Paper 9340-2016

Exact Logistic Models for Nested Binary Data in SAS®

Kyle Irimata, Arizona State University; Jeffrey Wilson, Arizona State University

ABSTRACT

The use of logistic models for independent binary data has relied first on asymptotic theory and later on exact distributions for small samples, as discussed by Troxler, Lalonde, and Wilson (2011). While the use of logistic models for dependent analysis based on exact analyses is not common, it is usually presented in the case of one-stage clustering. We present a SAS® macro that allows the testing of hypotheses using exact methods in the case of one-stage and two-stage clustering for small samples. The accuracy of the method and the results are compared to results obtained using an R program.

INTRODUCTION

Logistic regression models are commonly used in the analysis of binary outcome data across a number of different disciplines including medical research, social sciences and educational research. In many cases, it is also common to encounter correlation amongst the dichotomous outcomes as a result of clustering inherent in the data or collection scheme. Large sample, asymptotic approaches are the most frequently utilized in such situations and a variety of methods have been developed to address these associations.

In contrast, the use of the exact distribution in the case of correlated binary outcomes has been given relatively little attention, especially in the case of more than one level of clustering. Troxler, Lalonde and Wilson (2011) proposed an extension to exact techniques for hypothesis testing of data with second level clustering effects. This technique is useful in analyzing correlated binary data for sparse or small data sets and also offers flexibility and expandability in regards to the number of clustering levels it can address. We provide a macro implementation in SAS for hypothesis testing of significant predictors in the presence of two levels of clustering using the exact distribution. This macro can be applied to analysis of data with either one or two levels of clustering. Two illustrations are considered and the results are compared with those produced by an existing R program.

EXACT LOGISTIC REGRESSION MODELS

INDEPENDENCE MODEL

The logistic regression model is the most often used approach for relating a binary outcome to one or more covariates. This model is a member of generalized linear model with the logit link and systematic component consisting of the covariates such that

$$l\,n\left[\frac{p_j}{1-p_j}\right] = \alpha + \boldsymbol{x}_{\boldsymbol{j}}'\,\boldsymbol{\beta}$$

where $\boldsymbol{p_j}$ is the probability of success for the jth observation, α and the coefficients within the vector $\boldsymbol{\beta}$ are fixed but unknown parameters and $\boldsymbol{x_j}$ is the vector of covariate values for the jth observation.

Let $\boldsymbol{Y} = (\boldsymbol{Y_1}, \dots, \boldsymbol{Y_n})$ denote the vector composed of n binary observations. The joint probability mass function for this vector, under the assumption of independent observations is

$$P(Y_1 = y_1, \dots, Y_n = y_n) = ex\,p\left\{\alpha + \sum_{j=1}^{n} y_j \boldsymbol{x}_{\boldsymbol{j}}'\boldsymbol{\beta} + c\right\}$$

where c is a normalizing constant obtained by summing over all possible outcomes (Cox 1970).

Estimation of α and $\boldsymbol{\beta}$ is most commonly conducted through the use of large sample techniques such as maximum likelihood estimation, although such approaches may fail for sparse data sets. In such cases, alternative approaches offer a better method for large samples, as presented by Cox (1970). His method focuses on estimation of $\boldsymbol{\beta}$ in particular, through the use of sufficient statistics, while other nuisance parameters are treated through the use of ancillary statistics. This approach involves the exact computation of the conditional distribution of the sufficient statistics over all possible sets of outcomes that lead to the observed values of the ancillary statistics.

ONE-STAGE CLUSTERED LOGISTIC REGRESSION

When the assumption of independence amongst observations is not satisfied, it becomes necessary to account for association within groups of data using more complex models. In particular, these correlations often occur as a result of clustering, in which observations from the same cluster have some inherent association with each other. Many large sample approaches such as generalized estimating equations (GEE) have been developed (Liang & Zeger 1986); however comparatively few exact methods have been explored. Corcoran, et al. (2001) provided an exact approach for one-stage clustered data, which relies on conditioning arguments similar to those used by Cox (1970).

