Statistica Sinica **34** (2024), 1-16
doi:https://doi.org/10.5705/ss.202022.0228

ON COMBINING INDIVIDUAL-LEVEL DATA WITH SUMMARY DATA IN STATISTICAL INFERENCES

Lu Deng[1], Sheng Fu[2], Jing Qin[3] and Kai Yu[*2]

[1]*Nankai University,* [2]*National Cancer Institute, and* [3]*National Institue of Allergy and Infectious Diseases*

Abstract: Statistical models and inferences are typically based on measurements made on individual participants in a study (individual-level data). However, there is growing interest in improving statistical inference by taking advantage of aggregated summary-level data from other studies, such as statistics used in meta-analyses. Although the generalized method of moments (GMM) provides a flexible way of doing so, integrating external summary information does not always improve efficiency. Here, we provide a necessary and sufficient condition under which external summary information can be beneficial. We further extend the GMM to incorporate summary data generated from a population with a covariate distribution that is different from that of the individual-level data. Lastly, we compare the GMM with other integration procedures.

Key words and phrases: Empirical likelihood, generalized linear model, generalized method of moments, meta-analysis, summary statistics.

1. Introduction

Statistical inferences are usually conducted on detailed individual-level data observed on each participant in a study. Including relevant aggregated summary data from other studies would be preferred, although procedures for achieving such a goal might be not readily available. One exception is in the setting of meta-analysis, where estimates from comparable models established by different studies can be combined to form a more efficient estimate.

We consider a setting in which we use individual-level data (X, Y) from an internal study to investigate an underlying conditional model $f(Y \mid X; \theta)$, which specifies the conditional distribution of the outcome Y given the covariates X, with θ being the unknown parameter of interest. In addition, we assume we have summary data, represented by a set of estimates $\tilde{\beta}$, derived from external studies. The goal is to obtain a more efficient estimation of θ by combining the raw data (X, Y) from the internal study and $\tilde{\beta}$ from external studies. As in Qin (2000) and others (Imbens and Lancaster (1994); Qin et al. (2015); Chatterjee et al. (2016); Han and Lawless (2016); Cheng et al. (2018, 2019); Han and Lawless (2019); Kundu, Tang and Chatterjee (2019); Huang and Qin (2020); Zhang et al. (2020,

*Corresponding author.