Modeling Approaches to Binary Logistic Regression with Correlated Data

Cornell Statistical Consulting Unit

1 Introduction

Binary logistic regression is a versatile statistical tool for analyzing binary responses in a wide range of disciplines. Similar to ordinary least squares regression, one key assumption in regular binary logistic regression is that observations are independent of each other. Violations of the assumption of independence of observations may result in incorrect statistical inferences due to biased standard errors.

2 Examples of correlated data with binary outcomes

In practice many situations may give rise to correlated data. In longitudinal studies, subjects are measured over time on the outcome of interest. Measurements collected at different time points on the same subject are not independent from each other. Similarly, in studies with naturally occurring groups (e.g. persons within families) responses collected from members of the same group tend to be more similar than observations from different groups.

For example, suppose we have a data set with patients in 8 different hospitals. In each hospital, patients were assigned either to a specific treatment group or to a placebo (control) group. The binary outcome of the study is whether a subject's health has improved or not after a specified period of time. The data collected from patients from the same hospital are likely to be correlated. The correlation of these measurements needs to be accounted for in order to obtain valid inferences.

3 Statistical analysis

In practice, two types of statistical models are widely used to model binary data while accounting for correlation of the binary measurements in the statistical analysis. One is the marginal or population-averaged model. The second is the conditional or subject-specific model. They differ in the way the correlation of measurements is incorporated in the model, and also in the interpretation of estimated model parameters.

3.1 Marginal model

In the marginal model, the correlation of the measurements is accounted through a robust covariance matrix. A group effect is not explicitly included in the model. In contrast, in the

conditional or subject–specific model, the dependence of observations is modeled by including a group specific effect in the model.

For the marginal model inference, several approaches have been available in practice. One popular method is the Generalized Estimating Equations (GEE). It produces consistent estimators of the regression parameters and their variances. With the GEE approach, the estimation of standard errors can be done in two ways: model-based or empirical-based. The model-based variance estimation method allows the specification of an (assumed) correlation structure, to account for the non independence of observations. The empirical variance estimation method, also called the robust standard error method, uses the empirical sample dependence to adjust the standard errors obtained with a regular logistic regression model (that assumes independence of observations). The GEE estimation is a non likelihood or quasi-likelihood based method. The GEE method is implemented in many statistical analysis software packages, including SAS, SPSS, Stata and R.

3.2 Conditional model

The conditional or subject-specific logistic regression has a term in the model for each group, or hospital in our example. This group effect can then be modeled as either a fixed or random effect. Therefore, two approaches for conditional modeling can be used in practice. One approach treats the groups as fixed effects, while the second as random effects.

The fixed-effects logit model treats the group specific terms as fixed effects and uses a conditional maximum likelihood method for estimation. With the conditional maximum likelihood the group-specific effects are removed from the model, and model effects are estimated. The random effects approach to conditional modeling of correlated binary data treats the group effects as random and uses maximum likelihood methods for estimation. There is one important difference between the fixed and random conditional models. The fixed effects model only allows the intercepts to differ by group, but assumes that all observations are independent within and across groups. In contrast, the random effects approach directly models the correlation between observations within a group. The conditional model is implemented in several statistical software packages, including SAS, Stata, and R.

4 Interpreting regression coefficients

For linear models with normally distributed data, the effect of a predictor variable under marginal and conditional models has the same interpretation. With binary correlated outcomes, the regression coefficient of a predictor variable may have two different interpretations under the marginal and the conditional on random effects model, respectively, due to the non-linearity of the logistic regression function.

For the above hospital example, under the marginal model, the exponentiated treatment coefficient represents the odds of an average patient in the treated group to have improved health compared to an average patient in the control group. Under the conditional on random effects model, the exponentiated treatment coefficient represents the odds of a patient’s health being improved, for a treated person compared to the same person if he/she were not treated (control group).

Another illustration is to consider the smoking status of the patient as the predictor variable in the above example. In the marginal model, the exponentiated smoking status coefficient represents the odds of improved health for an average patient from the population of patients who smoke, compared with those who do not smoke. For a conditional on random effects model, the exponentiated smoking status coefficient represents the odds of improved health for an individual patient who is a smoker, compared to the same patient if he/she were not a smoker.

The choice of marginal versus conditional model depends on the objectives of the study. If the focus of the study is to predict the average rate of improved health, then a marginal model may be suitable. If the focus is on understanding the role of individual hospital characteristics in improving a patient's health, then the conditional model may be appropriate.

5 References

Agresti, A., An Introduction to Categorical Data Analysis, 2nd Edition, 2007

Pendergast J. et al., A Survey of Methods for Analyzing Clustered Binary Response Data, International Statistical Review, Vol. 64, No. 1. (Apr.1996), pp. 89-118.

Allison P., Logistic Regression using the SAS System: Theory and Applications, SAS Institute, 1999.

Created May 2021. Last updated April 2022.