Paper 081-31

Application of Proc Discrim and Proc Logistic in Credit Risk Modeling

Jin Li, Capital One Financial Service, Richmond, VA

ABSTRACT

PROC LOGISTIC is well known in credit card industry as a way to model binary variables such as 'response' or 'charge off'. However, there are many occasions that the dependent variable actually has more than two groups. For example, in addition to model whether or not an account will charge off, it is also important to know if it is going to attrite (voluntarily close account). For those that are going to charge off, we might interest to know whether they are going to charge off within a year, in 2nd year or in 2+ years. Answering these important questions helps us to better manage our risk, maximize our NPV and hence gain a competitive position in the market. In this article, PROC DISCRIM and PROC LOGISTIC are used to address these questions. The theory of each method is introduced in background section. Detailed SAS programs and results are shown in methodology/result section. The comparison between these two methods is outlined in the discussion section. Hopefully this discussion provides useful information for people who are interested in categorical data modeling using SAS.

INTRODUCTION

Binary logistic modeling is widely used in credit card industry. Indeed, a lot of times we find ourselves in a position to model a binary variable such as 'response' or 'charge off'. But there are also lots of occasions that we care more than two categories. For example, besides the interest in 'charge off' rate, we might also be interested in 'attrition' (voluntarily close account) at the same time. Instead of just modeling charge off, we are also interested to know whether the charge off will happen within 1 year of account open, or 2nd year or year 2+. SAS provides a great package of doing categorical data modeling. We will only discuss "PROC DISCRIM" and "PROC LOGISTIC" in this article. For other methods in categorical data modeling, please refer to Stokes et. al.[1].

Discriminant Analysis is an earlier alternative to *Logistic Regression*. Despite its strict restrictions on data distributions, it still has value when it comes to multiple group classification. Unlike binary *Logistic Regression*, *Discriminant Analysis* can be used to handle more than two categories. Computationally, it is very similar to analysis of variance (ANOVA). When given a group of variables, F tests are conducted to decide which variables are significant to differentiate between groups. SAS provides "PROC STEPDISC" to carry out the variable selection (F tests) in *Discriminant Analysis*. After the variables are selected, discriminant functions, aka. classification criterion, are developed to assign group membership. The classification criterion built in SAS takes into account the prior probabilities of groups (PRIORS statement in PROC DISCRIM) and it can be a linear function as well as a quadratic function. Option POOL in PROC DISCRIM can be used to specify which discriminant function is applied in the analysis.

It is very important to note that *Discriminant Analysis* demands a lot of assumptions on the data. Failure to meet these assumptions may lead to serious misclassification and meaningless results in the end. The DISCUSSION section of this paper covers some key assumptions in *Discriminant Analysis* and their impact on the analysis. SAS also provides nonparametric methods for *Discriminant Analysis* when the data does not have multivariate normal distribution. However, it is the author's belief that parametric method is sufficient for financial data analysis and therefore, nonparametric methods are not covered in this paper. For readers who are interested in nonparametric methods, please refer to Goldstein and Dillon[2], Hand[3] and SAS manuals for details.

In addition to popular application in binary logistic regression, PROC LOGISTIC can also be used to handle polytomous dependent variables through *Generalized Logit Models*. Specifically, SAS users can use LINK=glogit option in PROC LOGISTIC to carry out a *Generalized Logit Regression*. *Generalized Logit Regression* requires fewer assumptions than *Discriminant Analysis*. It can handle both categorical and continuous variables, and the predictors do not have to be normally distributed, linearly related, or of equal variance within each group[4].

