Paper 10220-2016

Correcting the Quasi-complete Separation Issue in Logistic Regression Models

Xinghe Lu, AmeriHealth Caritas Family of Companies, Philadelphia, PA

ABSTRACT

Quasi-complete separation is a commonly detected issue in logit/probit models. Quasi-complete separation occurs when the dependent variable separates an independent variable or a combination of several independent variables to a certain degree. In other words, levels in a categorical variable or values in numeric variable are separated by groups in a discrete outcome variable. Most of the time, it happens in categorical independent variable(s). Quasi-complete separation can cause convergence failures in the model, which consequently result in inflated coefficient estimates and standard errors. Therefore it potentially yields biased results.

This paper reviews various approaches for correcting Quasi-complete separation using a binary logistic regression model as the example. First it introduces the concept of Quasi–complete separation and how to diagnose the issue. Then it provides step-by-step guidelines for fixing the problem from straightforward data configuration approach to complicated statistical modeling approach, such as EXACT method and FIRTH method.

Keywords: Quasi-complete separation, logistic regression, Greenacre's method, FIRTH method and cluster analysis.

INTRODUCTION

Logistic regression is a statistical method used to measure the relationship between a dichotomous outcome variable and one or more independent variables. It is also called a logit model, because the log odds of the outcome variable is modeled as a linear combination of the predictor variables. Logistic regression is the special case of generalized linear model (GLM) with a logit function. Quasi-complete separation is a common issue in regression analysis when outcome variable is categorical variable, including logistic regression and probit regression. This paper describes several remedies ranging from simple data configuration to relatively complex statistical modeling.

WHAT IS QUASI-COMPELET SEPARATION

Quasi-complete separation occurs when the dependent variable separates an independent variable or a combination of several independent variables to a certain degree. In other words, at least one group of the dependent variable has zero frequency for at least one category of an independent variable. It can happen with continuous variables as well. However, it most often occurs with categorical variables. Here is an example.

EXAMPLE DATA

The following example is based on a randomly generated artificial dataset called DEMO and it does not refer to any business/company. The dataset (N=1,500) consists one ID variable, one dependent variable and three independent variables.

- Y: is the binary dependent variable (0 or 1)
- X1: is the continuous variable
- X2: is the nominal discrete variable (1-15)
- X3: is the ordinal discrete variable (1-8)

Table 1 shows the first 10 observations of the DEMO dataset.

ID	Y	X1	X2	X3
1	0	-0.043	2	8
2	0	0.648	2	3
3	0	-0.222	5	3
4	0	-0.731	14	6
5	0	-1.521	12	4
6	0	-0.430	11	6
7	0	0.176	5	1
8	0	-0.626	3	2
9	0	-0.644	15	3
10	0	2.387	2	6

DETECTION QUASI-COMPELET SEPARATION

First, run logistic regression model with PROC LOGISTIC procedure.

```
proc logistic data=demo desc;
   class x2 x3;
   model y = x1 x2 x3;
run;
```

