韓 國 軍 事 運 營 分 析 學 會 誌

第 25 卷, 第 2 號, 1999. 12. 31.

Application of GLIM to the Binary Categorical Data

Sok, Yong U*

Abstract

This paper is concerned with the application of generalized linear interactive modelling(GLIM) to the binary categorical data. To analyze the categorical data given by a contingency table, finding a good-fitting loglinear model is commonly adopted. In the case of a contingency table with a response variable, we can fit a logit model to find a good-fitting loglinear model. For a given 2^4 contingency table with a binary response variable, we show the process of fitting a loglinear model by fitting a logit model using GLIM and SAS and then we estimate parameters to interpret the nature of associations implied by the model.

* Sejong University

1. Introduction

The science of statistics deals with making decisions based on the observed data in the face of uncertainty. In these days, the study of statistical science has been accelerated by the development of computer science. By joining statistical theory and methodology to the computational capabilities, we could obtain very powerful tools in the analysis of a observed data such as categorical data. In practice, a categorical data set is given by a form of contingency table. The interpretation of a contingency table is based on analyzing associations between factors, and the ways of analyzing associations are found by fitting a good model.

It is well known that a good-fitting loglinear model for two-or three-way contingency table is very useful. But for the higher dimensional case, finding a good-fitting loglinear model has so many difficulties in application because of its complication. In the case of a contingency table with a response variable, we can fit a logit model instead of finding a good-fitting loglinear model. In the sense of dimension reductions, this methodology is very useful especially for the higher dimensional case with a response variable.

In this paper, for a given 2^4 contingency table with a binary response variable, we can show the process of fitting a loglinear model by fitting a logit model using GLIM and SAS, and then we estimate parameters to interpret the nature of associations implied by the model. Sok[8] also has studied the application of GLIM to the analysis of survival data. In Section 2, we discuss the way to find a good-fitting logit model for a high dementional contingency table with a binary response variable. In Section 3, we will find a good-fitting loglinear model for a given four-way contingency table with a binary response variable by fitting a logit model using GLIM and SAS. And finally, concluding remarks will be given in the final section.

2. Fitting a Logit Model

In many cases, the useful models for a given contingency table with a response variable are a small subset of all loglinear models. For instance, in the case that we can choose a response variable from explanatory variables, we know that modelling effects of explanatory variables on the response variable is more important than modelling relationships among these explanatory variables. The explanatory variables in the models may be continuous or categorical. When they are categorical, a subset of loglinear models which is equivalent to the set of all *logit models*(See Cox[3] or McCullagh and Nelder[5]) for the response variable is enough to study to find a good-fitting model. The logit models for associations between the response variable and the explanatory variables contain the same structure as the loglinear models.

For high dimensional case, Bishop[2] and Agresti[1] discussed the dimension reductions when there is a response variable. In the case of the higher dimensional contingency table with a response variable and several explanatory variables, we can find a good-fitting model for the given data set by fitting a logit model instead of a loglinear model directly. After finding a good-fitting logit model, we can choose a loglinear model which corresponds to the logit model. When the response is a binary variable which has two categories, this methodology is of the greatest interest. A good-fitting loglinear model describes effects of explanatory variables, explains associations and interactions between variables, and produces improved estimates of response probabilities.

The logit models generalize when there are several categorical factors. Suppose that there are three categorical factors A, B and C for the binary response. Let I denote the number of levels of A, J the number of levels of B and K the number of levels of C. Denote by M_{ijk1} the probability of response 1 when factor A is at level i, factor B is at level j and factor C is at level k, so the sum of response probabilities, $M_{ijk1} + M_{ijk2} = 1$.

Hence, the complete logit model for a four-way contingency table with a binary response variable can be written as follows :

$$\log\left(\frac{M_{ijk1}}{M_{ijk2}}\right) = \alpha + \beta_i^{A} + \beta_j^{B} + \beta_k^{C} + \beta_{ij}^{AB} + \beta_{ik}^{AC} + \beta_{jk}^{BC} + \beta_{ijk}^{ABC}$$

where $i = 1,\cdots,I,\ \ j = 1,\cdots,J,\ \ k = 1,\cdots,K,$ *and* A, B *and* C are the explanatory variables with I, J, K levels respectively.

