

Quasicomplete Separation in Logistic Regression: A Medical Example

Madeline J. Boyle, Carolinas Medical Center, Charlotte, NC

ABSTRACT

Logistic regression can be used to model the relationship between a dichotomous outcome variable and explanatory variables that can be either dichotomous or continuous. When using the LOGISTIC procedure in the SAS/STAT® software, one problem that can arise is complete or quasicomplete separation of the data points. An example from a blunt intestinal injury study completed at a major metropolitan hospital in the southeast will be presented. The quasicomplete separation of the data points will be presented, as will the steps taken in an attempt to remedy the problem.

INTRODUCTION

The LOGISTIC procedure in the SAS/STAT® software is useful in analyzing a binary response variable, when the response variable takes on one of two possibilities denoted by zero and one. For example, if the characteristic of interest was disease and the disease was not present, $Y=0$, and if the disease was present, $Y=1$. When performing logistic regression, possible hindrances to the data analysis arise from the data. The existence of maximum likelihood estimates for parameters of the logistic model depend on the configuration of the sample points in the observation space. If no finite maximum likelihood estimates exist, then you have the situation described here as "infinite parameters." The sample points can fall into "three mutually exclusive and exhaustive categories: complete separation, quasicomplete separation, and overlap" (So, 1993). These three categories will be discussed briefly, possible remedies for separation will be given, and finally, a medical example of quasicomplete separation will be given, along with the attempted solution.

INFINITE PARAMETERS

Infinite parameters refer to the situation when no finite maximum likelihood estimate exists, as can occur for a logistic regression model. The existence of these estimates depends on the configuration of the sample points in the observation space as mentioned before. The three types of configurations are complete separation, quasicomplete separation, and overlap. "A Tutorial on Logistic Regression," (So, 1993) gives more information on how infinite parameters arise in each of these configurations. A brief description of the configurations and possible remedies is found in Logistic Regression Examples Using the SAS System (SAS Institute Inc., 1995) and is summarized below.

Complete Separation

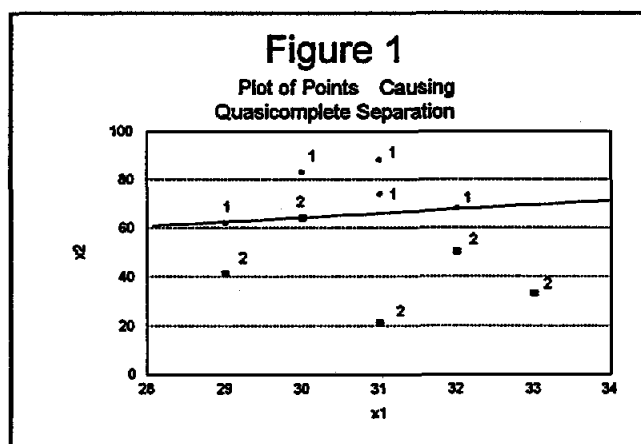
If a complete separation exists in the sample points, then the maximum likelihood estimate does not exist. In this case there exists a vector of pseudoestimates that correctly allocates all observations to their observed response groups. Such a data configuration gives an infinite set of nonunique estimates. At each iteration, the predicted probability that each observation belongs to its observed response group rapidly grows to one and the log likelihood diminishes to zero.

Quasicomplete Separation

If a quasicomplete separation exists in the sample points, then the maximum likelihood estimate does not exist. The data are not completely separated and a vector of

pseudoestimates correctly allocates all but a nonempty set of observations to their response groups. Such a data configuration also gives an infinite set of nonunique estimates. The log likelihood does not diminish to zero at each iteration, as it does in the case of complete separation. This is the separation that exists in the medical example to be discussed later.

Figure 1 displays the quasicomplete separation from the data used by So (1993). From the figure one can easily see that the two groups cannot be separated by a straight line.

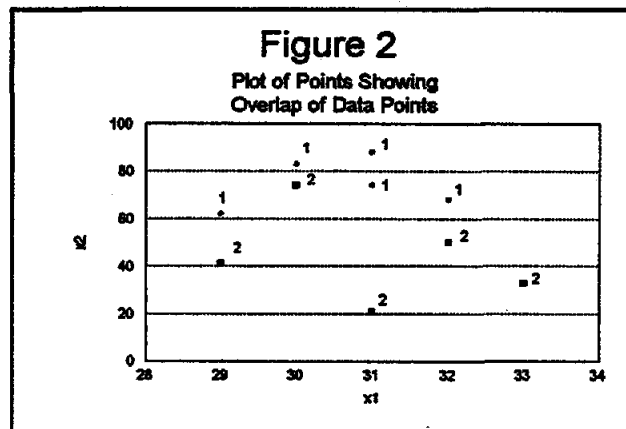


The line shown on the graph illustrates the quasicomplete separation of the two groups; if the value of "x2" for the point on the line from group number two was changed to any lower number (e.g., fifty), then a case of complete separation would exist. However, in this example since the data are not completely separated and at least one member from each group lies on the line, quasicomplete separation exists.

