Statistica Sinica **37** (2027), 1-19
doi:https://doi.org/10.5705/ss.202024.0204

IDENTIFICATION AND EFFICIENT ESTIMATION IN REGRESSION ANALYSIS WITH RESPONSE MISSING NOT AT RANDOM

Qinglong Tian, Donglin Zeng and Jiwei Zhao*

University of Waterloo, University of Michigan and University of Wisconsin-Madison

Abstract: Missing-data is a pervasive problem in regression analysis, compromising the accuracy and efficiency of parameter estimates. This paper focuses on the challenging scenario of missing not at random (MNAR) data, where the missingness of a value is linked to the value itself. Traditional approaches to addressing MNAR data confront a trade-off: imposing stringent assumptions about the missingness mechanism can enhance efficiency but curtail robustness, whereas accommodating model misspecification can bolster robustness but at the expense of efficiency. In addition, assuming a nonparametric MNAR mechanism will lead to model identifiability issues. We propose a novel approach that overcomes this limitation. Firstly, we address the model identifiability issue using the shadow variable. Then, by leveraging the sieve method, we can model the MNAR mechanism nonparametrically. This approach achieves the best of both worlds: it gains robustness by avoiding strict assumptions about the missingness mechanism while simultaneously achieving the semiparametric efficiency bound for the parameter of interest (meaning our estimator has the lowest possible asymptotic variance). The paper delves into the theoretical framework, outlining conditions for identifiability, constructing the semiparametric likelihood function, and rigorously proving the estimator's semiparametric efficiency. Additionally, we present an EM-type algorithm for practical implementation, discussing the E-step and M-step iterations and variance estimation methods. Finally, simulations and a real-data application demonstrate the effectiveness of our proposed method compared to existing approaches.

Key words and phrases: Efficient estimation, identification, method of sieves, missing data, missing not at random, semiparametric efficiency.

1. Introduction

Regression analysis is a cornerstone of statistical research, revealing relationships between a response variable and one or more covariates. However, real-world data often suffers from missing values. Imagine a sociological survey where participants skip questions or a healthcare study where patients miss appointments. These missing data points can significantly impact regression

*Corresponding author. E-mail: jiwei.zhao@wisc.edu