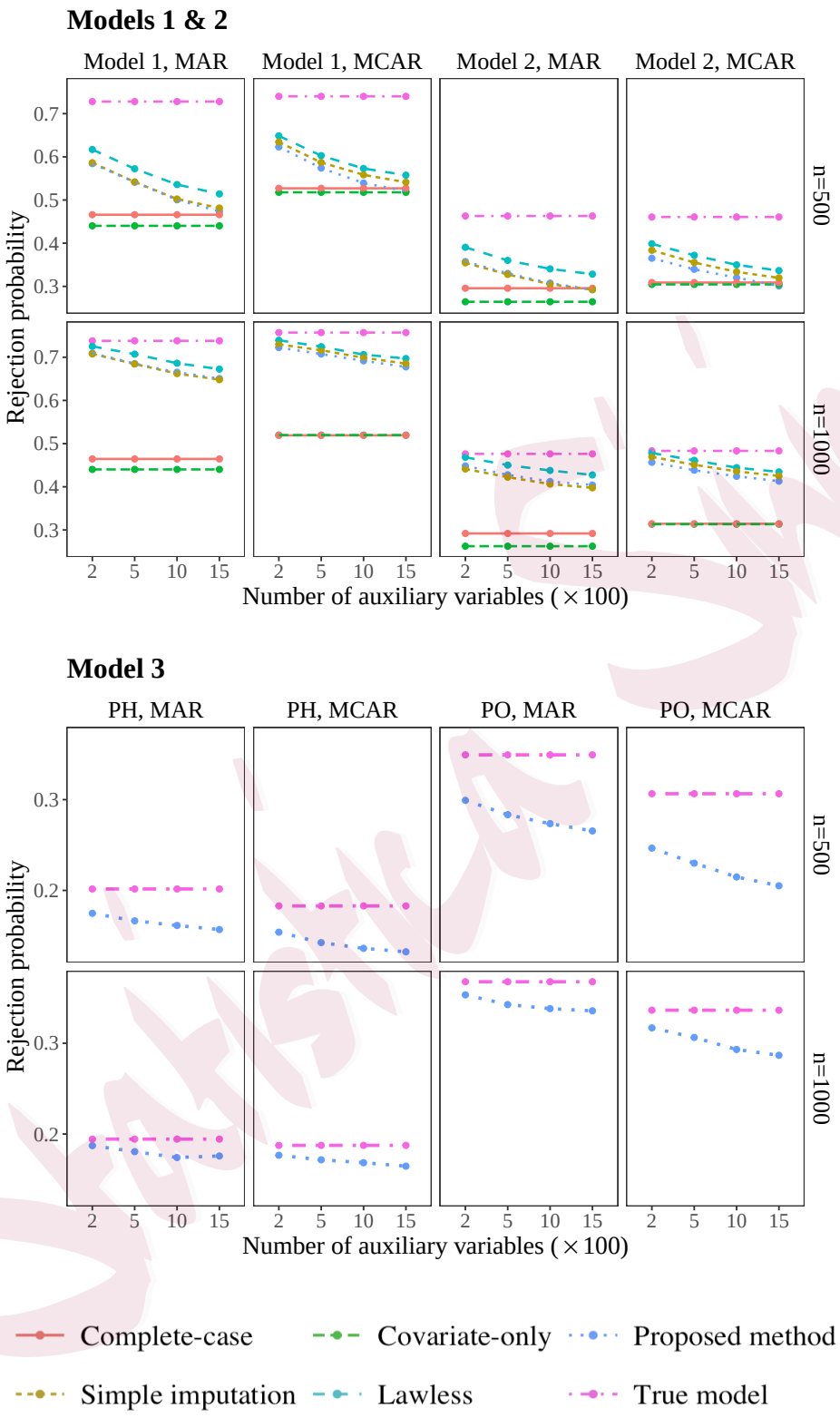

Figure 2: Study 1 — Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

we also present the results of the proposed single-model score test with the PH and PO models. The results under a missing proportion of 60% are plotted in Figures 3 and 4, and the results under a missing proportion of 30% are presented Figures S3 and S4 in the Supplementary Material. All three tests preserve the type I error under all three models. Under Models 1 and 2, the supremum test does not lose much power compared with the single-model test with the correct model specification. Under Model 3, the PO model is the most powerful, as it tends to yield a large value of the (standardized) score statistic. The supremum test, while having a test statistic value larger than or equal to that under the PO model, has a larger rejection threshold. Therefore, its power is slightly lower than the PO model. The PH model has the smallest test statistic value and thus is the least powerful. This illustrates that even when the outcome model is unknown or misspecified, we can perform the supremum test to achieve a relatively high power. In general, we recommend considering a few interpretable working models. It is difficult to derive a general rule on what or how many models to consider, but based on some additional simulation studies, the empirical results under the current simulation setting are similar for q varying from 2 to 6.