BACKGROUND

DISCRIMINANT ANALYSIS

PROC DISCRIM develops a classification criterion using a measure of generalized squared distance. Each observation is then classified into a group from which it has the smallest generalized squared distance. The generalized squared distance from x to group t is defined as:

$$D_t^2(x) = d_t^2(x) + g_1(t) + g_2(t)$$

where $d_t^2(x) = (x - \mu_t)' S^{-1} (x - \mu_t)$

$g_1(t) = \ln|S_t|$ if the within-group covariance matrices are used or

$= 0$ if the pooled covariance matrix is used

$g_2(t) = -2\ln(q_t)$ if prior probabilities are not all equal
$= 0$ if prior probabilities are all equal
μ_t denoting the vector containing variable means in group t
S denoting S_t if within-group covariance matrices are used and S_p if the pooled covariance matrix is used
S_t denoting the covariance matrix within group t
S_p denoting the pooled covariance matrix
q_t denoting the prior probability of group t

As the group-specific density estimate at x from group t is defined as:

$$f_t(x) = (2\pi)^{-\frac{p}{2}}|S|^{-\frac{1}{2}}\exp\left(-0.5d_t^2(x)\right)$$ where p is dimension of vector x

The posterior probability of x belonging to group t is then calculated according to Bayes' Theorem as:

$$p(t \mid x) = \frac{q_t f_t(x)}{\sum_{i=1}^{m} q_i f_i(x)} = \frac{q_t (2\pi)^{-\frac{p}{2}}|S|^{-\frac{1}{2}}\exp\left(-0.5d_t^2(x)\right)}{\sum_{i=1}^{m} q_i (2\pi)^{-\frac{p}{2}}|S|^{-\frac{1}{2}}\exp\left(-0.5d_i^2(x)\right)}$$

$$= \frac{q_t|S|^{-\frac{1}{2}}\exp\left(-0.5D_t^2(x)+0.5g_1(t)+0.5g_2(t)\right)}{\sum_{i=1}^{m} q_i|S|^{-\frac{1}{2}}\exp\left(-0.5D_i^2(x)+0.5g_1(i)+0.5g_2(i)\right)}$$

where i indicates the group number and there are total m groups

Let's consider $g_1(t) = \ln|S_t|$ and $g_2(t) = -2\ln(q_t)$ case here, all other cases can be derived the same way as shown below:

$$p(t \mid x) = \frac{q_t|S_t|^{-\frac{1}{2}}\exp\left(-0.5D_t^2(x)+0.5\ln|S_t|-\ln(q_t)\right)}{\sum_{i=1}^{m} q_i|S_i|^{-\frac{1}{2}}\exp\left(-0.5D_i^2(x)+0.5\ln|S_i|-\ln(q_i)\right)} = \frac{q_t|S_t|^{-\frac{1}{2}}\times\exp\left(-0.5D_t^2(x)\right)\times|S_t|^{\frac{1}{2}}\times\frac{1}{q_t}}{\sum q_i|S_i|^{-\frac{1}{2}}\times\exp\left(-0.5D_i^2(x)\times|S_i|^{-\frac{1}{2}}\times\frac{1}{q_i}\right)}$$

$$= \frac{\exp\left(-0.5D_t^2(x)\right)}{\sum\exp\left(-0.5D_i^2(x)\right)}$$

GENERALIZED LOGIT REGRESSION

For the convenience of notation, let's consider the case where the response variable has only three categories (Y=0, 1, 2). By choosing Y=0 as the reference category, the generalized logit models are given by:

$$\log\left(\frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}\right) = X'\beta_1$$

$$\log\left(\frac{P(Y=2 \mid X)}{P(Y=0 \mid X)}\right) = X'\beta_2$$

The maximum likelihood estimate of β_1 and β_2 can be calculated by maximizing

$$l(\beta_1,\beta_2) = \sum_{i=1}^{n}\log\left(\Pr(Y = y_i \mid x_i)\right)$$ where n is the number of subjects.

