Paper 2641-2015

A New Method of Using Polytomous Independent Variables with Many Levels for the Binary Outcome of Big Data Analysis

John Gao, ConstantContact;

Jesse Harriott, ConstantContact;

Lisa Pimentel, ConstantContact;

ABSTRACT

The paper discusses a new method in logistic regression for polytomous independent variables with many levels in big data. In the proposed method, the first step is to conduct an iterative statistical analysis from a SAS Macro program to search for the proper aggregation groups with statistically significant differences from all levels in a polytomous independent variable. Then the mean values at the new aggregation groups are used for a new independent variable, which is a numerical, can be regarded as a continuous and is going to replace the original variable in the following statistical analysis.

INTRODUCTION

In the big data, often variables are polytomous with many levels. The common method to deal with polytomous independent variable is to use a series of design variables which correspond to the option: "class" for the polytomous independent variable in SAS procedure : "proc logistic" if the outcome is binary[1][2]. This would potentially require thousands of design variables, increasing computation time with little help on the prediction of outcome should the data have many polytomous independent variables with many levels. This paper presents a new method for logistic regression with polytomous independent variables when analysis of big data is required.

In the proposed method, the first step is to conduct an iteration statistical analysis from a SAS Macro program similar to the algorithm in the creation of spline variables[3]. This analysis is to search for the proper aggregation groups with a statistically significant difference from all levels in a polytomous independent variable. Then the mean values at the new aggregation groups are used for a new independent variable, which is going to be a numerical and continuous independent variable and replaces the original variable in the following statistical analysis.

We have used the new method in predictive models on the polytomous independent variables, such as industry types and geographic locations. Both of them are significant drivers and the result is better and much simpler using the new method compared to using a "series of design variables".

THE CURRENT APPROACH FOR POLYTOMOUS INDEPENDENT VARIABLES

Usually, a polytomous variable is converted into a numerical variable by using a series of design variables. Then the new numerical variable replaces the original polytomous variable for the statistical analysis. For example, the polytomous variable X has N values: X_1,…., X_n. The design variable for X is going to be defined as

X_j <- a_j = (0,..,0,1,0,…,0) j=1,…,n-1

X_n <- a_n = (-1,…,-1) OR a_n = (0,…,0)

here a_n is called as the reference level.

In the SAS procedure of proc logistic, these design variables are going to replace the original polytomous variable from "class" or "by", such as

```
Proc logistic;
    Class X;
```

Model DV=X iv_1,…,iv_m ;

Run;

The output result from the program will be like

Class Level Information

Class	Value	Design Variables
X	a_1	1 0 0, …,0
	a_2	0 1 0, …,0
……		
	a_n	0 0 0, …,0