3. Application of GLIM and SAS

For simplicity we focus on a four-way contingency table problem with a binary response variable to illustrate for the higher dimensional case. Consider a following contingency table problem in Table 1.

Table 1. A 2^4 Contingency Table Problem.

		Accident Type			
		Collision		Rollover	
		Severity		**Severity**	
Car Weight	**Driver Ejected**	Not Severe	Severe	Not Severe	Severe
Small	No	350	150	60	112
	Yes	26	23	19	80
Standard	No	1878	1022	148	404
	Yes	111	161	22	265

We have four categorical variables such as

S = Severity of Accident ("Severe" or "Not Severe")

A = Accident Type ("Collision" or "Rollover"),

C = Car Weight ("Small" or "Standard") and

D = Driver Ejected ("No" or "Yes").

Since the variable S(Severity of Accident) may be influenced by the other variables(A, C or D), it is reasonable to consider S, which is a binary variable, as the response variable and the others as explanatory variables. So we will fit a logit model to our data with 4831 observations.

3.1 Application of GLIM

To find a good-fitting model, we can obtain the following Table 2 by the use of GLIM(See Healy[4] and Murray[6]).

From the Table 2, it is concluded that the model having the main effects only fit well among all possible models. To find significant interactions we can compute the scaled deviance, G^2, which is reduced by adding one interaction as follows :

$G^2(A+C+D \mid A+C+D+AC) = 7.31 - 5.53 = 1.78$,

$G^2(A+C+D \mid A+C+D+AD) = 7.31 - 5.53 = 1.78$ and

$G^2(A+C+D \mid A+C+D+CD) = 7.31 - 4.71 = 2.60$.

Thus the interactions are not significant under $\alpha = 0.05$. Therefore, we will choose

$$\log\left(\frac{M_{ijk1}}{M_{ijk2}}\right) = \alpha + \beta_i^A + \beta_j^C + \beta_k^D \quad \text{---------------------} \quad (3.1)$$

as the best model. The corresponding loglinear model is given by

$$\log(M_{ijkh}) = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^D + \lambda_h^S + \lambda_{ih}^{AS} + \lambda_{jh}^{CS} + \lambda_{kh}^{DS}$$
$$+ \lambda_{ij}^{AC} + \lambda_{ik}^{AD} + \lambda_{jk}^{CD} + \lambda_{ijk}^{ACD},$$

where $i=1,2,\ j=1,2,\ k=1,2,$ and $h=1,2$ ------------------- (3.2)

which is denoted by (AS, CS, DS, ACD).

Since we assume linear constraint on the parameters

$$\beta_1^A = \beta_1^C = \beta_1^D = 0$$

in GLIM coding, note that

$$\alpha = 2\lambda_2^S, \quad \beta_i^A = 2\lambda_{i2}^{AS},$$

$$\beta_j^C = 2\lambda_{j2}^{CS}, \quad \text{and} \quad \beta_k^D = 2\lambda_{k2}^{DS}.$$

Then the following estimates and the standard errors(s.e.) can be obtained from GLIM output in the following Table 3.

$\hat{\alpha} = -0.94$, s.e.($\hat{\alpha}$) = 0.08,

$\hat{\beta}_2^A = 1.64$, s.e.($\hat{\beta}_2^A$) = 0.08,

$\hat{\beta}_2^C = 0.34$, s.e.($\hat{\beta}_2^C$) = 0.09,

$\hat{\beta}_2^D = 1.03$, and s.e.($\hat{\beta}_2^D$) = 0.10.