The predicted probability for each category will then be:

$$P(Y=0 \mid X) = \frac{1}{1+e^{X'\beta_1}+e^{X'\beta_2}}$$

$$P(Y=1 \mid X) = \frac{e^{X'\beta_1}}{1+e^{X'\beta_1}+e^{X'\beta_2}}$$

$$P(Y=2 \mid X) = \frac{e^{X'\beta_2}}{1+e^{X'\beta_1}+e^{X'\beta_2}}$$

METHODOLOGY/RESULTS

The best way to illustrate *Discriminant Analysis* and *Generalized Logit Regression* in SAS is through examples. Suppose we need to build a model to predict the status of accounts after 2 years of opening. The target (dependent variable) has been grouped in three main categories of interest: charge off ('C'), attrition ('A') and still open ('O'). There are 3 independent variables available for modeling: average credit limit in other credit cards, total number of trades (credit cards, mortgage, loans etc) in credit file, utilization of other credit cards (utilization >1.0 means over credit limit). The build sample would look like something below:

```
DATA build;
      INPUT credit_limit number_of_trades utilization Target $ @@;
      DATALINES;
1300 22 0.41 O    400  9  0.88 C   2500 29 0.57 O    4400 49 0.29 A
4900 43 0.86 A    5200 49 0.36 A   2500 35 1.02 A    600  9  0.83 C
1100 28 0.88 O    600  11 0.50 C   500  33 0.12 C    1600 26 0.90 O
4000 49 0.43 A    400  11 1.09 C   2400 36 0.22 O    2500 29 0.76 O
5400 53 0.30 A    2700 32 0.68 O   500  8  1.09 C    2000 36 0.54 O
1600 35 0.54 O    500  8  1.10 C   650  13 1.00 C    5000 46 0.71 O
5100 52 0.17 A    2200 37 0.60 O   1400 25 0.50 O    2200 31 0.37 O
4700 49 0.75 A    1500 33 0.25 O   1600 29 0.63 O    2200 33 0.25 O
1500 30 0.27 O    1600 38 0.58 O   1800 39 0.45 O    2100 37 0.37 O
1700 36 0.68 O    2100 28 1.00 O   1600 29 0.65 O    600  10 0.50 C
1300 25 0.22 O    1900 24 0.18 O   1900 33 0.49 O    2600 30 0.53 O
2300 24 0.12 O    1200 34 0.30 O   5400 52 0.47 A    2600 35 0.33 O
1700 24 0.65 O    500  8  0.73 C   600  7  0.40 C    4400 47 0.10 A
2200 31 0.50 O    1400 34 0.64 O   5100 47 0.71 A    2000 31 0.17 O
2300 30 0.30 O    1700 32 0.59 O   1000 30 0.11 O    1400 33 0.23 A
2400 32 0.90 C    3300 30 0.87 A   2300 35 0.47 O    2800 35 0.58 O
500  12 0.48 C    2700 37 0.69 O   2200 28 0.38 O    4500 54 0.29 A
4900 50 0.29 A    550  11 0.41 C   1900 25 0.20 O
;
RUN;
```

DISCRIMINANT ANALYSIS

If we want to carry out *Discriminant Analysis* on this data, generally, it is a good practice to run PROC STEPDISC before executing PROC DISCRIM to find out which variables yield biggest difference between target groups and eliminate those insignificant variables. Following code is written for this variable selection purpose:

```
PROC STEPDISC DATA=build;
CLASS target;
VAR credit_limit number_of_trades utilization;
RUN;
```

Partial SAS output is shown below:

```
                                 The STEPDISC Procedure
                           The Method for Selecting Variables is STEPWISE

                 Observations        71          Variable(s) in the Analysis        3
                 Class Levels         3          Variable(s) will be Included       0
                                                 Significance Level to Enter     0.15
                                                 Significance Level to Stay      0.15

                                 Stepwise Selection: Step 3
                            Statistics for Removal, DF = 2, 67
                                          Partial
                    Variable              R-Square     F Value     Pr > F

                    credit_limit            0.3098       15.04     <.0001
                    number_of_trades        0.3642       19.19     <.0001

                                    No variables can be removed.
                                  Statistics for Entry, DF = 2, 66
                                           Partial
                    Variable              R-Square     F Value     Pr > F     Tolerance
                    utilization             0.0128        0.43     0.6530        0.2180

                               Stepwise Selection Summary
        Number                                   Partial                         Wilks'    Pr <
  Step    In     Entered        Removed      R-Square F Value   Pr > F     Lambda   Lambda

     1     1   number_of_trades               0.7447   99.15   <.0001  0.25534412  <.0001
     2     2   credit_limit                   0.3098   15.04   <.0001  0.17624169  <.0001
```

```
                                                      Average
                                                      Squared
            Number                                  Canonical  Pr >
    Step      In  Entered            Removed      Correlation  ASCC

     1        1  number_of_trades                  0.37232794    <.0001
     2        2  credit_limit                      0.50056710    <.0001
```