4.3 Study 3 — Single model tests with strong and weak signals

In this subsection, we consider a case with a larger degree of variability in the model selection step. In particular, we adopt the setting in Section 4.1 but generated S with a mixture of strong and weak signals of the auxiliary variables. Specifically, we set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components, 0.02 at the subsequent 80 components, and 0 at the remaining components. The results are presented in Figures S5 – S8 in the Supplementary

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

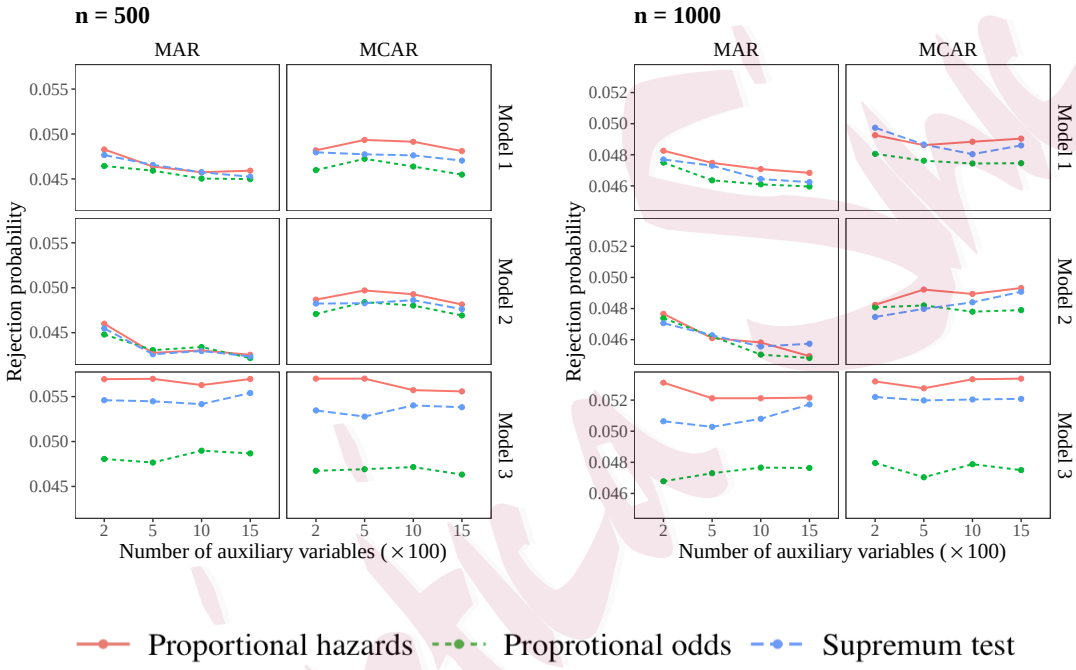

Figure 3: Study 2 — Rejection probabilities under a missing proportion of 60% and the null hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