For one-stage clustered data, we let Y_{ij} be the j^th^ observation $(j = 1, \ldots, n_i)$ within the i^th^ cluster $(i = 1, \ldots, N)$ and let $\boldsymbol{Y_i} = \left(Y_{i1}, \ldots, Y_{in_i}\right)$ denote the vector composed of n_i outcomes for the i^th^ cluster and let $\boldsymbol{x} = (x_1, \ldots, x_N)$ be the vector of these covariate values. Define $Z_i = \sum_{j=1}^{n_i} Y_{ij}$ to be the sum of outcomes in the i^th^ cluster with a corresponding probability mass function (Wilson and Lorenz 2015). We can further express this probability mass function in terms of the sufficient statistics.

TWO-STAGE CLUSTERED BINARY MODELS

The one stage clustered logistic regression model may be further extended to account for two stages of clustering, in which the data consist of first-stage clusters, each of which in turn contain second-stage clusters. Let Y_{ijk} denote the k^th^ $\left(k = 1, \ldots, n_{ij}\right)$ observation within the j^th^ $(j = 1, \ldots, J_i)$ secondary cluster nested within the i^th^ $(i = 1, \ldots, I)$ primary cluster. We assume that the amount of within cluster association is constant for all second-stage clusters and similarly that the amount of within cluster association is constant for all first-stage clusters. Further, let $\boldsymbol{x_i} = \left(x_{i1}, \ldots, x_{iJ_i}\right)$ be the vector composed of the covariate values x_{ij} in the ij^th^ secondary cluster and let $\boldsymbol{x} = (x_1, \ldots, x_I)$.

Define $Z_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}$ to be the sum of the binary outcomes in the ij^th^ second cluster and consider the probability mass function for the vector $\boldsymbol{Z_i} = \left(Z_{i1}, \ldots, Z_{iJ_i}\right)$ (Wilson and Lorenz 2015). Under the assumption of independence between the first-stage clusters, the probability mass function for the overall response vector $\boldsymbol{Z} = (Z_1, \ldots, Z_I)$ can be expressed as a function of the sufficient statistics.

HYPOTHESIS TESTING

In many cases we may be interested in testing for significant covariate effects using the hypothesis:

$$H_0\!: \beta = 0 \text{ vs. } H_a\!: \beta > 0$$

For $\beta = 0$, the conditional probability mass function under H_0 is given by

$$P(\boldsymbol{Z} = \boldsymbol{z} | \boldsymbol{x}; s_1, s_2, s_3) = \frac{\left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}\right]}{\sum_{\boldsymbol{z}^* \in \Gamma(s_1, s_2, s_3)} \left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}\right]}$$

where s_1, s_2, s_3 are sufficient statistics.

A one-sided p-value for testing the given hypothesis based on the likelihood ratio for $\beta = 0$ versus the alternative $\beta > 0$ is then given by

$$\Pr(t > t_{obs} | H_0, \boldsymbol{x}, s_1, s_2, s_3) = \sum_{\boldsymbol{z}^* \in \Gamma(s_1, s_2, s_3):\, t(\boldsymbol{z}^*) \geq t_{obs}} \left[\frac{\left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}\right]}{\sum_{\boldsymbol{z}^* \in \Gamma(s_1, s_2, s_3)} \left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}\right]} \right]$$

where we have that the likelihood ratio is decreasing in t. We therefore reject H_0 and can conclude that β is significantly larger than zero when this p-value is less than our chosen significance level of α.

SAS MACRO

The approaches discussed previously for conducting hypothesis tests for exact logistic regression can be utilized in SAS through the general macro call:

```
%exactlogistic(data=, levels=1);
```

The first argument, *data*, contains the SAS dataset to be analyzed by the macro. The data must contain the first-stage labels in the first variable (or a constant term if the data is only one-stage), the number of observations in each second-stage cluster in the second variable, the observed counts for each second-stage cluster in the third variable and the covariate values as the fourth variable.