Overlap

An overlap of the sample points exists when neither a complete nor quasicomplete separation of the data exists. If there is overlap in the sample points, then the maximum likelihood estimate exists and is unique. Figure 2 displays the overlap of the data points from the data used

by So (1993). Every straight line that can be drawn on this graph will always have a sample point from each of the two groups on the same side of the line; therefore there is overlap of the data.



Remedies

If there is complete or quasicomplete separation in your data, the maximum likelihood estimates do not exist, and although version 6.11 of the SAS® system will continue to run, the statistics from the model may not be valid. Various remedies are available; first, examine the original raw data for errors and if any are found, repeat the analysis to see if the separation still exists. If this does not work, there are some options involving the data: 1) categorize quantitative variables, 2) use fewer or different explanatory variables, or 3) collect more data. "With increasing sample size, the probability of observing a set of separated data points tends to zero, no matter what the sample scheme" (Albert and Anderson, 1984). The model may also be altered to remedy the separation. Try reclassifying the response variable, or in a model with a selection setting if you encounter complete separation when you use, for example, backwards elimination selection method, try using forward or stepwise selection instead.

Complete and quasicomplete separation usually occur with small samples and qualitative data.

However, complete or quasicomplete separation can occur for any type of data or sample size. An important note to keep in mind is that the more explanatory variables your model contains, the greater the likelihood of encountering complete or quasicomplete separation.

MEDICAL EXAMPLE

A retrospective study over six years of patients with blunt intestinal injury was completed at the Carolinas Medical Center in Charlotte, NC. The study objective was to identify factors associated with a delay of more than six hours between the time of injury and therapeutic laparotomy. The statistical analysis included a stepwise logistic regression to determine whether a set of explanatory variables could predict the outcome of delayed laparotomy, having a life-threatening injury, and the location of injury (small bowel or colon). The analyses were completed using the SAS® system for Windows, version 6.11.

The original explanatory set of thirty-three variables contained categorical, dichotomous, and continuous variables. These variables included mechanism of injury, abdominal exam results, fractures, Computerized Tomography (CT) exam results, blood alcohol level, Diagnostic Peritoneal Lavage (DPL) exam results, and hypotensive status. The sample included sixty-one patients who were confirmed by laparotomy to have sustained blunt intestinal injury with thirty of those patients having a laparotomy more than six hours post injury.

An obvious drawback to a stepwise logistic regression with such a small sample size and so many explanatory variables was the lack of ability in the model to replicate the results for another set of patients. A rule of thumb proposed by Harrell, et. al., (1985) was that "...one should not attempt a stepwise regression when there are fewer than ten times as many events in the training sample as there are

candidate predictor variables." When the response variable is binary, the limiting sample size is the sample size of the less frequent response category. In this example this is the thirty patients with a delayed laparotomy. Using Harrell's rule of thumb, three explanatory variables could be introduced into the model. However, since the focus of this example is on the quasicomplete separation of the data and because the data was collected for a six year period, we will not comment further on the sample size.

When we attempted to use the LOGISTIC procedure on our model for delayed laparotomy with thirty-three candidate explanatory variables, an intercept and three other variables were entered into the model, and then a warning message that a quasicomplete separation of the data existed and the Maximum Likelihood Estimates did not exist was printed on the output. At this point the procedure continued fitting the model and statistics, but at each step noted the model validity was questionable; this step is new to the latest release, version 6.11, of the SAS® system. In the previous versions, the model fitting stopped as soon as the separation was found and a warning was indicated in the output. The same result was returned when modeling whether or not the patient had a life-threatening illness. When the location of injury (small bowel or colon) was examined as a response variable, the stepwise logistic failed to find an adequate model, based on the low sensitivity and specificity of the model. The highest sensitivity and specificity found were 22.5% and 14.3%, respectively. However, this response variable did not encounter the problem with quasicomplete separation of the sample data points.

The initial steps taken in an attempt to remedy the quasicomplete separation of the data points and the weakness of the third model included verifying the data, and combining explanatory variables to reduce the number entered into the model to seven. The new set of explanatory

variables included mechanism of injury, DPL gross and micro exam results, blood alcohol level, and four groups depending on any injuries or fractures found on initial examination, also depending on what variable was being modeled, two of the three following variables were included: location of injury, delay of more than six hours before surgery or not, and life-threatening illness or not, when appropriate.

Reducing the number of explanatory variables eliminated the quasicomplete separation in the two response variables, delay and injury type. The third response variable however, location of injury, now exhibited the problem of quasicomplete separation of the sample data points.

