Statistica Sinica **35** (2025), 1-19
doi:https://doi.org/10.5705/ss.202023.0082

ASYMPTOTIC ANALYSIS OF MIS-CLASSIFIED LINEAR MIXED MODELS

Haiqiang Ma[1] and Jiming Jiang*[1,2]

[1] *Jiangxi University of Finance and Economics and*
[2] *University of California, Davis*

Abstract: We study impact of class misspecification on the analysis of linear mixed models. Here, the misclassification means that some of the classes or groups associated with the random effects are mismatched. Such misclassification problems are becoming increasingly common in modern data science, including intentional and unintentional misclassifications. One important case of intentional misspecification is related to differential privacy; while a case of unintentional misspecification arises in classified mixed model prediction. Our study shows that standard asymptotic properties of the maximum likelihood and restricted maximum likelihood estimators, including consistency and asymptotic normality, remain valid under the misclassification provided that the proportion of the misclassified group numbers is asymptotically negligible in a suitable sense. Empirical results of simulation studies fully support our theoretical findings. A real-data example is considered.

Key words and phrases: Asymptotic behavior, differential privacy, linear mixed models, misclassification, random effects, robustness.

1. Introduction

Mixed effects models (Jiang and Nguyen, 2021) are widely used in practice. These models explore heteroscedasticity within the population, such as subpopulations or groups. These groups, characterized by the random effects, are of fundamental importance, and a main reason for the broad application of mixed effects models. It should be noted that the groups depend on the model we define, namely the mixed effects model—the data itself is not necessarily grouped, or grouped according to the mixed effects model.

A classical mixed effects model assumes that the group classifications are correctly specified. For example, there are 58 counties in the state of California, USA. Thus, if data are clustered according to those counties, numbered from 1 to 58, it is assumed that the county number is correctly specified for each data record. Specifically, consider a linear mixed model (LMM) with county-level random effects, expressed as $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, 58$, $j = 1, \ldots, n_i$, where y_{ij} is jth value of the response variable from county i, x_{ij} is an associated

*Corresponding author. E-mail: jimjiang@ucdavis.edu