ADJUSTMENT ON WILLIAM'S METHOD WHEN NEGATIVE VARIANCE COMPONENTS OCCUR IN UNDERDISPERSED LOGISTIC MODELS

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

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

Introduction

In many practical situations it is identified that the variance of the random component of logistic regression model exhibits a different value than the usually assumed binomial variability. This phenomenon can be simply identified by the value of dispersion parameter of the model with the absence of systematic deficiencies. When the value of the dispersion parameter exceeds one, the model is said to be overdispersed, while a value less than one indicates an underdispersed model. Williams (1982) developed a quasi likelihood estimation method, which utilizes the relationship between the expectation and variance of random components, using iterated reweighted least squares equation to estimate parameters of overdispersed logistic regression. Engel,*et.al.* (1993) have showed Williams' method can also be successfully used to analyse underdispersed logistic regression models when the constraints on the variance component hold. In this paper we propose an intuitive method which is equivalent to William's method, however which can be used for underdispersed logistic regression models when no constraints are dictated on variance component.

Theory and Methods

Let Y_j be the number of successes in n_j, not necessarily independent trials. Define Y_j's as random components of logistic regression models with $p-1$ explanatory variables, $X_{j1},...X_{j,P-1}$ and $X_{j0}=1$ for $j=1,2,....n$. Then the model has the form $\text{logit}(p_j)=\lambda_j=\sum_{s=0}^{P-1}x_{js}\beta_s$. In addition, let $Z_{j1},Z_{j2}.......Z_{jn_j}$ be the correlated Bernoulli trials with success probability p_j whose sum is Y_j. Let $Corr(Z_{ji},Z_{ji'})=\alpha_j$ $i,i'=1,2.......n_j$ and $i \neq i'$. Hence $Var(Y_j)=n_jp_j(1-p_j)(1+\alpha_j(n_j-1))$ (Collett, 1994). Positive correlation results $Var(Y_j)>n_jp_j(1-p_j)$ and leads to overdispersion, whereas negative correlation results $Var(Y_j)<n_jp_j(1-p_j)$ and leads to underdispersion. In the case of negative correlation in order to have a positive value for the variance of the random component, a restriction on α_j is imposed so that $-(n_j-1)^{-1}\leq\alpha_j\leq 1$. To incorporate Williams method of estimation, $Var(Y_j)$ is written

as $Var(Y_j) = v_j w_j^{-1}$, where $v_j = n_j p_j (1 - p_j)$ and $w_j^{-1} = (1 + \alpha(n_j - 1))$, and a common value α is assumed for α_j, which is unknown and needs to be estimated. Williams proposed a method to estimate α and the model parameters, in which he restricts α>0 and accounts only for overdispersion. Engel, *et.al.* have showed Williams' method works for underdispersed models whenever $-(n_j - 1)^{-1} \le \delta_j$. By contrast there are practical situations identified with negative correlation with $\alpha < -(n_j - 1)^{-1}$.This happens if the numbers of trials of one or more binomial observation are comparably large. Further, due to the removal of the above constraint, weights introduced in the variance components may have negative values. Gao, *et.al.* (2008) found negative variances may occur if no constraints are imposed in the estimation of variance components, and they suggested several methods to deal with these situations. Williams' method fails to analyse the occurrences of negative variances when underdispersion is present.

In this paper we suggest to use the absolute values of the weights when updating the model with new weights and in all consequent steps in Williams method whenever w_j<0.

The proposed adjusted Williams method is given here:

Step 01:

Assume α = 0 and estimate the parameters of logistic regression with an assumption of binomial random component and evaluate the value of the Pearson's goodness of fit statistics X^2.