The second argument, *levels*, specifies which of the models to utilize. By default, the p-value for the hypothesis test is calculated for a one-stage model with one level of clustering. Changing this value to *levels = 2* will produce a p-value for the hypothesis test under the two-stage framework.

Each call of the macro returns a p-value for the hypothesis test of $H_0: \beta = 0$ vs. $H_a: \beta > 0$. This macro is available at http://www.public.asu.edu/~jeffreyw.

This macro relies on SAS IML and first calculates the values of the sufficient statistics for the data set for the designated model. The calculations are completed using an iterative approach in which each possible set of outcomes are enumerated, conditioned on the sample sizes for each cluster. Each of these possible outcome vectors are saved and the sufficient statistics for each set are calculated to be used in the calculation of the p-value. Due to the computationally intensive nature of this approach, we also incorporate a practical check based on one of the sufficient statistics to reduce the required memory and processing time for the macro.

DATA EXAMPLE

COMPARISON TO ANALYSES IN R

To illustrate the use of the %exactlogistic macro we analyzed a correlated data set with identifiers removed, Troxler, et al. (2011) and Have, et al. (1999). These data have generic labels and a nested correlation structure, wherein group denotes the first stage of clustering, n_{ij} denotes the number of binary observations within each second stage cluster, z_{ij} denotes the count of positive outcomes in each second-stage cluster and x denotes the value of the covariate for each second-stage cluster. The data analyzed for this illustration is given in Table 1.

Group	n_{ij}	z_{ij}	x
1	4	2	2
1	2	0	4
1	6	2	3
2	5	0	0
2	5	0	1
3	3	1	7
3	7	6	6
3	4	2	7
3	2	1	5

Table 1: Counts and Single Predictor for Generic Hierarchical Data Set

The data was first analyzed using the %exactlogistic macro for one-stage and two-stage clustered logistic regression for evaluating the hypothesis $H_0: \beta = 0$ vs. $H_a: \beta > 0$. The one-stage model was fit using the macro call:

```
%exactlogistic(data=ASU, levels=1);
```

and the two-stage model was fit using:

```
  %exactlogistic(data=ASU, levels=2);
```

For the one-stage model, the one-sided p-value was 0.0332, while the p-value for the two-stage model was 0.0959. Thus, similarly to Troxler, et al. we see that the test statistic calculated when ignoring the additional level of clustering for group is inflated due to the association present at the group level.

For comparison, the data was also analyzed using the R program developed by Troxler, et al (2011). For the one-stage model, the one-sided p-value was 0.0332, while the p-value for the two-stage model was 0.0959. Thus we can see that the SAS macro provides results consistent with those produced by the R program.

BRITISH SOCIAL SURVEY

We also evaluated the use of the %exactlogistic macro on data from the British Social Attitudes Survey which began in 1983 and concluded in 1986 (McGrath and Waterton 1986). We utilized a subset of this data with measurements from three districts. Each of these districts contain between two and five individuals, each of whom are measured at four different time points. At each time point, the individual is evaluated on whether he or she overall agrees with abortion based on the results of a questionnaire. Thus, the variable district represents the first-stage of clustering while number denotes the count of positive responses for a given individual for the four time points. We also considered the covariate religion, which has can take four values, in which ‘1’ denotes Roman Catholic, ‘2’ denotes Protestant, ‘3’ denotes Other and ‘4’ denotes None. This data are reproduced in Table 2.

