Statistica Sinica **37** (2027), 1-25
doi:https://doi.org/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 and*
[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.

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

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