As indicated in the output, a stepwise selection is conducted in this step, the significance level of retaining a variable or adding a variable is 0.15. Similar to PROC LOGISTIC, variable selection method can be changed to backward selection or forward selection using "METHOD =" option. The significance level of variable selection can also be specified through "SLENTRY =" and "SLSTAY =" option. In this example, utilization is found to be insignificant to differentiate different target groups (p value = 0.6530). Therefore, only "credit limit" and "number of trades" will be used for final analysis conducted in PROC DISCRIM.

The SAS code for PROC DISCRIM is shown below:

```
PROC DISCRIM DATA=build TESTDATA=build POOL=test OUT=disc;
PRIORS prop;
CLASS target;
VAR   credit_limit number_of_trades;
RUN;
```

In the above code, option "POOL=test" enables SAS to test whether linear discriminant functions or quadratic discriminant functions are appropriate for this analysis. SAS default is using linear discriminant functions (pool=yes). You can also choose quadratic discriminant functions by specifying pool = no. However, it is always good to test the homogeneity of the within-group covariance matrices before choosing discriminant functions. That is why we use "POOL=test" here.

The PRIORS statement in the code specifies the prior probabilities of group membership. As there is no previous knowledge about the actual distribution among three target groups, it is assumed that prior probabilities proportional to the sample sizes in this code.

The SAS output from PROC DISCRIM summarizes the data distribution at the beginning, then it shows the model prediction results on build sample at the end:

```
                     The DISCRIM Procedure
       Observations      71          DF Total                70
       Variables          2          DF Within Classes       68
       Classes            3          DF Between Classes       2

                        Class Level Information
             Variable                                                  Prior
Target       Name         Frequency       Weight     Proportion    Probability

A            A                   15      15.0000       0.211268       0.211268
C            C                   14      14.0000       0.197183       0.197183
O            O                   42      42.0000       0.591549       0.591549


                     The DISCRIM Procedure
       Classification Summary for Calibration Data: WORK.BUILD
     Resubstitution Summary using Quadratic Discriminant Function

               Generalized Squared Distance Function
          2           _          -1    _
         D (X) = (X-X )' COV   (X-X ) + ln |COV | - 2 ln PRIOR
          j          j      j      j          j              j

          Posterior Probability of Membership in Each Target
                                  2                     2
            Pr(j|X) = exp(-.5 D (X)) / SUM exp(-.5 D (X))
                                j        k            k
```

```
Number of Observations and Percent Classified into Target

From
Target              A           C           O       Total

A                  12           0           3          15
                80.00        0.00       20.00      100.00
C                   0          12           2          14
                 0.00       85.71       14.29      100.00
O                   1           0          41          42
                 2.38        0.00       97.62      100.00
Total              13          12          46          71
                18.31       16.90       64.79      100.00

Priors        0.21127     0.19718     0.59155


             Error Count Estimates for Target

                         A           C           O       Total

Rate                0.2000      0.1429      0.0238      0.0845
Priors              0.2113      0.1972      0.5915
```

According to the output above, we can see that among 71 records in the build sample, there are 6 observations (8.45%) misclassified into "wrong" groups. Specifically, 3 attrited accounts ('A') are classified as open accounts ('O'), 2 charge off accounts ('C') are classified as open accounts ('O'), and 1 open accounts ('O') is classified as attrited account ('A'). All the other accounts are classified correctly.