Some warning messages from LOG indicate that the model is not adequate and questionable.

```
WARNING: There is possibly a quasi-complete separation of data points.
The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above
warning. Results shown are based on the last maximum likelihood
iteration. Validity of the model fit is questionable.
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.9729	27.8511	0.0319	0.8583
x1		1	-0.2316	0.1048	4.8794	0.0272
x2	1	1	1.3249	16.1333	0.0067	0.9346
x2	2	1	1.0359	16.1352	0.0041	0.9488
x2	3	1	0.5199	16.1350	0.0010	0.9743
x2	4	1	0.3724	16.1380	0.0005	0.9816
x2	5	1	0.5070	16.1366	0.0010	0.9749
x2	6	1	0.8538	16.1346	0.0028	0.9578
x2	7	1	1.1387	16.1342	0.0050	0.9437
x2	8	1	0.4046	16.1366	0.0006	0.9800
x2	9	1	0.8057	16.1351	0.0025	0.9602
x2	10	1	1.2411	16.1347	0.0059	0.9387
x2	11	1	1.3379	16.1337	0.0069	0.9339
x2	12	1	-12.0529	225.8	0.0028	0.9574
x2	13	1	0.8858	16.1351	0.0030	0.9562
x2	14	1	1.0258	16.1342	0.0040	0.9493
x3	1	1	1.7404	22.7056	0.0059	0.9389
x3	2	1	1.5500	22.7057	0.0047	0.9456
x3	3	1	2.1626	22.7053	0.0091	0.9241
x3	4	1	1.2059	22.7062	0.0028	0.9576
x3	5	1	1.7343	22.7055	0.0058	0.9391
x3	6	1	-11.4435	158.9	0.0052	0.9426
x3	7	1	1.7112	22.7057	0.0057	0.9399

From the estimate table, we can see that standard errors are extremely large when x2=12 and X3=6 compared to other levels in x2 and X3 variables. And the magnitudes of point estimate of these two levels are also very large. All of these abnormal signs convince us that the estimates of coefficients and standard errors from the model are not reliable. We can validate this assumption easily by running 2 by 2 frequency tables shown as below,

		X2														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	0	106	82	122	92	89	102	87	96	92	76	81	96	85	97	91
	1	13	7	7	4	5	8	9	5	7	8	11	0	7	9	6

		X3							
		1	2	3	4	5	6	7	8
Y	0	171	173	165	178	188	177	169	173
	1	16	14	23	10	17	0	15	11

When X2=12 or X3=6, the cell frequency of Y=1 are all 0. In a 2 by 2 table, at least one cell's frequency is zero, which is a typical example of a quasi-complete separation. A quasi-complete separation can cause infinite parameter estimation in the model and failure of convergence of maximum likelihood estimation (MLE).

FIXING QUASI-COMPELET SEPARATION

COLLECTING MORE DATA

Most of the time quasi-complete separation is associated with small sample sizes. Recruiting more samples in the development data could be a possible remedy by filling the empty cells. However, it is not practical because additional data may not be available.

EXCLUDING CASES

Excluding the level(s) causing the issue and applying the rest of data into the model is another way to fix quasi-complete separation.

```
proc logistic data=Demo desc;
   class x2 x3;
   model y = x1 x2 x3;
   where x2 not in (12) and x3 not in (6);
run;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4772	0.1122	487.6614	<.0001
x1		1	-0.2316	0.1048	4.8794	0.0272
x2	1	1	0.4639	0.2989	2.4097	0.1206
x2	2	1	0.1749	0.3862	0.2051	0.6506
x2	3	1	-0.3411	0.3789	0.8103	0.3680
x2	4	1	-0.4884	0.4900	0.9936	0.3189
x2	5	1	-0.3539	0.4423	0.6400	0.4237
x2	6	1	-0.00711	0.3607	0.0004	0.9843
x2	7	1	0.2778	0.3452	0.6476	0.4210
x2	8	1	-0.4563	0.4411	1.0697	0.3010
x2	9	1	-0.0552	0.3836	0.0207	0.8855
x2	10	1	0.3802	0.3678	1.0683	0.3013
x2	11	1	0.4770	0.3208	2.2114	0.1370

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
x2	13	1	0.0248	0.3846	0.0042	0.9485
x2	14	1	0.1648	0.3444	0.2290	0.6323
x3	1	1	0.1056	0.2494	0.1792	0.6721
x3	2	1	-0.0848	0.2606	0.1059	0.7448
x3	3	1	0.5278	0.2191	5.8051	0.0160
x3	4	1	-0.4288	0.2974	2.0789	0.1493
x3	5	1	0.0995	0.2420	0.1690	0.6810
x3	7	1	0.0765	0.2551	0.0899	0.7644

After applying WHERE statement, there is no more warning message and all statistical inferences are in the normal range.