Step 02:

Compare the X^2 value with the corresponding degrees of freedom of the model, $n-p$, and if $\frac{X^2}{n-p} >> 1$ conclude that overdispersion exists, and if $\frac{X^2}{n-p} << 1$ conclude that underdispersion exists. Then estimate

$$\hat{\alpha} = \frac{X^2 - (n-p)}{\sum_j \{(n_j - 1)(1 - v_j{}^* q_j)\}}.$$

Step 03:

Calculate the new weights $w_j = \frac{1}{(1 + \hat{\alpha}(n_j - 1))}$, which may have negative values when $\hat{\alpha} < -(n_j - 1)^{-1}$. Since variance of the random component must be a positive value when refitting the model by updating the weights, use the following absolute weighted least squares equations

$\boldsymbol{X^T W' V^* X\hat{\beta} = X^T W' V^* Y^*}$, where $\boldsymbol{W'} = diag(|w_i|)$, $\boldsymbol{X}$ is the regressor matrix of rank p, $\boldsymbol{V^*} = diag(v_j{}^*)$,

$$y_j{}^* = \lambda_j{}^* + \frac{(y_j - n_j p_j{}^*)}{v_j{}^*},$$

and $p_j{}^*$, $\lambda_j{}^*$, $v_j{}^*$ are values of p_j, λ_j, v_j corresponding to $\boldsymbol{\beta}^*$ initial estimates of $\boldsymbol{\beta}$. Then recalculate X^2 for the updated model.

Step 04:

If recalculated X^2 is close enough to $n-p$, the estimates of $\hat{\alpha}$ and β are satisfactory. If not re-estimate α as

$$\hat{\alpha} = \frac{X^2 - \sum_j |w_j|\left(1 - |w_j| v_j^* q_j\right)}{\sum_j \left\{ |w_j| (n_j - 1)\left(1 - |w_j| v_j^* q_j\right) \right\}}$$ and

go to step 03.

Results

We attempt to illustrate the necessity of the adjustments given in this paper, and its ability for producing better fits for underdispersed logistic regression models by means of an example. The data (D.Collett, 1994) refer to the number of fasteners failing out of a number tested subjected to varying pressure loads. The binary outcome of this example is whether or not a particular aluminum fastener fails when a given load is applied to it. When a logistic regression model is fitted to this data the following results have been obtained. Pearson chi-square statistic X^2=0.3706 and the deviance value is 0.3719, on 8 degrees of freedom. Hence the value of the dispersion parameter is 0.0464<<1. This suggests an underdispersion and may be explained as binary responses from the fasteners subjected to a given load are negatively correlated, or the data is artificially generated. Moreover initial estimate of α is $\hat{\alpha}$ = -0.0142 and the value of $-(n_{\max} - 1)^{-1}$ is -0.0101. Thus, clearly $\hat{\alpha} < -(n_j - 1)^{-1}$ and Williams' method cannot be used for this situation. Since X^2 is far less than the degrees of freedom, we try to increase the value of X^2 with adjusted Williams method described above. To carry out the proposed method an R function, "adjusted.williams", is written. By employing this method, a substantial increment is gained on both Pearson's goodness of fit statistics and the deviance of the model after 7 iterations. At the 7th iteration X^2=6.0393 and the deviance is 6.0124 on 7 degrees of freedom. Above goodness of fit test statistic and deviance values would be taken as satisfactory values.

Conclusion

The adjusted William's method proposed in this paper holds considerable promise on modeling underdispersion, when negative weights occur for the variance components. Indeed this novel method can be used as a generalization to Williams' method without imposing any constraints on the correlation between Bernoulli trials in responses.

References

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

Engel, B. and Brake, J.T. (1993). Analysis of Embryonic Development with a Model for Under or Overdispersion Relative to Binomial Variation. Biometrics, 49(1): 269-279.

Gao, B., Li, S., Li, W., Li, S. and Wang, X. (2008). Statistical Analysis of Negative Variance Components in the Estimation of Variance Components. International Association of Geodesy Symposia, 132(III): 293-296.

Williams, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. Applied Statistics, 31(2):144-148.