The actual probability of each observation falls into each category ('A', 'C', 'O') can be found in dataset "disc".

GENERALIZED LOGIT REGRESSION

In PROC LOGISTIC, model statement option "LINK=glogit" can be use to conduct a *Generalized Logit Regression* to attack previous problem. The SAS code is shown below:

```
PROC LOGISTIC DATA=build;
MODEL target(ref='O') = credit_limit number_of_trades utilization /
SELECTION=stepwise LINK=glogit;
RUN;
```

Note that stepwise selection is used in above code to select variables. SAS uses 0.05 as default significance level in this case. The final model is shown in the output:

```
                            The LOGISTIC Procedure

NOTE: No (additional) effects met the 0.05 significance level for entry into the
model.

                                  Summary of Stepwise Selection
               Effect                                Number        Score         Wald
Step   Entered          Removed             DF          In    Chi-Square   Chi-Square    Pr > ChiSq

1   number_of_trades                         2           1       52.8706         .          <.0001
2   credit_limit                             2           2        7.0584         .          0.0293
3                  number_of_trades          2           1         .         4.4079       0.1104


                                  Type III Analysis of Effects
                                                         Wald
                      Effect                DF     Chi-Square     Pr > ChiSq

                      credit_limit           2        30.0739         <.0001
```

Analysis of Maximum Likelihood Estimates

Parameter	Target	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	A	1	-6.6500	1.4785	20.2293	<.0001
Intercept	C	1	4.6242	1.4147	10.6847	0.0011
credit_limit	A	1	0.00189	0.000466	16.3740	<.0001
credit_limit	C	1	-0.00470	0.00126	13.9788	0.0002

Odds Ratio Estimates

Effect	Target	Point Estimate	95% Wald Confidence Limits	
credit_limit	A	1.002	1.001	1.003
credit_limit	C	0.995	0.993	0.998

It may seem that among all three predictors, only credit_limit is significant. However, if we loosen up the significance level to 0.15 to match up with what we've done in *Discriminant Analysis*, we shall find that "number_of_trades" is the second important predictor. SAS code below is used to generate a generalized logit model on two predictors: credit_limit, number_of_trades:

```
PROC LOGISTIC DATA=build;
MODEL target(ref='O') = credit_limit number_of_trades / LINK=glogit;
RUN;
```

If the build sample is scored on the generalized logit model generated from above code, and the group membership is assigned according to the maximum probability, then we will find that this model gives same results as *Discriminant Analysis*: 3 attrited accounts ('A') are classified as open accounts ('O'), 2 charge off accounts ('C') are classified as open accounts ('O'), and 1 open accounts ('O') is classified as attrited account ('A'). All the other accounts are classified correctly.

SAS code used to manually score the model, analyze the results and their corresponding outputs are provided below:

```
DATA logit;
    SET build;
 phat_A=-9.1618+credit_limit*0.00116+number_of_trades*0.1263;
 phat_C=6.3220+credit_limit*(-0.00272)+number_of_trades*(-0.1629);
 prob_O = 1/(1+exp(phat_A)+exp(phat_C));
 prob_A = prob_O*exp(phat_A);
 prob_C = prob_O*exp(phat_C);
 max = max(prob_A, prob_C, prob_O);
 IF prob_O = max THEN pred = 'O';
 ELSE IF prob_A = max THEN pred = 'A';
 ELSE pred = 'C';
RUN;

PROC FREQ DATA=logit;
TABLES target * pred /LIST;
RUN;
```

The FREQ Procedure

Target	pred	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	A	12	16.90	12	16.90
A	O	3	4.23	15	21.13
C	C	12	16.90	27	38.03
C	O	2	2.82	29	40.85
O	A	1	1.41	30	42.25
O	O	41	57.75	71	100.00