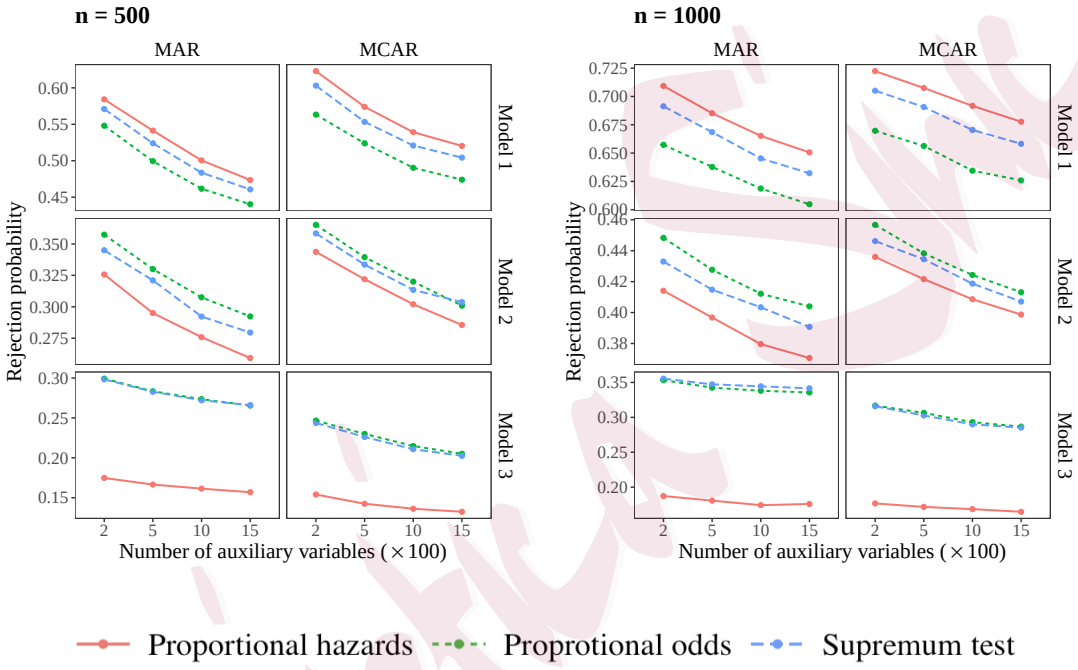

Figure 4: Study 2 — Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Material. Similar to the results of Study 1, the proposed method yields satisfactory performance under this setting. Under the alternative hypothesis, the true model method tends to have high power, but the proposed method is more powerful than the true model method in some scenarios. This is because the true model contains many auxiliary variables with weak signals, and the extra information contained in the variables does not compensate for the variability involved in the estimation of their effects, so it is beneficial to select a subset of variables with stronger signals. Therefore, to select the model of S, we could perform feature screening with a relatively large selection threshold or use penalization methods, such as lasso, with a large penalty. These tuning parameters could be selected based on conservative information criteria or methods with false discovery rate control.

4.4 Study 4 — Supremum tests with discrete S

In this subsection, we further evaluate the robustness of the proposed methods under a misspecified distribution of S. Specifically, we adopt the setting in Section 4.1 but generated S from a binomial distribution with the number of trials equals 2 and success probability

$$p = \frac{\exp(\boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}}\boldsymbol{A} + \boldsymbol{\gamma}_{A,2}^{\mathrm{T}}\boldsymbol{A}^2)}{1 + \exp(\boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}}\boldsymbol{A} + \boldsymbol{\gamma}_{A,2}^{\mathrm{T}}\boldsymbol{A}^2)}.$$

Such a covariate distribution would arise when S is a genotype. The values of $\boldsymbol{\gamma}_X$, $\boldsymbol{\gamma}_A$ and $\boldsymbol{\gamma}_{A,2}$ are equal to those in Section 4.1. Note that in this setting, S does not depend linearly on $\boldsymbol{X}$. The null hypotheses being tested are $\beta = 0$ for the single-model test and $\beta^{(j)} = 0$ for $j = 1, \ldots, q$ for the supremum test. The results are presented in Figures S9–S12. Similar to the results in Section 4.2, the proposed method preserves the type I

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

error across all three models. Under Models 1 and 2, the supremum test demonstrated power comparable to the single-model test under the correct model. Under Model 3, the power of the supremum test is substantially larger than that of the single-model test with PH and is comparable to that of the single-model test with PO.

5. Real Data Analysis

5.1 TCGA: Bladder Urothelial Carcinoma

We analyze a dataset of patients with bladder urothelial carcinoma (BLCA) from TCGA (The Cancer Genome Atlas Network, 2014). In the study, most subjects had available clinical variables, including sex, age at diagnosis, time to tumor progression, and time to death since the initial diagnosis. The expressions of 18224 genes, generated by RNA sequencing, were measured for most subjects. The expressions of 208 proteins or phospho-proteins are available for 82% of the subjects. After removing subjects with missing clinical data, the sample size is 348. The median follow-up time was about 1.3 years, and about 49% of the patients were lost to follow-up before tumor progression or death.

We aim to identify protein expressions that are associated with the time to tumor progression or death, whichever occurs first. The covariates in $\boldsymbol{X}$ include age at diagnosis, sex and stage N. In the sample, 26.44% of patients are female. Stage N is classified into N0 (64.08%), N1(12.93%), N2(21.26%) and N3(1.72%) and is represented by a single variable with values 0, 1, 2, and 3, respectively. In a single analysis, we set the covariate of interest S to be the expression of a protein or phospho-protein. We set the gene expressions as auxiliary variables. About 6% of the gene expression values are missing, and we impute

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Table 1: p-values and references of significant proteins in the TCGA BLCA analysis.

