PATTERNS OF DENSITY PLOTS OF RANDOM EFFECTS IN MIXED EFFECTS LOGISTIC REGRESSION MODELS

C. Manoj[1], P.Wijekoon[2] and Roshan D. Yapa[2]

[1]Postgraduate Institute of Science, University of Peradeniya
[2] Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya

Introduction

The logistic regression model is the widely used generalized linear model for modeling binary response with one or more explanatory variables which can be quantitative or qualitative. Generally this model assumes a binomial distribution for the random component. However in many practical situations it can be realized that the variability of the random component is greater than the assumed binomial variability, which is commonly known as overdispersion. This phenomenon can be simply identified by the value of dispersion parameter of the model with the absence of systematic deficiencies. Generally, overdispersion is caused by the violation of the hypotheses of the binomial distribution which are (i) correlation between the individual Bernoulli trials, (ii) variability of the success probability within an explanatory level and (iii) loss of information on some explanatory variables. The way of analyzing the last cause of overdispersion is developing a mixed effect logistic regression model, and analyzing its random effects. The mixed effect logistic regression model is given by, $\text{logit}(p_i) = \eta_i + \gamma Z_i$, where p_i is the success probability in the ith explanatory level, η_i is the ith systematic component, $i = 1......n$, and Z_i's are random effects whose mean is zero and variance is one and the coefficient $\gamma \geq 0$. The method of maximum likelihood is applied to obtain estimates of model parameters. It is generally assumed that the random effects Z_i's are independent standard normal variables, and with this assumption Z_i's are simply integrated out from the likelihood function. Then the resultant marginal likelihood function is used to obtain estimates of unknown parameters in the model by means of numerical methods. Although it is assumed that Z_i's are normal for simplicity and generality, some other alternative distributions may also exist (Collett, 1994). Handayani and Kurnia (2006) discussed the effects of overdispersion by taking an example, and they have shown that mixed effect logistic regression model can be well used to model the overdispersion.

In this paper, we try to identify the patterns of density plots of random effects in mixed effect logistic regression models first by taking two different examples. Then a simulation study is done to understand the cause of the patterns of density plots of random effects.

Methodology

Two examples are taken to understand the distribution of random effects. The first dataset (Cell data) is the same one from Handayani and Kurnia (2006) and the second dataset (Satisfaction data) is taken from *A handbook of small datasets* (Hand *et al.*, 1994). The cell data contains dependent variable y = number of cells survived out of 400 cells after the irradiation (Binomial) and two independent variables Dish and Occasion. For all observations, it is noted that the number of successes is less than the number of failures. The satisfaction data concerns the levels of satisfaction with housing. Five households have been randomly selected from 18 clusters by means of cluster sampling. Each household has given a single response (unsatisfied, satisfied or very satisfied) about their satisfaction with their home. For the purpose of analyzing this with the binary logistic regression model, the trichotomous satisfaction variable is converted to a dichotomous variable by combining the responses satisfied and very satisfied together. In this case for 7 observations out of 18, the number of successes is less than the number of failures. For the rest, the number of successes is greater than the number of failures. For both datasets, first the standard logistic regression models are fitted, and then the overdispersion was checked. In order to deal with overdispersion, mixed effect logistic regression models are fitted and random effects are extracted from the models and tested for normality. For the cell data, the random effects agree with normal assumption. For the satisfaction data, the random effects depart from normality. To understand a more appropriate distribution, a kernel density plot of random effects is plotted, and the outcome is analyzed further. To confirm the results obtained from the cell data, a simulation study is performed by simulating 20 binomial responses from the standard model, and refitting the mixed effect models with the new simulated responses. Random effects of simulated models are also tested for normality. R package *lme4* and the function *simulate (stat)* are used for the analysis.

Results and Discussion

In cell data, the standard logistic regression model shows that overdispersion is present, (dispersion parameter=19.0627). The p- values (> 0.01) of the three normality tests (Anderson-Darling, Lilliefors and Shapiro-Wilk) for random effects of original and simulated mixed effects logistic regression models demonstrate that the random effects are normal at 1% significance level. The kernel density plot of random effects (Figure 1) also shows a unimodal distribution.

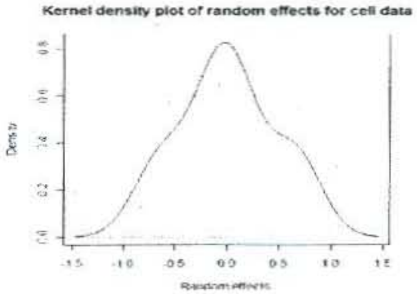

Figure 1. Kernel density plot of random effect for cell data

The dispersion parameter value (2.5812) of the standard logistic regression model for the satisfaction data also indicates the presence of overdispersion. However, the random effects do not agree with the normality assumption since p- vales are p<0.01 in all three normality tests.

Moreover, kernel density plot (Figure 2) of random effects for satisfaction data suggests a bimodal density curve for random effects.

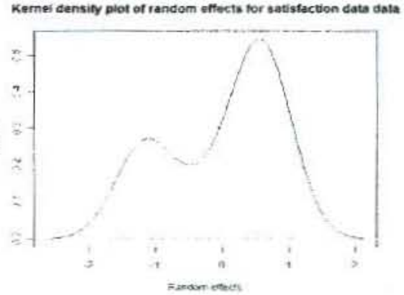

Figure 2. Kernel density plot of random effect for satisfaction data

With a deep look on this kernel density plot, it can be realized that one peak arises for negative random effects and another peak arises for positive random effects. Furthermore, negative random effects represent the cases where number of unsatisfied is less than the number of satisfied whereas positive random effects represent the cases where number of unsatisfied is greater than the number of satisfied. However, when testing the normality separately for negative and positive random effects it can be concluded that both random effects, individually follow normality at 1 % significance level, and together they satisfy the conditions of bimodal normal distribution, a mixture of two unimodel normal distributions.

Conclusion

The random effects logistic regression model holds considerable promise on modeling overdispersion when there is loss of information on some explanatory variables. Our illustrative examples indicate that the normality assumption of random effects has a relation with the number of successes greater than or less than the number of failures presented in each explanatory levels. Two important results achieved from the above two examples are if for each explanatory level we have number of successes less than number of failures (or number of failures less than number of successes), the random effects agree with the unimodel normal assumption. On the other hand, if there are considerable amount of both number of successes greater than number of failures and number of failures greater than number of successes, the random effects have a *bimodal normal distribution* with the mixture of two unimodel normal distributions.

References

Collett, D. (1994). Modelling Binary Data, Chapman and Hall, London.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E.(1994). A Handbook of Small Data Sets, Chapman and Hall, London.

Handayani, D. and Kurnia, A.(2006). Mixed Effect Model Approach for Logistics Regression Model with Overdispersion, Proceeding at The First International Conference on Mathematics and Statistics (ICoMS-1), Bandung, June 19-21.