DISCUSSION

PRESENCE OFCONTROL GROUP

At first glance, both *Discriminant Analysis* and *Generalized Logit Regression* provide membership prediction and can be used interchangeably. However, a closer look reveals that these two methods do not use the same approach to compare the groups. *Discriminant Analysis* compares all the groups simultaneously, while *Generalized Logit Regression* compares each group with a reference group. In our case of example, *Discriminant Analysis* compares 'A', 'C', 'O' groups at the same time, while *Generalized Logit Regression* compares 'A' with 'O' and 'C' with 'O'. If we are interested in the comparison between group 'A' and 'C', then a different *Generalized Logit Regression* model (using 'A' or 'C' as reference group) should be built. Consequently, *Discriminant Analysis* selects the set of variables that vary significantly across all the groups, while *Generalized Logit Regression* only selects the set of variables that differentiate between test groups and reference group.

ASSUMPTIONS

As mentioned in previous sections, *Discriminant Analysis* usually demands more restricted data structure than *Generalized Logit Regression*. However, not all assumptions for *Discriminant Analysis* are equally important. The author has found that PROC DISCRIM works well on some occasions when multivariate normal assumption is violated. For example, if we plot the Q-Q plot of our original build sample (Figure 1, Figure2), we can see while there is obvious deviation from multivariate normal distribution*, *Discriminant Analysis* still provides same result with *Generalized Logit Regression*.

* The actual code for Q-Q plot can be found in Chapter 2 of Khattree and Naik[5]

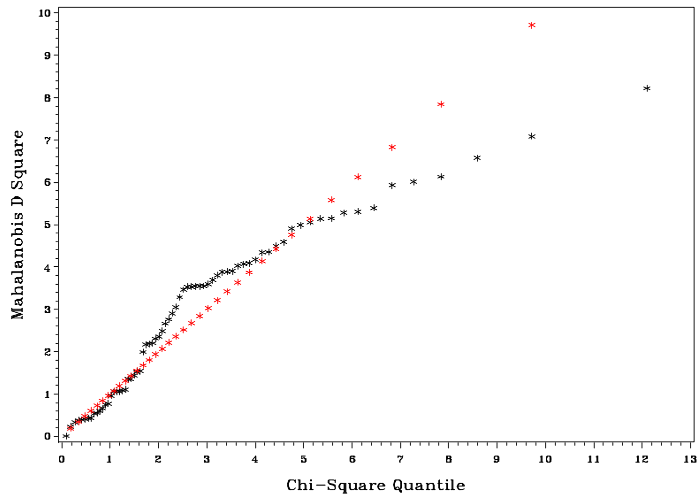

Figure 1: Q-Q plot for assessing normality with 3 variables: credit_limit, number_of_trades, utilization

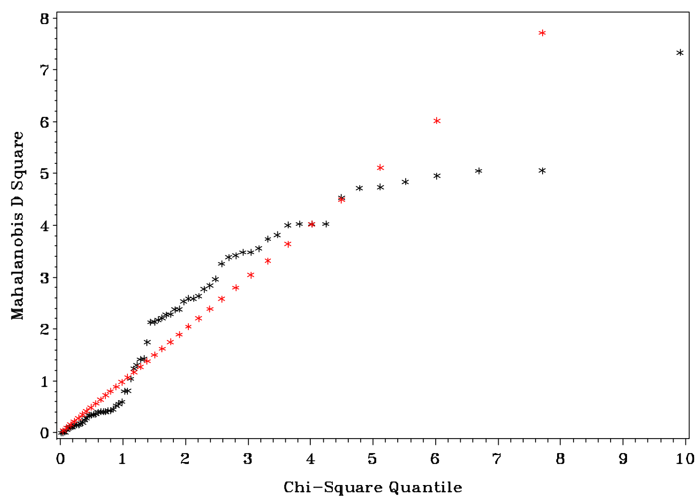

Figure 2: Q-Q plot for assessing normality with 2 variables: credit_limit, number_of_trades