District	Number	Count	Religion
1	4	2	2
1	4	1	2
2	4	4	4
2	4	4	4
2	4	0	2
2	4	4	4
2	4	3	2
3	4	4	2
3	4	4	2
3	4	3	3

Table 2. Counts and Single Predictor for British Social Attitudes Survey

The data was analyzed using the %exactlogistic macro for both the one-stage and two-stage clustered approaches for evaluating the hypothesis $H_0: \beta = 0$ vs. $H_a: \beta > 0$. The one-stage model was fit using the macro call:

```
  %exactlogistic(data=SOCATT, levels=1);
```

and the two-stage model was fit using:

```
  %exactlogistic(data=SOCATT, levels=2);
```

For the one-stage model which ignores district, the one-sided p-value for religion was 0.0489, while the p-value for the two-stage model was 0.107. Thus, the test statistic is inflated when the higher levels of clustering are ignored. In this case, there is enough correlation at the district level that the results of our hypothesis test are affected. When the data were analyzed using the R program, we found that the results were in agreement with those obtained using the %exactlogistic macro.

COMPARISON OF RUN TIMES

We also compared the run times of the %exactlogistic macro and the R program for both data examples. Both approaches were run on a Windows machine with a 2.10 GHz Intel Core i7-3687U CPU and 8.00 GB of RAM. For the analysis of the data provided by Troxler, et al. (2011), we saw that the %exactlogistic macro was able to produce a p-value for the one-stage model in 5 seconds, while the two-stage model took 6 seconds. In comparison, the R program took considerably longer, with the one-stage model taking approximately 20 seconds, while the two-stage model took approximately 82 seconds.

In the analysis of the British Social Attitudes Survey discussed by McGrath and Waterton (1986), we found that the %exactlogistic macro took 29 seconds to produce a p-value for the one-stage model, while the two-stage model took about 198 seconds. For fitting the same models in R, we found that the one-stage model took 109 seconds to run, while the two-stage model took 448 seconds. The results of both these analyses are summarized in Table 3.

	Troxler, et al.		McGrath & Waterton	
Approach	**One-stage**	**Two-stage**	**One-stage**	**Two-stage**
SAS	5	6	29	198
R	20	82	109	448

Table 3. Comparison of Run Times (in seconds)

CONCLUSION

Large sample approaches are most commonly used in analyzing binary response data; however in the case of smaller sample sizes or sparse data, it is often necessary to utilize exact approaches. Although a variety of methods are available and well developed for analyzing data using asymptotic theory, approaches which rely on exact methodology are comparatively less investigated. In particular, there is a significant gap in easily accessible implementations for analyzing these data for the case of correlated responses.

We present the %exactlogistic macro in SAS which incorporates the techniques discussed by Troxler, et al. (2011), for testing for significant covariates through the use of hypothesis testing. This macro provides results which are consistent with those produced by a previous R implementation, while requiring significantly less time to complete the analysis. The %exactlogistic macro offers an easy to use implementation for utilizing exact theory for binary responses in the presence of two stages of clustering.

REFERENCES

Corcoran CL, Ryan PS, Mehta C, Patel N, Monenbergs G (2001). An exact trend test for correlated binary data. Biometrika. 57: 941-948.

Cox DR (1970). Analysis of Binary Data. Chapman & Hall: New York.

Have TR, Kunselman AR, Tran L (1999). A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. Statistics in Medicine. 18: 947-960.

Liang K-Y, Zeger SL (1986). Longitudinal data analysis using generalized linear models. Biometrika. 73: 13-22.

McGrath K, Waterton J (1986). British Social Attitudes 1983-1986 panel survey. London, Social and Community Planning Research.

Pepe MS, Anderson GL (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics. 23: 939-951.

Troxler S, Lalonde T, Wilson, JR (2011). Exact logistic models for nested binary data. Statistics in Medicine. 30: 866-876.

Wilson JR, Lorenz KA (2015). Modeling binary correlated responses using SAS, SPSS and R. Springer: New York.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kyle Irimata
Arizona State University
Kyle.irimata@asu.edu

Jeffrey Wilson
Arizona State University
Jeffrey.wilson@asu.edu