COLLAPSING THE LEVELS

When categorical variable with quasi-complete is ordinal, one might think of redefining that categorical variable by collapsing the levels based on the order in that variable. For example, there is an ordinal categorical variable called age group, which contains five levels (0-18, 19-30, 30-50, 50-60 and 60+). Let's suppose all Y are equal to 1 when age group=30-50. In other words, the frequency cell of Y=0 and age group=30-50 is 0. As we already knew from prior example, quasi-complete will occur and model will fail to converge. What you can do is to combine age group of 30-50 with group 50-60 together and create a new age group, 30-60. You also may combine age group 30-50 with 19-30. In the demo example, X3 is the ordinal categorical independent variable. So level 6 of X3 can be combined with either level 5 or 7. In the next example, level 6 was combined with level 7.

```
Data demo2;
   set demo;
   if x3 = 6 then x3 = 7;
run;

proc logistic data=demo2 desc;
   class x3;
   model y = x1 x3;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6087	0.1071	593.0077	<.0001
x1	1	-0.2337	0.1025	5.1984	0.0226
x3 1	1	0.2137	0.2448	0.7624	0.3826
x3 2	1	0.0517	0.2574	0.0404	0.8407
x3 3	1	0.6171	0.2156	8.1913	0.0042
x3 4	1	-0.3264	0.2944	1.2295	0.2675
x3 5	1	0.1903	0.2387	0.6359	0.4252
x3 7	1	-0.5617	0.2463	5.1998	0.0226

The demo data has been reconstructed and the result showed the quasi-complete separation issue has been solved.

So far, I have introduced multiple ways of fixing separation issue in non-statistical approaches. In the next section, several statistical approaches will be introduced.

GREENACRE'S METHOD

Greenacre's method is a data-driven method to collapse levels in the categorical variable. It is a repeatable iteration process, which means, in each iteration, the Chi-square statistics from all possible

combinations of two levels are produced and compared. The levels giving the least difference in Chi-square statistic will be merged together. This method is a fast and efficient way to reduce the dimension of the categorical input since only univariate association was considered.

In the sample dataset DEMO, independent variable X2 has 15 levels. First, frequencies and the proportions of event by each level in X2 were calculated. For example, X2=1, FREQ=119 and PROP=0.109, means there are total 119 cases when X2=1 and among them, about 10.9% (13/119) cases have the events(Y=1).

x2	FREQ	PROP
1	119	0.109
2	89	0.079
3	129	0.054
4	96	0.042
5	94	0.053
6	110	0.073
7	96	0.094
8	101	0.050
9	99	0.071
10	84	0.095
11	92	0.120
12	96	0.000
13	92	0.076
14	106	0.085
15	97	0.062

Next step is to apply Greenacre's method by using PROC CLUSTER with the WARD method. WARD method or Ward's minimum variance method is a commonly used hierarchical cluster technique that merges two clusters together based on smallest decrease in the ANOVA sum of squares or Chi-square statistic. Ward's method tends to join clusters with a small number of observations, and is strongly biased toward producing clusters with roughly the same number of observations. It is also very sensitive to outliers (Milligan 1980).

```
ods output clusterhistory=clus_history;
proc cluster data=demo_freq method=ward;
   freq _freq_;
   var propor;
   id x2;
   run;
ods listing;