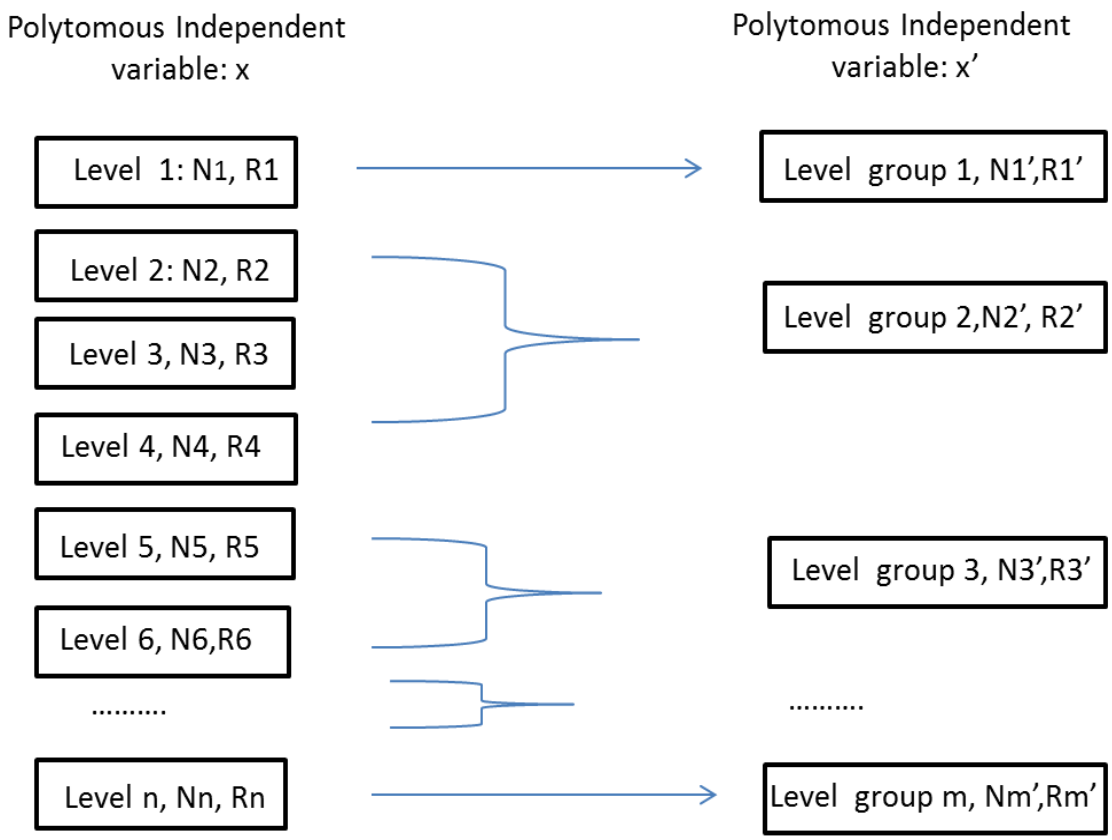
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.6872	1.3697	31.4984	<.0001
iv_1		1	0.1438	0.0236	37.0981	<.0001
……						
X	a_1	1	-0.9204	0.4897	3.5328	0.0602
	a_2	1	-0.3839	0.3975	0.9330	0.3341
……						
	a_{n-1}	1	-1.1083	0.1132	16.1923	<0.0001

in the approach, you have to keep the result of all levels in the polytomous variable even though some of them are not significant. Usually, big data sets have many polytomous variables, and each of them has many levels. This approach would potentially require thousands of design variables and increase the complexity in both computation and the prediction of outcome.

A NEW METHOD OF USING POLYTOMOUS INDEPENDENT VARIABLES FOR BINARY OUTCOME IN BIG DATA

In the proposed method for big data, there are 2 steps. The first step is to conduct an iteration statistical analysis to search for the proper aggregation groups with a statistically significant difference from all levels in a polytomous independent variable. In the second step, the original polytomous variable is replaced by a piecewise variable. The each value in the piecewise variable is the average rate of the binary outcome of each aggregated group derived from all levels in the step 1.



here Ri are the outcome means at level I (i=1,…n) with

$$R1<=R2<=R3<=R4<=R5<=R6<=\ldots<=Rn$$

And Ri'(i=1,…,m) are the outcome means at group level j after regrouping all levels into new level groups (with statistical significant difference on the expectation) gotten from a search processing

$$R1'<R2'<R3'<\ldots.<Rm'$$

and

$$X' = R_1' \quad \text{if } x\epsilon \text{ Level Group 1}$$
$$= R_2' \quad \text{if } x\epsilon \text{ Level Group 2}$$
$$\ldots\ldots\ldots$$
$$= R_m' \quad \text{if } x\epsilon \text{ Level Group m}$$

The new defined independent variable X' is monotonically variable and can be regarded as a continuous variables for big data. In the SAS Macro program, an iteration processing of searching new level groups with statistical significant differences has been developed. The first is from the level 1 with the smallest value of the outcome means. Then we can conduct statistical test for the level 1 group with level 2 group with second smallest value of the outcome mean. If these 2 groups have a statistically significant difference, we can start to test level 2 with level 3 groups. If level 1 and level 2 do not have statistical significant difference, we can combine them together into new level group 1. Then we are going to test new level group 1 with level 3. The processing will be repeated until all levels have been tested.

Then we can replace the original level values of the polytomous variable by the new level values with a statistically significant difference and also can be described by these means of all new levels because of the "1 to 1" equivalence relationship of a piecewise function in logit from polytomous's levels to outcome means. It is very easily to approve that the conditional mean of at outcome y given a polytomous variable x is very good approximation based on a maximum likelihood analysis. Therefore, it can be directly used in model development. It is noted that the coefficient for the new independent variable X' is always positive due to the monotonic nature of the variable.