Protein expression	**Proposed method**	**Complete-case**		**Covariate-only**		**Reference**
		PH	PO	PH	PO	
GATA3	3.40E−05	1.12E−04	5.00E−05	1.04E−04	4.40E−05	Higgins et al. (2007)
Src	1.48E−04	8.20E−04	4.86E−04	8.33E−04	4.62E−04	Xu et al. (2021)
TAZ	1.80E−04	1.38E−03	5.51E−04	1.09E−03	4.32E−04	Gao et al. (2014)

them using k-nearest neighbor imputation with $k = 10$.

We perform the supremum test with $q = 2$ and the two transformation functions corresponding to the PH and PO models. The working model of S is selected in two steps: first, select 1000 gene expressions by the correlation-based marginal screening procedure, and then perform lasso on the selected gene expressions; the tuning parameter in lasso is selected by BIC. For comparison, we also perform the complete-case analysis and the covariate-only method described in the simulation studies under the PH and PO models.

Under a (family-wise) significance level of 0.05 and the Bonferroni correction, i.e., an individual significance level of $0.05/208 = 0.00024$, three proteins are identified to be significantly associated with progression-free survival time under at least one of the five tests. All three protein expressions are more significant under the proposed method than under other methods with either outcome model. Also, the three proteins have been identified to be related to the progression of BLCA in previous studies. The p-values under all methods of the significant protein expressions and some relevant references are given in Table 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

5.2 METABRIC

We also apply the proposed method to analyze data from the Molecular Taxonomy Of Breast Cancer International Consortium (METABRIC) study (Curtis et al., 2012) to investigate the association between gene expressions and the relapse-free survival time of breast cancer patients. The data are available through the cBioPortal for Cancer Genomics (https://www.cbioportal.org/study/summary?id=brca_metabric). The study contains data of clinical variables, gene expressions and copy number alterations (CNAs). For the analysis, we select patients with subtypes Luminal A and Luminal B as study subjects. Also, we select 1500 genes with the largest variances as the study variables. After removing subjects with missing clinical data, the sample size is 1119. The median follow-up time was about 119 months, and 35% of the patients were lost to follow-up before tumor progression or death. We artificially introduce 50% of missingness with the MAR mechanism described in the simulation studies for the gene expressions to demonstrate the proposed method.

The covariates in $\boldsymbol{X}$ include age at diagnosis, Her2 status, indicator of chemotherapy, indicator of hormone therapy, and indicator of radiotherapy. Her2 status is classified into loss (6.08%), neutral (77.57%) and gain (16.35%) and is represented by a single variable with values 0, 1 and 2, respectively. In a single analysis, we set the covariate of interest S to be a single gene expression. We set the CNAs as auxiliary variables. For each CNA, if there exists another CNA such that they have more than 95% same values, then we delete it from the analysis. After deletion, the dimension of CNA is 385.

We perform the supremum test with $q = 2$ and the two transformation functions

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

corresponding to the PH and PO models. The auxiliary variables of CNA are selected by lasso, and the tuning parameter of lasso is selected using BIC. For comparison, we include the results under the complete-case analysis and the covariate-only method described in the simulation studies with the PH and PO models. Also, we perform score tests using all available gene expressions under the PH and PO models, and we refer to it as the complete-data analysis. The results of the complete-data analysis can be viewed as the gold standard since it uses all values of S.

There are seven gene expressions identified to be significantly associated with progression-free survival time at the (Bonferroni-corrected) significance level of $0.05/1500 = 3.33 \times 10^{-5}$ under the complete-data analysis with either outcome model. Among these gene expressions, all of them are most significant under the complete-data analysis with the PO model, and 5 are more significant under the proposed method than under the complete-case analysis and the covariate-only method with either outcome model. This suggests that the proposed method is more powerful than the other two methods. The p-values under all methods of the significant gene expressions are given in Table 2. As the genes tend to show higher significance under the PO model, the results highlight the advantage of considering multiple outcome models. If we relied solely on the "default choice" of the Cox model, then we might have missed some identified associations.