proc print data=clus_history;
run;
```

Step	NumberOfClusters	Idj1	Idj2	FreqOfNewCluster	SemipartialRSq	RSquared
1	14	3	5	223	0.0001	1
2	13	7	10	180	0.0001	1
3	12	6	9	209	0.0002	1
4	11	2	13	181	0.0002	0.999
5	10	CL14	8	324	0.0011	0.998
6	9	CL11	CL12	390	0.0025	0.996
7	8	1	11	211	0.0046	0.991
8	7	CL13	14	286	0.005	0.986
9	6	CL10	15	421	0.0055	0.981

10	5	CL6	4	517	0.0109	0.97
11	4	CL9	CL7	676	0.0375	0.932
12	3	CL8	CL4	887	0.1399	0.793
13	2	CL5	12	613	0.1832	0.609
14	1	CL3	CL2	1500	0.6094	0

The NumberOfClusters field gives the total number of cluster after merging. They are in descending order and the last value is 1, because all levels will be grouped into one big cluster eventually. Idj1 and Idj2 are the name of the levels used in the clustering. Any level starting with CL means the previously collapsed levels and the numbers following CL is the resulting cluster. For example, at step 5, Idj1 is CL14, which represents the new cluster of level 3 and level 5 at step1; Idj2 is 8, the level 8 in variable X2. FreqOfNewCluster indicates the number of observation after merging. SemipartialRSq is the semi-partial R-squares, which is the disease in the proportion of variance accounted for by joining two clusters. The value of SemipartialRSq at step1 is 0.0001, which is difference between the overall Chi-square statistic and Chi_square statistic after collapsing level 3 and 5. They are in increasing order, because WALD method merges two clusters with smallest variation from all possible combinations of any two clusters. The last column is squared multiple correlation (RSQ), representing the proportion of variance remaining after the levels are collapsed.

```
data cutoff;
   if _n_=1 then set chi;
   set clus_history;
   chisquare = _pchi_*rsquared;
   degfree = NumberOfClusters-1;
   logpvalue = logsdf('CHISQ',chisquare,degfree);
run;
```

The next step is to calculate overall Chi-square statistics.

```
proc freq data=demo;
   tables x2*y / chisq;
   output out=chi(keep=_pchi_) chisq;
run;
```

Result: _PCHI_=18.36

Then get the logarithm value by multiplying overall Chi-square statistic with proportion of variance remaining by individual level.

