Statistica Sinica **35** (2025), 1-24
doi:https://doi.org/10.5705/ss.202023.0075

HIGH-DIMENSIONAL SUBGROUP REGRESSION ANALYSIS

Fei Jiang*, Lu Tian, Jian Kang and Lexin Li

*University of California at San Francisco, Stanford University,
University of Michigan and University of California at Berkeley*

Abstract: Classical regression generally assumes that all subjects follow a common model with the same set of parameters. With ever advancing capabilities of modern technologies to collect more subjects and more covariates, it has become increasingly common that there exist subgroups of subjects, and each group follows a different regression model with a different set of parameters. In this article, we propose a new approach for subgroup analysis in regression modeling. Specifically, we model the relation between a response and a set of primary predictors, while we explicitly model the heterogenous association given another set of auxiliary predictors, through the interaction between the primary and auxiliary variables. We introduce penalties to induce the sparsity and group structures within the regression coefficients, and to achieve simultaneous feature selection for both primary predictors that are significantly associated with the response, as well as the auxiliary predictors that define the subgroups. We establish the asymptotic guarantees in terms of parameter estimation consistency and cluster estimation consistency. We illustrate our method with an analysis of the functional magnetic resonance imaging data from the Adolescent Brain Cognitive Development Study.

Key words and phrases: Adolescent Brain Cognitive Development Study, functional magnetic resonance imaging, group Lasso, high-dimensional regressions, subgroup analysis.

1. Introduction

Classical regression modeling generally assumes that all the subjects follow a *common* regression model with the same set of model parameters. In numerous applications, however, there may exist subgroups of subjects, and each group follows a *different* regression model with a different set of parameters. With ever advancing capabilities of modern technologies to collect more subjects and more covariates information, such data heterogeneity is becoming increasingly common. It thus becomes imperative to effectively identify subgroups of subjects and properly account for data heterogeneity in regression modeling (Ma and Huang, 2017).

Our motivation is the Adolescent Brain Cognitive Development (ABCD) Study, which plans to follow the brain development and health of over 10,000

*Corresponding author. E-mail: fei.jiang@ucsf.edu