A BAYESIAN DESIGN-BASED GENERALIZED LINEAR MIXED MODEL FOR TWO-STAGE SURVEY DATA

BY

OLAKIIGBE, ABIODUN KANMI (PG14/0017)

PGD, M.Sc. (FUNAAB, ABEOKUTA)

A Thesis Submitted to the Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Statistics.

November, 2025

ABSTRACT

Applied statistical analyses have increasingly featured model-based inference approaches tailored for complex survey designs. While incorporating survey, design information has been widely implemented in regression and Generalized Linear Models (GLMs). Its integration into Generalized Linear Mixed Models (GLMMs) has remained limited. This research focused on developing a Bayesian design-based generalized linear mixed model (GLMM) to accommodate the complexities of two-stage survey sampling, specifically addressing binary outcome data. The proposed binomial outcome, GLMM, was formulated by incorporating design sampling weights within a latent Gaussian model. In this framework, the binary response for each household in cluster i of stratum h followed a Bernoulli distribution, conditional on the N×1 vector w of survey weights. The conditional expectation of the response was linked to a linear predictor with both fixed and random effects, forming a latent field vector. The components of this field were assumed to be conditionally independent, given a structure of entity-specific stochastic dependence. Consequently, the latent field satisfied the Gaussian Markov random field (GMRF) property, characterized by a sparse precision matrix. Priors were assigned to the hyperparameters of this latent Gaussian field. Due to the inclusion of design weights, the resulting finite population likelihood became analytically intractable. The Integrated Nested Laplace Approximation (INLA) method was adopted to address this challenge, leveraging Laplace transformations and a second-order Taylor expansion for computationally efficient inference. The performance of INLA-based GLMM (INLA-GLMM) was compared with classical and Bayesian approaches through simulations involving various sample sizes: n = 10, 20, 25, 150, 500, 2000, 5000, and 10,000. In addition, a prior sensitivity analysis was conducted. The model was applied to real-world data from the 2021 Nigerian Malaria Indicator Survey (NMIS). The result

demonstrated the impact of including design information in the analysis. For instance, the estimated proportion of respondents with malaria incidence differed when design effects were excluded and when they were included, with 8.1% and 7.4%, respectively. Similarly, estimates changed from 25.7% to 27.5% for households reporting no malaria in children. Simulations showed that INLA-GLMM consistently achieved lower log-likelihood values compared to classical Adaptive Gauss-Hermite Quadrature (AGHQ) and weighted AGHQ (WTD-AGHQ) methods for all population sizes, except the smallest (n = 10, 20, 25). Deviance Information Criterion (DIC) values for INLA-GLMM were also lower than those from Markov Chain Monte Carlo (MCMC) GLMM for most population sizes. The prior sensitivity analysis revealed that while fixed effect estimates remained consistent across priors, random effect precision estimates were more sensitive in smaller samples. On the NMIS dataset, INLA-GLMM performed better, achieving a lower log-likelihood (-6425.06) compared to AGHQ (-6327.10) and WTD-AGHQ (-6333.00). Likewise, the INLA-GLMM DIC value (12383.76) is better than that of the MCMC-GLMM (12388.94). The INLA-GLMM also recorded the shortest running time, completing in just 3 minutes, whereas MCMC-GLMM took 47 minutes, making it highly suitable for large-scale survey data analysis. The findings from the study supported the use of design-based Generalized Linear Mixed Models using Integrated Nested Laplace Approximation for analyzing binary outcomes in large survey datasets.