Compared with design variables, the new piecewise variable based on the information of all levels as a single independent variable can capture the impact of all levels with much simpler way. We have used this method in predictive models on the polytomous variables: state, business type and customer claim type and etc. All of these polytomous variables have significant improvement in the prediction of binary outcome than without using them or without using design variables in the model development. The appendix shows the comparison of 2 methods for the implementation of the polytomous variables: state in a predictive model.

REFERENCES

[1]. Hosmer, David W.; Lemeshow, Stanley (2000). *Applied Logistic Regression*. New York: Wiley. ISBN 0-471-61553-6.

[2] Paul D. Allison, (March 2012) *Logistic Regression Using SAS®: Theory and Application*, Second Edition : SAS Institute

[3]Jian Gao (2005*) An Optimal Spline Logistic Regression Method* in International conference of statistics". Hawaii, Jan 9-Jan 11 2005.

APPENDIX

1. SAS output by using class state

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-16.3131	7.9313	4.2304	0.0397
D1		1	3.5956	0.9707	13.7206	0.0002
D2		1	15.7928	4.3573	13.1364	0.0003
D3		1	6.1159	0.7777	61.8473	<.0001
D4		1	5.7765	1.8707	9.5346	0.002
D5		1	6.8723	0.3326	426.9956	<.0001
D6		1	4.0675	0.2152	357.2887	<.0001
D7		1	2.6533	0.4792	30.6613	<.0001
D8		1	7.74	0.5588	191.8524	<.0001
D9		1	3.5059	0.3387	107.1386	<.0001
D10		1	2.539	0.6677	14.4578	0.0001
D11		1	2.4802	0.3269	57.5539	<.0001
D12		1	2.8351	0.5774	24.1135	<.0001
D13		1	2.5392	0.6536	15.0943	0.0001
D14		1	3.1401	0.6314	24.7345	<.0001

STATE	**AA**	1	12.1094	322.1	0.0014	0.97
STATE	**AB**	1	0.4846	7.8619	0.0038	0.9508
STATE	**AE**	1	-8.4431	210.5	0.0016	0.968
STATE	**AK**	1	-0.0878	7.869	0.0001	0.9911
STATE	**AL**	1	0.4453	7.8613	0.0032	0.9548
STATE	**AR**	1	0.2071	7.8641	0.0007	0.979
STATE	**AZ**	1	0.6436	7.8602	0.0067	0.9347
STATE	**BC**	1	0.272	7.8618	0.0012	0.9724
STATE	**CA**	1	0.3865	7.8595	0.0024	0.9608
STATE	**CO**	1	0.1864	7.8602	0.0006	0.9811
STATE	**CT**	1	0.4981	7.8604	0.004	0.9495
STATE	**DC**	1	0.5715	7.8625	0.0053	0.9421
STATE	**DE**	1	0.3911	7.8657	0.0025	0.9603
STATE	**FL**	1	0.5829	7.8596	0.0055	0.9409
STATE	**GA**	1	0.1939	7.86	0.0006	0.9803
STATE	**HI**	1	0.2448	7.8653	0.001	0.9752
STATE	**IA**	1	0.1574	7.8636	0.0004	0.984
STATE	**ID**	1	0.7486	7.8642	0.0091	0.9242
STATE	**IL**	1	0.5117	7.8598	0.0042	0.9481
STATE	**IN**	1	0.589	7.8608	0.0056	0.9403
STATE	**KS**	1	0.6014	7.8618	0.0059	0.939
STATE	**KY**	1	0.6446	7.8624	0.0067	0.9347
STATE	**LA**	1	0.5915	7.8609	0.0057	0.94
STATE	**MA**	1	0.535	7.8598	0.0046	0.9457
STATE	**MB**	1	1.2073	7.873	0.0235	0.8781
STATE	**MD**	1	0.2934	7.8601	0.0014	0.9702
STATE	**ME**	1	0.5421	7.8637	0.0048	0.945
STATE	**MI**	1	0.5849	7.8601	0.0055	0.9407
STATE	**MN**	1	0.8103	7.8603	0.0106	0.9179
STATE	**MO**	1	0.8207	7.8607	0.0109	0.9168
STATE	**MS**	1	0.1966	7.8643	0.0006	0.9801
STATE	**MT**	1	0.7019	7.8685	0.008	0.9289
STATE	**NB**	1	0.2799	7.9378	0.0012	0.9719
STATE	**NC**	1	0.5311	7.8601	0.0046	0.9461
STATE	**ND**	1	1.8263	7.8719	0.0538	0.8165
STATE	**NE**	1	0.6379	7.8629	0.0066	0.9353
STATE	**NH**	1	0.9669	7.8614	0.0151	0.9021
STATE	**NJ**	1	0.255	7.8599	0.0011	0.9741
STATE	**NL**	1	-8.18	118.9	0.0047	0.9451
STATE	**NM**	1	0.8421	7.8623	0.0115	0.9147
STATE	**NS**	1	0.1363	7.8882	0.0003	0.9862