Table 2. GLIM Output for All Possible Models

Model	G^2	df	p_value
None	737.89	7	0.00
A	136.47	6	0.00
C	737.14	6	0.00
D	451.90	6	0.00
A+D	22.84	5	0.00
A+C	122.10	5	0.00
C+D	449.15	5	0.00
A+C+AC	120.67	4	0.00
A+D+AD	21.12	4	0.00
C+D+CD	448.00	4	0.00
A+C+D	**7.31**	**4**	**0.12**
A+C+D+AC	5.53	3	0.14
A+C+D+AD	5.53	3	0.14
A+C+D+CD	4.71	3	0.20
A+C+D+AC+AD	3.59	2	0.17
A+C+D+AC+CD	3.61	2	0.17
A+C+D+AD+CD	1.89	2	0.40
A+C+D+AC+AD+CD	0.67	1	0.41

The fact that the estimated average logit α is negative indicates an overall tendency in Severity of Accident to be less "Severe" outcome than "Not Severe" outcome. Classification of Accident Type as "Rollover" has a positive effect on the logit for Severity of Accident. Thus it is estimated that "Rollover" of Accident Type have a larger chance to be "Severe" outcome on Severity of Accident than "Collision" of Accident Type, controlling Car Weight and Driver Ejected. The effect on the logit is positive when Car Weight is classified "Standard" when Accident Type and Driver Ejected are controlled. And the effect on logit is positive when Driver Ejected is classified "Yes", controlling Accident Type and Car Weight.

The estimated A-S, C-S and D-S partial odd's ratios are given as follows :

$e^{-\beta_2^A} = e^{-1.64} = 0.19,$

$e^{-\beta_2^C} = e^{-0.34} = 0.71$ and

$e^{-\beta_2^D} = e^{-1.03} = 0.36$ respectively.

A partial GLIM program and its output is shown in the followings and also the complete GLIM program and the corresponding output can be accessed from the author.

Table 3. A Partial Program Listing and Output Using GLIM

```
[i] ? $unit 8
[i] ? $factor a 2 c 2 d 2
[i] FAC? $data a c d s n
[i] DAT? $read
[i] REA? 1 1 1  150   500 2 1 1 112 172
[i] REA? 1 1 2   23    49 2 1 2  80  99
[i] REA? 1 2 1 1022  2900 2 2 1 404 552
[i] REA? 1 2 2  161   272 2 2 2 265 287
[i] ? $yvar s
[i] ? $err b n
[i] ? $fit : +a $
```

```
[o] scaled deviance = 737.89 at cycle 3
[o]             d. f.  =   7
[o] scaled deviance = 136.47(change = -601.4) at cycle 3
[o]             d. f.  =   6   (change =    -1 )
[i] ? $dis e r $
[o]             estimate          s.e.         parameter
[o]   1         -0.5562          0.03406          1
[o]   2           1.797          0.07955          A(2)
[o]   scale parameter 1.000
[o] unit    observed    out of      fitted        residual
[o]  1         150        500       182.21        -2.993
[o]  2         112        172       133.42        -3.915
[o]  3          23         49        17.86         1.527
[o]  4          80         99        76.79         0.773
[o]  5        1022       2900      1056.81        -1.343
[o]  6         404        552       428.17        -2.466
[o]  7         161        272        99.12         7.796
[o]  8         265        287       222.62         5.997
[i] ? $fit a+c+d $
[o] scaled deviance = 7.3090 at cycle 3
[o]             d. f.  =   4
[i] ? $dis e r $
[o]             estimate          s.e.         parameter
[o]   1         -0.9401          0.08284          1
[o]   2           1.639          0.08281          A(2)
[o]   3          0.3367          0.08612          C(2)
[o]   4           1.030          0.09891          D(2)
[o]   scale parameter 1.000
.
.
.
```

3.2 Application of SAS

Using the procedure "PROC CATMOD;" in SAS(See SAS/STAT User's Guide[7]), we can also find a good-fitting logit model (3.1) to the given data set. Note that, here in SAS coding, we assume that a linear constraint on the parameters,

$$\sum_i \beta_i^A = \sum_j \beta_j^C = \sum_k \beta_k^D = 0 \text{ instead of } \beta_1^A = \beta_1^C = \beta_1^D = 0$$

in GLIM coding. Then we get the loglinear model (3.2) with

$$\alpha = \lambda_2^S, \quad \beta_i^A = \lambda_{i2}^{AS},$$

$$\beta_j^C = \lambda_{j2}^{CS}, \text{ and } \quad \beta_k^D = \lambda_{k2}^{DS} .$$

