Multilevel logistic regression with complete separation

Robert Clark, Australian National University

1. Introduction

Logistic regression is a widely used technique for modelling the effect of a set of covariates on a binary outcome variable. It can answer questions such as “how do the context and grammatical features of a sentence affect the probability that a language learner will make a correct grammatical choice?” in the context of linguistics, and “how do environmental and/or treatment conditions affect the likelihood of observing a particular species (and its abundance) at a location?” in the context of ecology.

Scientists applying logistic regression sometimes encounter a surprising phenomenon, where some estimated regression coefficients are startlingly large and some or all of the accompanying standard errors are even larger. The p-values (one indicator of the evidence of each covariate’s effect) typically indicate that there is no evidence of effects, even though the researcher can see clear signs of a strong effect in simple descriptive analyses of the data. This situation is typically due to *complete separation*, where some maximum likelihood estimates are literally infinite (see for example [3] pp. 147-149). Paradoxically, it occurs most often when covariate effects are very substantial (although it can also be due to over-fitting).

In principle, infinite parameter estimates are meaningful on the logit scale used in logistic regression; they correspond to a fitted probability of 0 or 1 for some covariate values. In practice, they are inconvenient because they make commonly used normal-based confidence intervals, p-values and Wald confidence intervals invalid. The problem arises not just with the canonical logit link, but also inherently with other link functions such as the probit and complementary log-log link functions.

2. Approaches

When faced with complete or quasi separation, many applied researchers erroneously discard the predictor causing the complete separation just to aesthetically make the model look “better”. Aside from this, two common solutions are also often used to remedy the large coefficients: penalised likelihood [1] and Bayesian analysis (e.g [2]) making use of regularising or weakly informative priors.

There is limited literature on combining these approaches with multilevel modelling, which is commonly required in both linguistics and ecology.

3. Overview of Research

The aim of this research is to review and evaluate methods for complete and quasi separation in statistical linguistic and ecology, particularly for multi-level data. Guidelines will be developed for applied researchers become, to increase awareness of this commonly occurring feature of binary data and of how it can be remedied.

References

(1) Firth D. (1993) Bias reduction of maximum likelihood estimates. Biometrika;80:27–38

(2) Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. Bayesian Analysis, 13(2), 359-383.

(3) Hosmer, Lemeshow and Sturdivant (2013), Applied Logistic Regression, 3rd Edition, Wiley: New York.