```
proc sql;
   create table cutoff as
   select a.*, b.*,
   logsdf('CHISQ',_pchi_*rsquared,numberofclusters-1) as logpvalue
   from chi a, clus_history b;
quit;
```

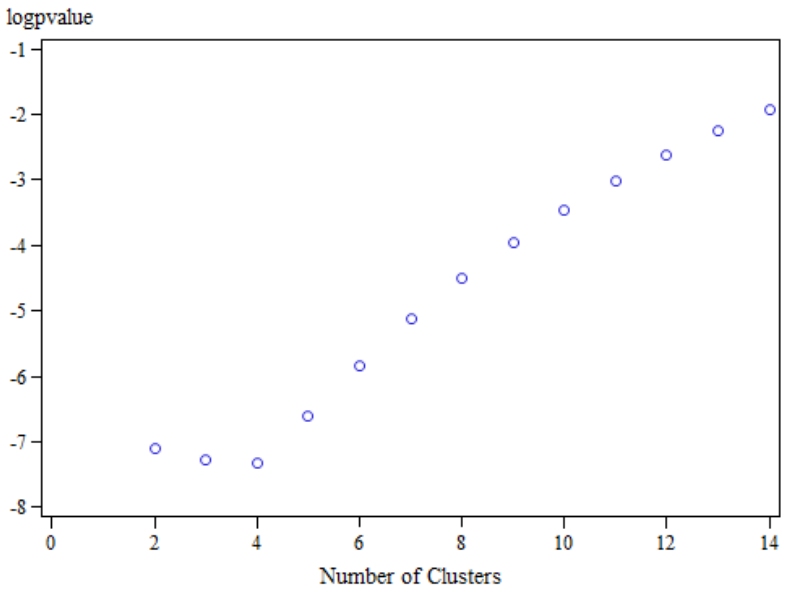

Figure 1. Plot of number of cluster vs. logPvalue

From the Figure 1, we can see that when clusters are 4, logvalue has the lowest value. So the optimum number of clusters is 4.

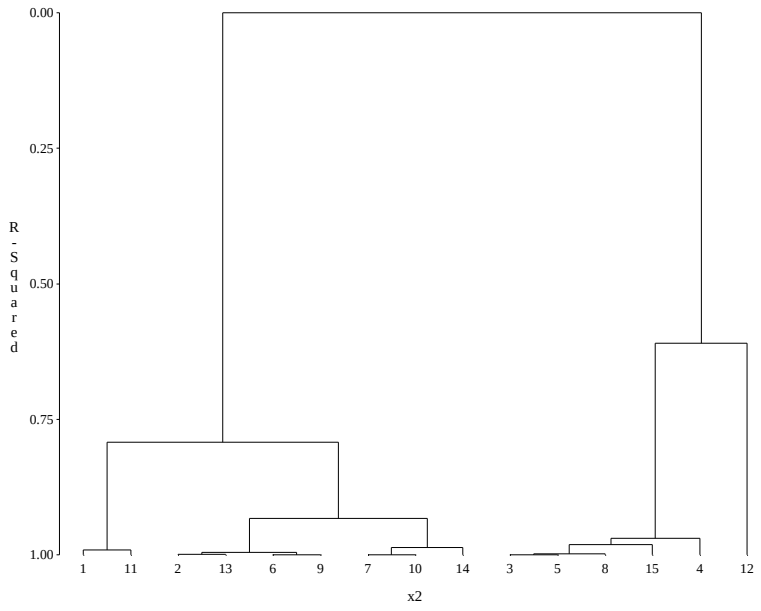

Figure 2. Dendrogram of clusters for variable X2.

The following summary table is based on dendrogram shown above.

CLUSTER	x2
1	3
	5
	8
	15
	4
2	7
	10
	6
	9
	2
	13
	14
3	1
	11
4	12

EXACT METHOD

Conventional logistic regression generates coefficient estimates through the maximum likelihood estimate (MLE) method. However, MLE is suitable for large sample estimation. If the sample is too small, standard error or P-value from MLE can be biased. Exact logistic regression can handle small sample size, skewed data or sample with quasi-complete separation. You may consider it as some generation format of Fisher's exact test. This method was originally introduced by Cox (1970), and the computation method has been employed in PROC LOGISTIC procedure since SAS® 8.1 version. Please keep in mind that logistic conditional exact methods is a time and memory consuming process and requires quite amount of computational time. Only 480 observations were randomly selected out of total 1500 from demo dataset.

```
proc logistic data=Demo_Sample desc;
   class x3 / param=ref;
   model y = x3;
   exact x3 / estimate=both;
run;
```

The following is the parameter estimate table from EXACT method.

Exact Parameter Estimates					
Parameter	Estimate	Standard Error	95% Confidence Limits		Two-sided p-Value
x3	1 0.9617	0.8545	-0.8974	3.3517	0.4390
x3	2 -6.3E-15	1.0128	-2.6551	2.6551	1.0000
x3	3 0.4194	0.9280	-1.7856	2.9362	1.0000
x3	4 0.4194	0.9280	-1.7856	2.9362	1.0000
x3	5 0.7224	0.8827	-1.2648	3.1619	0.6794
x3	6 -0.8933 *	.	-Infinity	1.2410	0.4958
x3	7 0.7224	0.8827	-1.2648	3.1619	0.6794

Note: * indicates a median unbiased estimate.

PENALIZED LIKELIHOOD METHOD OR FIRTH METHOD

Logistic regression with exact method can handle small sample size and quasi-complete separation data set. However, it can be very computationally demanding and not very efficient. An alternative method that can be used in large data set or quasi-complete separation is called penalized likelihood method or

FIRTH method, which was introduced by Firth (1993). In this method, Newton-Raphson algorithm was used to calculate the vector of first derivatives.

$$U(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_i x_i y_i - \sum_i x_i \hat{y}_i - \sum_i h_i x_i (.5 - \hat{y}_i)$$

Where, h_i is the i^{th} diagonal element of the "hat" matrix H.

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2} \text{ and } W = diag\{\hat{y}_i (1 - \hat{y}_i)\}$$

FIRTH method is implemented in PROC LOGISTIC procedure by using "FIRTH" option and it is recommended to use it with CLPARM option as well. This is because conventional confidence interval from Wald test produces biased and misleading result. CLPARM=PL means computation of confident interval is based on profile likelihood; while CLPARM=WALD, means the confidence intervals are individual Wald test based.