Then the computational results are obtained from SAS output in the following Table 4. From this table, we have obtained the fact that the likelihood ratio chi-square statistic value, which is equal to 7.31 in SAS output, is the same as the deviance, G^2 in GLIM output for the main effect model.

And the complete SAS program and the corresponding output for three models such as Model H = I J K, Model H = I J K I*K and Model H = I J K I*J J*K I*K. will be given in Appendix A.

$\hat{\alpha}$ = −0.56, s.e.($\hat{\alpha}$) = 0.06,

$\hat{\beta}_2^A$ = 0.82 = − $\hat{\beta}_1^A$, s.e.($\hat{\beta}_2^A$) = 0.04,

$\hat{\beta}_2^C$ = 0.17 = − $\hat{\beta}_1^C$, s.e.($\hat{\beta}_2^C$) = 0.04

$\hat{\beta}_2^D$ = 0.515 = − $\hat{\beta}_1^D$, and s.e.($\hat{\beta}_2^D$) = 0.05.

Table 4. A Partial SAS Output for the Main Effect Model

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	93.28	0.0000
I	1	15.28	0.0001
J	1	391.54	0.0000
K	1	108.51	0.0000
LIKELIHOOD RATIO	4	7.31	0.1204

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

				Standard	Chi-
Effect	Parameter	Estimate	Error	Square	Prob
INTERCEPT	1	-0.5628	0.0583	93.28	0.0000
I	2	0.1683	0.0431	15.28	0.0001
J	3	0.8193	0.0414	391.54	0.0000
K	4	0.5152	0.0495	108.51	0.0000

4. Concluding Remarks

We have shown the process of fitting a logit model to find a good-fitting loglinear model. A four-way cross-classification of variables A, C, D and S were considered to illustrate models for the higher dimensions. We have fitted a model for the given 2^4 contingency table problem using GLIM and also the same model has been fitted for the given problem using SAS in section 3.1 and 3.2 respectively. Consequently, we can fit the same model and the same interpretation to our data set in either case of applying GLIM or applying SAS.

Hence, we found the same good-fitting loglinear model for the given data set by fitting the best logit model. In summary, it can be concluded that the best logit model contains only the main effect terms for the factors, but no interaction terms. That is, the model can be suggested for our data set as the following logit model

$$\log\left(\frac{M_{ijk1}}{M_{ijk2}}\right) = \alpha + \beta_i^A + \beta_j^C + \beta_k^D,$$

where the scaled deviance, G^2 = 7.31 with degree of freedom 4.

Therefore, we can say that "Rollover" of Accident Type makes more likely "Severe" of Severity of Accident, controlling other two factors, and so do "Standard" of Car Weight and "Yes" of Driver Ejected.

In theortical sense, this methodology could be extended to the higher (more than four) dimensional cases, but it could not be applicable to practical situations because of the dramatic increase in the number of cells and the complication caused by possible interaction patterns.

Appenxix A:The SAS Program and the Edited Output for Three Models

```
/* LOGIT MODEL USING SAS */
DATA ONE;
 DO I=1 TO 2;
   DO K=1 TO 2;
     DO J=1 TO 2;
       DO H=1 TO 2;
INPUT COUNT @@; OUTPUT;
 END; END; END; END;
CARDS;
      350  150   60  112
       26   23   19   80
     1878 1022  148  404
      111  161   22  265
      ;
PROC CATMOD;
 WEIGHT COUNT;
MODEL H = I J K / ML NOGLS;
RUN;
MODEL H = I J K I*K / ML NOGLS;
RUN;
MODEL H = I J K I*J J*K I*K /ML NOGLS;
RUN;
```