Attempts to eliminate the quasicomplete separation of the data for the variable location of injury were unsuccessful. Backward stepwise regression was attempted, as well as reclassifying the location of injury. Since the quasicomplete separation of the data was unable to be resolved, the statistics from this stepwise logistic regression model for this variable was not interpretable.

The table below displays the quasicomplete separation in this example for the response variable, location of injury by looking at the number of patients with pelvic injuries.

Table of Location of Injury for Patients with Pelvic Injuries

Location of Injury	Colon	Small Bowel
Pelvic Injury	4	0
Other Injury	27	30

As seen in the table, none of those patients who have small bowel injuries had pelvic injuries. This results in a quasicomplete separation of the data. Those patients who had a pelvic injury were exclusively in the group of patients who had a colon injury. The patients with another type of injury had their location of injury as either the small bowel or colon. Therefore, quasicomplete separation of the sample points existed; the data are separated into two groups with the exception of a nonempty set of observations. In the simple example illustrated in Figure 1, the majority of points were correctly allocated to their groups. Only three points in that set were not correctly allocated. In this case the majority of patients were in the non-allocated set, while four were correctly allocated to the group who had a location of injury at their colon.

The partial output for this example is given in Appendix A. In this output the warning of quasicomplete separation of the data can be seen in the fourth step of the procedure, as well as the log likelihood that does not diminish to zero as it does in complete separation. The variables entered into the model are as follows: Threaten-(whether the patient had a life-threatening illness or not), Alcohol-(whether patient's alcohol level was greater than zero or no alcohol was found/test not done), MVAU-(whether patient was involved in an unrestrained motor vehicle accident) and Grp_2-(whether the patient had a pelvic injury or other type of injury). The odds ratios and other statistics are calculated for this model with a warning given about questionable model validity, which refers to the existence of quasicomplete separation of the data points. These additional statistics are not based on the maximum likelihood estimates because these values do not exist; therefore, these statistics should not be used until the model validity has been determined. Running the model with another set of data is one manner of verifying model validity.

CONCLUSION

Quasicomplete separation occurs when the data are not completely separated and a vector of pseudoestimates correctly allocates all but a nonempty set of observations to their response groups. This was illustrated with a medical example.

Some remedies can be attempted to relieve the separation of the sample data, including increasing the sample size, categorizing quantitative variables and reducing the number of explanatory variables. The latter proved useful in two of the three logistic regression models attempted in our example. For the third model using location of injury, as the response variable, the quasicomplete separation could not be eliminated and a successful model could not be achieved.

ACKNOWLEDGEMENTS

SAS and SAS/STAT are registered trademarks or trademarks of the SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES

- Albert A, Anderson JA (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71: 1-10.
- Harrell FE, Lee KL, Matchar DB, Reichert TA (1985). Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions. *Cancer Treatment Reports*, 69: 1071-7.

SAS Institute Inc., Logistic Regression Examples Using the SAS System, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1995.

SAS Institute Inc., SAS/STAT® Software: Changes and Enhancements through Release 6.11, Cary, NC; SAS Institute Inc., 1996.

So, Y. (1993). A Tutorial on Logistic Regression. *Proceedings of the Eighteenth Annual SAS Users Group International*, 1290-1295.

Address correspondence to:

Madeline J. Boyle
Department of Biostatistics
Research Office Building, Room 312
Carolinas Medical Center
P.O. Box 32861
Charlotte, NC 28232-2861
Work: 704-355-4159
Fax: 704-355-1880
Email: mjboyle@med.unc.edu

APPENDIX A:
PARTIAL OUTPUT FROM PROC LOGISTIC FOR THE QUASICOMPLETE EXAMPLE

The LOGISTIC Procedure

Data Set: INJURIES

Response Variable: LOCATE 1=small bowel & 0=colon

Response Levels: 2

Number of Observations: 61

Link Function: Logit

Response Profile

Ordered Value	LOCATE	Count
1	0	31
2	1	30

...

Step 4. Variable Grp_2 entered:

Maximum Likelihood Iteration Phase

Iter	Step	-2Log L	Intercept	Threaten	Grp_2	Alcohol	MVAU
0	INITIAL	84.54756	0.0327	0	0	0	0
1	IRLS	63.17109	1.1055	-1.6892	1.3942	-1.3805	0.6904
...							
11	IRLS	60.98928	1.5249	-2.2162	11.476	-2.0950	1.0144
12	IRLS	60.98925	1.5249	-2.2162	12.476	-2.0950	1.0144

WARNING: There is possibly a quasicomplete separation in the sample 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 in question.

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr> Chi-Square
1	Threaten		1	11.6344	.	0.0006
2	Alcohol		2	3.9853	.	0.0459
3	MVAU		3	4.2257	.	0.0398
4	Grp_2		4	1.8067	.	0.1789
5		Grp_2	3	.	0.00129	0.9714