6. Discussion

In this paper, we develop a score test for the presence of association between a potentially right-censored survival outcome and an incomplete covariate, where the missing values of the incomplete covariate can be imputed using high-dimensional auxiliary variables.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Table 2: p-values of significant gene expressions in the METABRIC data analysis.

Gene expression	**Proposed method**	**Complete-data**		**Complete-case**		**Covariate-only**	
		PH	PO	PH	PO	PH	PO
CDCA5	2.92E−03	7.60E−06	7.57E−08	3.30E−02	1.24E−02	1.51E−02	1.37E−02
FAM164A	4.44E−03	2.30E−05	2.16E−05	1.44E−02	3.68E−02	3.17E−02	3.74E−02
S100P	4.39E−03	4.47E−05	3.87E−06	3.15E−03	6.02E−03	9.66E−03	9.31E−03
NFKBIZ	6.50E−03	2.73E−04	1.99E−05	1.35E−02	7.01E−03	3.08E−02	9.39E−03
PTTG1	4.08E−03	7.41E−05	2.22E−05	6.85E−02	1.73E−02	6.33E−02	2.06E−02
CCNB2	9.80E−05	1.50E−04	9.71E−06	8.40E−04	1.38E−04	1.07E−03	2.06E−04
AURKA	1.24E−02	7.94E−04	2.04E−05	1.07E−01	3.56E−03	5.85E−02	1.10E−02

We propose to select a subset of auxiliary variables before performing the score test. We consider a flexible transformation model for the survival outcome and propose a supremum test that combines multiple outcome models to improve power. Our theoretical development requires only that S is linearly associated with covariates $\boldsymbol{X}$, and the validity of the score test does not depend on the correctness of the working model.

In the simulation studies, we demonstrate that under alternative hypotheses, the sign of the score statistic consistently aligns with the sign of β whenever the null hypothesis is rejected. However, this result is merely empirical, and a theoretical guarantee has yet to be established. The theoretical challenge lies in the fact that under (contiguous) alternatives, the least-squares estimator $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} \equiv \mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})^{-1}\mathrm{E}(RS\boldsymbol{W}_{\mathcal{K}})$ converges to a limit that depends on β through the distribution of R, which is unspecified. It is unclear what the contribution of $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ to the asymptotic mean of the score statistic $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ is, except in the simple case of MCAR, where the limit of $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ does not depend on β.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

We assume that R is independent of $(S, \boldsymbol{A})$ given $(Y, \Delta, \boldsymbol{X})$, which is more restrictive than the conventional MAR assumption that allows R to be independent of S given $(Y, \Delta, \boldsymbol{X}, \boldsymbol{A})$. If some components of $\boldsymbol{A}$ are not selected into the working model of S on $\boldsymbol{W}_{\mathcal{K}}$, and R depends on these unselected components of $\boldsymbol{A}$, then S would be missing not at random under the working model. To allow for the dependence of R on $\boldsymbol{A}$, we would need to identify the components of $\boldsymbol{A}$ that are associated with R and account for them when fitting the working model of S.

Under the working model of S, we estimate $\boldsymbol{\gamma}_{\mathcal{K}}$ using the complete cases only. Under H_0 (and contiguous alternatives), inverse probability weighting is possible but not required, because estimation of $\boldsymbol{\gamma}_{\mathcal{K}}$ is consistent regardless of whether weighting is used. In fact, weighting may create a superpopulation under which the least-squares estimator of $\boldsymbol{\gamma}_{\mathcal{K}}$ is less efficient than the unweighted estimator. In addition, inclusion of weights would affect the expansion of the score statistic (2.7), and the proof for Theorem 1 may need to be substantially modified.

We focus on hypothesis testing rather than regression analysis in general. Although association testing typically can be done under a regression framework, there are issues of interest in hypothesis testing that are not pertinent under a general regression analysis framework. First, in estimation, strong assumptions concerning the distribution of the incomplete variables are usually made to ensure desirable theoretical properties, such as consistency of the estimators. By contrast, in hypothesis testing, we are primarily concerned with the theoretical properties of estimators under the null hypothesis (or contiguous alternatives), and correct specification of the full model is not required for a test to be valid. As a result, much more relaxed model assumptions could be considered for

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

association tests than for regression analyses in general. Second, estimation is generally performed under a specific model, whereas in hypothesis testing, we may only be interested in the existence of association between a covariate and an outcome variable, not necessarily under particular models. The proposed methods make use of the flexibility of the hypothesis testing problem and are not simple by-products of an inferential procedure under a general regression setting.