MAXIMUM LIKELIHOOD ANALYSIS

teration	Sub Iteration	-2 Log Likelihood	Convergence Criterion
0	0	6697.1881	1.0000
1	0	5945.462	0.1122
2	0	5933.9881	0.001930
3	0	5933.9422	7.7374E-6
4	0	5933.9422	1.685E-10

Parameter Estimates

Iteration	1	2	3	4
0	0	0	0	0
1	-0.4307	0.1385	0.7274	0.4148
2	-0.5532	0.1668	0.8139	0.5086
3	-0.5627	0.1683	0.8193	0.5151
4	-0.5628	0.1683	0.8193	0.5152

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	91.94	0.0000
I	1	17.12	0.0000
J	1	267.30	0.0000
K	1	108.86	0.0000
I*K	1	1.80	0.1799
LIKELIHOOD RATIO	3	5.53	0.1371

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-0.5559	0.0580	91.94	0.0000
I	2	0.1971	0.0476	17.12	0.0000
J	3	0.7861	0.0481	267.30	0.0000
K	4	0.5161	0.0495	108.86	0.0000
I*K	5	-0.0639	0.0476	1.80	0.1799

CATMOD PROCEDURE

Response: H　　Response Levels (R)= 2
Weight Variable: COUNT　　Populations (S)= 8
Data Set: ONE　　Total Frequency (N)= 4831
Observations (Obs)= 16

POPULATION and RESPONSE PROFILES

Sample	I	J	K	Sample Size
1	1	1	1	500
2	1	1	2	49
3	1	2	1	172
4	1	2	2	99
5	2	1	1	2900
6	2	1	2	272
7	2	2	1	552
8	2	2	2	287

Response H

1	1
2	2

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	93.28	0.0000
I	1	15.28	0.0001
J	1	391.54	0.0000
K	1	108.51	0.0000
LIKELIHOOD RATIO	4	7.31	0.1204

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-0.5628	0.0583	93.28	0.0000
I	2	0.1683	0.0431	15.28	0.0001
J	3	0.8193	0.0414	391.54	0.0000
K	4	0.5152	0.0495	108.51	0.0000

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	72.76	0.0000
I	1	18.22	0.0000
J	1	194.35	0.0000
K	1	60.94	0.0000
I*J	1	1.23	0.2680
J*k	1	2.85	0.0912
I*K	1	2.93	0.0872
LIKELIHOOD RATIO	1	0.67	0.4134

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-0.5323	0.0624	72.76	0.0000
I	2	0.2718	0.0637	18.22	0.0000
J	3	0.8577	0.0615	194.35	0.0000
K	4	0.4925	0.0631	60.94	0.0000
I*J	5	-0.0537	0.0485	1.23	0.2680
J*K	6	-0.0964	0.0571	2.85	0.0912
I*K	7	-0.1076	0.0629	2.93	0.0872

REFERENCES

[1] Agresti, A., *Categorical Data Analysis,* Wiley, N. Y., 1990.

[2] Bishop, Y. V. V., Effects of collapsing multidimensional contingency table, *Biometrics* Vol. 25, pp. 119 - 128, 1971.

[3] Cox, D. R., The *Analysis of Binary data,* Chapman and Hall, London, 1970.

[4] Healy, M. J. R., *GLIM : An Introduction*, Clarendon Press, Oxford, 1988.

[5] McCullagh, P. and Nelder, J. A., *Generalized Linear models,* 2nd ed., Chapman and Hall, London, 1989.

[6] Murray, A. et el., Statistical Modelling in *GLIM* Clarendon Press, Oxford, 1989.

[7] *SAS/STAT user's Guide*, version 6, 4th ed., Cary, N.C., SAS Institute Inc., 1994.

[8] Sok, Y. U., Application of GLIM to Survival Data, *Journal of the Military Operations Research Society of Korea* Vol. 21, No. 2. pp.102-115, 1995.

[99년 8월 20일 접수, 99년 11월 3일 최종수정]