Contribution ID: **184** Type: **Poster**

Effect of Class Imbalance and Multicollinearity on Parameter Estimation in Binary Logistic Regression

Thursday 24 July 2025 11:45 (15 minutes)

Poster

Effect of Class Imbalance and Multicollinearity on Parameter Estimation in Binary Logistic Regression

Author

María Lucía Feo-Serrato

Affiliation

Complutense University of Madrid

Abstract

Class imbalance—or, more broadly, rare events data—presents a common challenge in binary logistic regression (BLR), particularly in contexts where the phenomenon of interest has low prevalence (e.g., rare diseases or risk of abuse). The likelihood function employed in BLR tends to underestimate the probability of rare events (Oommen et al., 2011). Moreover, rare events exert a disproportionate influence on the variance-covariance matrix used to compute the standard errors of the estimated coefficients; consequently, any error in estimating the probabilities of these events can be amplified within this matrix, ultimately compromising the precision of the coefficient estimates (King and Zeng, 2001).

The primary objective of this study is to analyze the effects of class imbalance and multicollinearity on parameter estimation in binary logistic regression, and to assess which techniques might mitigate their impact, thereby yielding more consistent and efficient model estimates.

Monte Carlo simulations were employed to examine the influence of class imbalance (ranging from 50:50 to 95:5), multicollinearity (none, moderate, and high), and sample size (from 250 to 10,000) on parameter recovery. The estimation methods compared in the study included BLR, Ridge, LASSO, Elastic Net, and SMOTE, with performance evaluated in terms of average bias, standard error (SE), and root mean square error (RMSE).

The results indicate that BLR maintains moderate bias under conditions of low or moderate imbalance (≥70:30) and large sample sizes (>1,000). However, under extreme imbalance (95:5), both SE and RMSE increase significantly. Ridge and Elastic Net demonstrated greater stability in scenarios characterized by pronounced imbalances and small sample sizes, whereas SMOTE exhibited high variability and substantial bias in cases of severe imbalance.

From a theoretical perspective, this study contributes to a nuanced understanding of the impact of class imbalance in explanatory contexts, which are prevalent in psychology and other behavioral sciences. Moreover, it underscores the importance of selecting appropriate methods tailored to the specific conditions of the data. Specifically, regularization methods—particularly Ridge and Elastic Net—emerge as promising tools for managing imbalances in practical applications. These methods are especially valuable in contexts where the stability of estimates is paramount, such as in public health studies or security risk analyses. Nevertheless, their application should be accompanied by rigorous evaluations using key metrics such as RMSE, SE, and probabilistic calibration indices.

In summary, while BLR proves effective under moderate imbalances and large samples, Ridge and Elastic Net are preferable in more complex scenarios. Validating these findings with real-world data and exploring advanced approaches, such as Bayesian methods, is recommended to further enhance the reliability of explanatory models.

Keywords

class-imbalance, multicollinearity, binary logistic regression

Primary authors: Mrs FEO-SERRATO, María Lucía (Complutense University of Madrid); Mr OLMOS ALBACETE, Ricardo (Autonomus University of Madrid); CUEVAS UREÑA, Clara

Presenters: Mrs FEO-SERRATO, María Lucía (Complutense University of Madrid); Mr OLMOS ALBACETE, Ricardo (Autonomus University of Madrid); CUEVAS UREÑA, Clara

Session Classification: Poster Session 3

Track Classification: Statistical analyses: Statistical analyses