In the proposed method, the model of Y can be misspecified without compromising the validity of the test. The proposed score test preserves the type I error with or without correct specification of the outcome model, since the variance of score statistic is derived using the sum of squares of the individual score statistics instead of the Hessian of the log-likelihood. As expected, when the outcome model is misspecified, the power is adversely affected. Nevertheless, the loss in power is relatively small based on our simulation studies.

Our work can be extended to allow for different types of outcome variables. First, the survival data considered in this paper is right-censored. It is of interest to consider other types of censoring, such as interval censoring, where the event of interest is known only to occur within a time interval. For example, in HIV/AIDS studies, blood samples are taken from study subjects periodically to look for evidence of HIV seroconversion. Then one subject's event time is only known to fall between two blood drawings. Second, in the current study, the time-to-event outcome is univariate. In genomic studies, we may encounter multivariate survival data, where each subject may experience more than one event. In this case, the interested events may be correlated with each other. We may consider modeling a joint survival function and performing a score test for multiple parameters.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Supplementary Material

The online Supplementary Material provides proofs of technical results and additional simulation results.

Acknowledgments

This research was partially supported by the Guangdong Basic and Applied Basic Research Foundation (Project No. 2021A1515110048) and the Hong Kong Research Grants Council (Grant No. 15303422).

Appendix: Details of the variance estimator of the score statistic

In this Appendix, we give the expressions in the Taylor series expansion of the score statistic presented in Section 2.1, which are used in the variance estimator of the score. Define $\eta_i(\boldsymbol{\zeta}) = G_i'''(\boldsymbol{\zeta})/G_i'(\boldsymbol{\zeta}) - \{G_i''(\boldsymbol{\zeta})/G_i'(\boldsymbol{\zeta})\}^2$, with $G_i'''(\boldsymbol{\zeta}) = G'''\{\xi_i(\boldsymbol{\zeta})\}$ and G''' being the third derivative of G. Let

$$\widehat{\boldsymbol{I}}_{\alpha\alpha} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\right\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}},$$

and $\widehat{\boldsymbol{I}}_{\alpha\lambda}$ be a $(\|\boldsymbol{\alpha}\|_0 \times m)$-matrix with the kth column being

$$(\widehat{\boldsymbol{I}}_{\alpha\lambda})_k = -\frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t_k)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) + \Delta_i\psi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\big\}\exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\boldsymbol{X}_i.$$

Let $\widehat{\boldsymbol{I}}_{\lambda\lambda}$ be an $(m \times m)$-matrix with the (j,k)th element being

$$(\widehat{\boldsymbol{I}}_{\lambda\lambda})_{jk} = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\Big[-\frac{I(Y_i=t_j)}{\lambda_j^2} + I(Y_i \geq t_j)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\big\}\exp(2\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\Big] & \text{if } k = j, \\ -\frac{1}{n}\sum_{i=1}^{n} I\{Y_i \geq \max(t_k, t_j)\}\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\big\}\exp(2\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i) & \text{if } k \neq j. \end{cases}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

The matrix $\widehat{\boldsymbol{I}}_{\zeta\zeta}$ is defined as

$$\widehat{\boldsymbol{I}}_{\zeta\zeta} = \begin{pmatrix} \widehat{\boldsymbol{I}}_{\alpha\alpha} & \widehat{\boldsymbol{I}}_{\alpha\lambda} \\ \widehat{\boldsymbol{I}}_{\alpha\lambda}^{\mathrm{T}} & \widehat{\boldsymbol{I}}_{\lambda\lambda} \end{pmatrix}.$$

The vector $\widehat{\boldsymbol{I}}_{\beta\zeta}$ is defined as $\widehat{\boldsymbol{I}}_{\beta\zeta} = (\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}, \widehat{\boldsymbol{I}}_{\beta\lambda}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\widehat{\boldsymbol{I}}_{\beta\alpha} = -\frac{1}{n}\sum_{i=1}^{n}\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\big\}$$
$$\times\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\big\}\boldsymbol{X}_i,$$

and $\widehat{\boldsymbol{I}}_{\beta\lambda}$ is an m-vector with the kth component being

$$(\widehat{\boldsymbol{I}}_{\beta\lambda})_k = -\frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t_k)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) + \Delta_i\psi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\big\}\exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)$$
$$\times\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}.$$

Finally,

$$\widehat{\boldsymbol{I}}_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} R_i\boldsymbol{W}_{\mathcal{K},i}\boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}},$$
$$\widehat{\boldsymbol{I}}_{\beta\gamma} = -\frac{1}{n}\sum_{i=1}^{n}\big\{\Delta_i + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\big\}(1-R_i)\boldsymbol{W}_{\mathcal{K},i}.$$

References

Bachoc, F., Leeb, H. and Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics* **47**, 1475–1504.

