Statistica Sinica **35** (2025), 1-20
doi:https://doi.org/10.5705/ss.202023.0118

LINEAR HYPOTHESIS TESTING FOR HIGH DIMENSIONAL TOBIT MODELS

Tate Jacobson* and Hui Zou

Oregon State University and University of Minnesota

Abstract: Few methods have been developed for conducting statistical inference in high-dimensional left-censored regression. Among the methods that do exist, none are flexible enough to test general linear hypotheses—that is, all hypotheses of the form $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}^*_{\mathcal{M}} = \boldsymbol{t}$. To fill this gap, we introduce partial penalized Tobit tests for testing general linear hypotheses in high-dimensional left-censored data. In particular, we develop partial penalized Wald, score, and likelihood ratio tests for high-dimensional Tobit models. We derive approximate distributions for the partial penalized Tobit test statistics under the null hypothesis and local alternatives in an ultra high-dimensional setting, finding that the tests achieve their nominal size asymptotically and that they are approximately equivalent for large n. In addition, we derive the tests' approximate power in this setting. We propose an alternating direction method of multipliers algorithm to compute the partial penalized test statistics. Through an extensive empirical study, we show that the partial penalized Tobit tests achieve their nominal size and that they are consistent in a finite sample setting. As an application, we analyze data from the AIDS Clinical Trials Group, using our partial penalized Tobit tests to test whether certain HIV mutations are significant predictors of HIV viral load.

Key words and phrases: Censoring, high-dimensional statistical inference, hypothesis testing.

1. Introduction

As it has become easier to collect large amounts of data, high dimensional modeling problems have become increasingly common in many domains. For researchers analyzing data with a left-censored response—common in some economic and medical applications—the availability of high dimensional data leads to the challenge of dealing with two modeling complications at once. As a motivating example, we consider the problem of modeling the relationship between human immunodeficiency virus (HIV) viral load and mutations in the HIV genome. The assays used to measure HIV viral load cannot detect the virus if its concentration is below a certain (known) threshold. Rather than discarding these observations, researchers simply record that the viral load is less than or equal to the detection threshold. As a result, the observed viral load is left-censored at the threshold value. At the same time, the number of participants

*Corresponding author. E-mail: jacobtat@oregonstate.edu