STATE	**NT**	1	-8.8326	102.4	0.0074	0.9313
STATE	**NU**	1	-9.3961	210.5	0.002	0.9644
STATE	**NV**	1	0.725	7.8609	0.0085	0.9265
STATE	**NY**	1	0.593	7.8596	0.0057	0.9399
STATE	**OH**	1	0.6143	7.86	0.0061	0.9377
STATE	**OK**	1	0.55	7.8623	0.0049	0.9442
STATE	**ON**	1	0.6149	7.86	0.0061	0.9376
STATE	**OR**	1	0.1639	7.8611	0.0004	0.9834
STATE	**PA**	1	0.681	7.8598	0.0075	0.931
STATE	**PE**	1	-8.7044	210.5	0.0017	0.967
STATE	**QC**	1	0.4097	7.8703	0.0027	0.9585
STATE	**RI**	1	0.3199	7.8633	0.0017	0.9676
STATE	**SC**	1	0.4871	7.8609	0.0038	0.9506
STATE	**SD**	1	0.9869	7.8718	0.0157	0.9002
STATE	**SK**	1	0.2469	7.8725	0.001	0.975
STATE	**TN**	1	0.3094	7.8609	0.0015	0.9686
STATE	**TX**	1	0.417	7.8596	0.0028	0.9577
STATE	**UT**	1	0.5292	7.8624	0.0045	0.9463
STATE	**VA**	1	0.5867	7.86	0.0056	0.9405
STATE	**VT**	1	0.8175	7.864	0.0108	0.9172
STATE	**WA**	1	0.507	7.8603	0.0042	0.9486
STATE	**WI**	1	0.5716	7.8607	0.0053	0.942
STATE	**WV**	1	-0.4423	7.8823	0.0031	0.9552
STATE	**WY**	1	1.1055	7.8686	0.0197	0.8883
STATE	**WA**	1	0.3903	12.1946	0.001	0.9745
STATE	**WI**	1	0.5691	12.1947	0.0022	0.9628
STATE	**WV**	1	-0.0301	12.1997	0	0.998
STATE	**WY**	1	0.9065	12.1989	0.0055	0.9408

2. SAS output by using the new approach.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.8508	1.0588	224.1324	<.0001
D1	1	3.5843	0.9655	13.7814	0.0002
D2	1	14.8121	4.3359	11.6703	0.0006
D3	1	6.0666	0.7746	61.3303	<.0001
D4	1	5.4104	1.8611	8.4514	0.0036
D5	1	6.9033	0.3313	434.3062	<.0001
D6	1	4.028	0.214	354.2839	<.0001
D7	1	1.9025	0.4627	16.9034	<.0001
D8	1	7.8272	0.5571	197.3952	<.0001
D9	1	3.564	0.3365	112.1831	<.0001
D10	1	2.5464	0.6653	14.6476	0.0001
D11	1	2.2841	0.3235	49.8523	<.0001
D12	1	2.7966	0.5737	23.7628	<.0001
D13	1	2.528	0.6472	15.2582	<.0001
D14	1	2.7046	0.6204	19.0073	<.0001
ST	1	2.9314	0.5391	29.5644	<.0001