Bachoc, F., Preinerstorfer, D. and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics* **48**, 440–463.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Bjørnland, T., Bye, A., Ryeng, E., Wisløff, U. and Langaas, M. (2018). Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Statistics in Medicine* **37**, 4234–4251.

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

Cheng, S., Wei, L. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

Cheng, S., Wei, L. and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association* **92**, 227–235.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* **34**, 187–202.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y. et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1–23.

Derkach, A., Lawless, J. F. and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* **102**, 988–994.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70**, 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Fithian, W., Sun, D. and Taylor, J. (2017). Optimal inference after model selection. *Preprint*. Available at

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

arXiv:1410.2597v4.

Gao, Y., Shi, Q., Xu, S., Du, C., Liang, L., Wu, K., Wang, K., Wang, X., Chang, L. S., He, D. et al. (2014). Curcumin promotes KLF5 proteasome degradation through downregulating YAP/TAZ in bladder cancer cells. *International Journal of Molecular Sciences* **15**, 15173–15187.

Higgins, J. P., Kaygusuz, G., Wang, L., Montgomery, K., Mason, V., Zhu, S. X., Marinelli, R. J., Presti Jr, J. C., van de Rijn, M. and Brooks, J. D. (2007). Placental S100 (S100P) and GATA3: Markers for transitional epithelium and urothelial carcinoma discovered by complementary DNA microarray. *The American Journal of Surgical Pathology* **31**, 673–680.

Hu, Y.-J., Li, Y., Auer, P. L. and Lin, D. Y. (2015). Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations. *Proceedings of the National Academy of Sciences* **112**, 1019–1024.

Lánczky, A., Nagy, Á., Bottai, G., Munkácsy, G., Szabó, A., Santarpia, L. and Győrffy, B. (2016). mirpower: A web-tool to validate survival-associated mirnas utilizing expression data from 2178 breast cancer patients. *Breast Cancer Research and Treatment* **160**, 439–446.

Lawless, J. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis* **24**, 28–44.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 907–927.

Pettitt, A. (1984). Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society: Series C* **33**, 169–175.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Sieh, W., Koebel, M., Longacre, T. A., Bowtell, D. D., DeFazio, A., Goodman, M. T., Høgdall, E., Deen, S.,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

Wentzensen, N., Moysich, K. B. et al. (2013). Hormone-receptor expression and ovarian cancer survival: An ovarian tumor tissue analysis consortium study. *The Lancet Oncology* **14**, 853–862.

The Cancer Genome Atlas Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111**, 600–620.

Välk, K., Vooder, T., Kolde, R., Reintam, M.-A., Petzold, C., Vilo, J. and Metspalu, A. (2011). Gene expression profiles of non-small cell lung cancer: Survival prediction and new biomarkers. *Oncology* **79**, 283–292.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Wong, K. Y. and Feng, J. (2023). Score tests with incomplete covariates and high-dimensional auxiliary variables. *Statistica Sinica* **33**, 1483–1505.

Wong, K. Y., Zeng, D. and Lin, D. Y. (2019). Robust score tests with missing data in genomics studies. *Journal of the American Statistical Association* **114**, 1778–1786.

Xu, W., Anwaier, A., Ma, C., Liu, W., Tian, X., Palihati, M., Hu, X., Qu, Y., Zhang, H. and Ye, D. (2021). Multi-omics reveals novel prognostic implication of src protein expression in bladder cancer and its correlation with immunotherapy response. *Annals of Medicine* **53**, 596–610.

Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93**, 627–640.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0391

censored data. *Journal of the Royal Statistical Society: Series B* **69**, 507–564.

Zeng, D., Lin, D. Y. and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica* **18**, 355–377.

Jiahui Feng

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

E-mail: jia-hui.feng@connect.polyu.hk

Kin Yau Wong

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Hong Kong Polytechnic University Shenzhen Research Institute, China

E-mail: kin-yau.wong@polyu.edu.hk