```
proc logistic data=demo desc;
   class x3;
   model y = x1 x3 / firth clparm=pl;
run;
```

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9000	0.2002	209.8496	<.0001
x1	1	-0.2369	0.1013	5.4641	0.0194
x3 1	1	0.5362	0.2994	3.2071	0.0733
x3 2	1	0.3785	0.3097	1.4939	0.2216
x3 3	1	0.9294	0.2762	11.3231	0.0008
x3 4	1	0.0143	0.3400	0.0018	0.9666
x3 5	1	0.5119	0.2944	3.0235	0.0821
x3 6	1	-3.0067	1.2430	5.8509	0.0156
x3 7	1	0.4856	0.3044	2.5445	0.1107

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		Estimate	95% Confidence Limits	
Intercept		-2.9000	-3.5218	-2.5875
x1		-0.2369	-0.4385	-0.0361
x3	1	0.5362	-0.0388	1.2498
x3	2	0.3785	-0.2235	1.1027
x3	3	0.9294	0.4153	1.6196
x3	4	0.0143	-0.6679	0.7708
x3	5	0.5119	-0.0505	1.2203
x3	6	-3.0067	-7.2396	-1.3151
x3	7	0.4856	-0.1023	1.2043

CONCLUSION

Quasi-complete separation is common in regression analyses where the outcome variable is categorical. It causes the failure of convergence in the model and yields unreliable inference and misleading results. This paper demonstrated multiple methods to remedy this issue ranging from data configuration to statistical approaches. Greenacre's method is good when the categorical variable(s) with problem has (have) a lot of levels; EXACT method is good when the sample is small and FIRTH method is suitable for small or big data.

REFERENCES

Allison, Paul, D (2008), Convergence Failures in Logistic Regression, Philadelphia, PA, SAS Global Forum 2008, Paper 360-2008.

Allison, Paul D. (2012), Logistic Regression Using SAS: Theory and Application, Second Edition. Cary, NC: SAS Institute Inc.

Derr, Robert, (2009), Perform Exact Logistic Regression with the SAS® System, SAS Institute Inc., Cary,NC, Paper P254-25.

Predictive Modeling Using Logistic Regression, Course Note, 2008, Cary, NC: SAS Institute Inc.

SAS OnlineDoc 9.2, SAS/STAT(r) User's Guide, Second Edition.

Webb, Mandy, Wilson,Jeffrey and Chong, Jenny(2004) An Analysis of Quasi-complete Binary Data with Logistic Models: Applications to Alcohol Abuse Data, Journal of Science, 273-285.

ACKNOWLEDGMENTS

I would like to acknowledge Jim Jones, VP Healthcare Analytics, AmeriHealth Caritas, and Wanzhen Gao, Director, Health Care Analytics, AmeriHealth Caritas for their support and encouragement.

I would like to acknowledge Seungyoung Hwang from Johns Hopkins University for assistance with reviewing; Scott Leslie, SGF Mentoring Program Coordinator, for help with grammar checking.

Amerihealth Caritas is the nation's leader in the health care solutions with more than 30 years of experience managing care for individuals and families in publicly funded program.

CONTACT INFORMATION

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

Xinghe Lu
AmeriHealth Caritas Family of Companies
3rd Floor, 200 Stevens Drive, Philadelphia, PA 19113.
E-mail: xlu@amerihealthcaritas.com
Web: http://www.amerihealthcaritas.com/