But when outliers are present in the data, *Discriminant Analysis* does not seem to handle them as well as *Generalized Logit Regression*. For example, if we change the last observation in build sample to “1900 100 0.20 O”, and go through the same model building exercise as before, we can see that error rate for *Discriminant Analysis* is 8.45%, while it is 7.04% for *Generalized Logit Regression*. A closer look at the Q-Q plot (Figure 3) indicates that the last observation is an outlier with significant large ‘Robust Mahalanobis D Square’. In this case, “PROC STEPDISC” still considers both “credit_limit” and “number_of_trades” as significant predictors and hence results in higher error rate. On the other hand, “PROC LOGISTIC” only finds “credit_limit” to be significant and minimizes the effect of outlier.

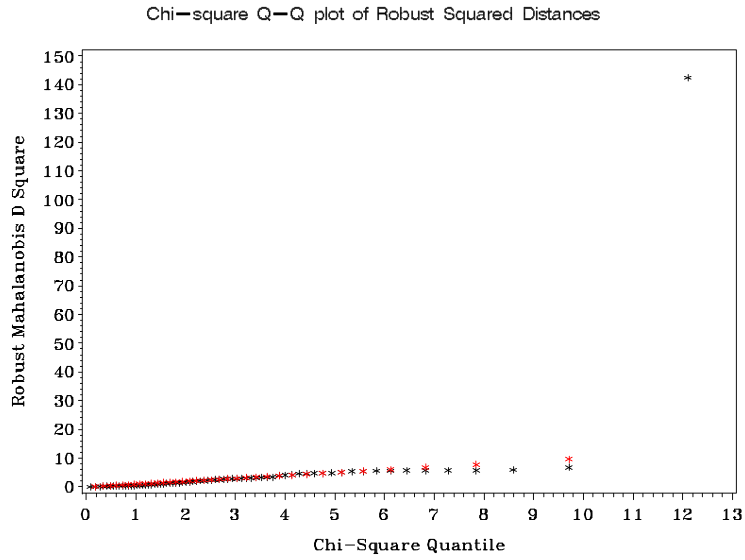

Figure 3: Q-Q plot for assessing outliers with 3 variables: credit_limit, number_of_trades, utilization

CONCLUSION

Generalized Logit Regression is generally more robust than *Discriminant Analysis*, however, assumptions should not be the only reason for choosing *Generalized Logit Regression* over *Discriminant Analysis*. As always, thorough understanding of the problem should provide guidance on which methodology to choose.

REFERENCES

1. Stokes, M.E., Davis, C.S. and Koch, G.G 2000. Categorical Data Analysis Using The SAS System. SAS Institute and Wiley
2. Goldstein, M. and Dillon, W.R.(1978). Discrete Discriminant Analysis. New York: Wiley.
3. Hand, D.J. (1982). Kernel Discriminant Analysis. Chichester: Research Studies Press.
4. Tabachnick, B.G. and L.S.Fidell. 1996. Using Multivariate Statistics. Harper Collins College Publishers: New York.
5. Khattree, R. and Naik, D 1999. Applied Multivariate Statistics with SAS software. SAS Institute and Wiley

ACKNOWLEDGMENTS

The author would like to express her appreciation to the following parties:

Capital One Financial Service for making this opportunity possible
John Zhan for his review and encouragement of this paper
Lu Su for her review of this paper and continuous support
Joerg Rothenbuehler for his help on conducting discriminant analysis with SAS

RECOMMENDED READING

1. http://www.statsoft.com/textbook/stathome.html
2. http://www.med.monash.edu.au/spppm/research/rda/Comparisons%20across%203%20methods.htm

CONTACT INFORMATION

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

Name: Jin Li
Enterprise: Capital One Financial Service
Address: 15030 Capital One Drive
City, State ZIP: Richmond, VA 23238
Work Phone: (804) 284-5136
Fax:
E-mail: jin.j.li@gmail.com
Web: