Paper 92

%DISCIT Macro: Pre-screening continuous variables for subsequent binary logistic regression analysis through visualization

Mohamed S. Anany, Videa (Cox Media Group subsidiary)

ABSTRACT

In binary logistic regression, when there are many independent variables that could potentially be included in the model, it is always a good practice to perform bivariate analysis between the dichotomous variable (dependent) and the independent variables. The paper presents a procedure to identify candidate continuous variables by performing several bivariate analyses and summarizing it into a graphical panel. The analysis is based on a two-sample *t* -test, and a graphical panel to visualize the relationship between the continuous variable and the dichotomous dependent variable. The analysis recodes the continuous variable(s) into two different ordinal forms, adjusts their scale through odds and log odds transformations if needed, and collapses similar buckets of the recoded form to improve their linear relationship if exists. Furthermore, we make use of the information value as a measure of the predictive power of the independent variable in capturing the dichotomous variable. The %DISCIT macro makes this prescreening process easier and less time consuming for analysts.

INTRODUCTION

When building binary logistic regression models, the analyst usually ends up with a large number of variables that needs to be prescreened. This becomes a tedious and time consuming task for SAS programmers. One approach is to perform bivariate analysis between the dichotomous variable (Y) and the independent variables (Xs). Independent variables come in many forms; binary, continuous, nominal categorical, or ordinal categorical variables. This paper is concerned with identifying candidate continuous variables only. The %DISCIT macro produces a graphical panel to visualize the relationship between the dependent variable, and continuous predictor. The graphical representation will help determine the kind of relationship (increasing /decreasing, linear/nonlinear) between the dependent variable and the continuous variable. Furthermore, it gives us some insight on whether we need to use the continuous variable in its original form in the model or if we need to apply a transformation and/or re-code it. The reason behind doing these transformations is to improve the relationship between the response and the predictor. Ultimately, the significance of this improvement will be determined when the different forms of the variable are introduced in the model where their predictive powers are compared.

PRESCREENING

OVERVIEW

The prescreening process involves a two-sample *t* –test of the continuous variable grouped by the dichotomous variable as a measure of the independent variable significance. Besides testing for significance, the macro will produce side by side box-plots and two-way comparative histograms to visualize the distribution of the continuous variable distribution across the two values of the dependent variable (1 and 0). LOWESS (locally weighted scatterplot smoother) smoothing is also used as a graphical assessment of linearity of the original form of the continuous variable in the logit. Another important part of the prescreening process is the discretization process described in the following section.

DISCRETIZATION

First the continuous variable is transformed into two ordinal forms by binning it into groups based on 1) user defined cut points and 2) groups of equal sizes (usually 10). The mean of the dependent variable along with the confidence interval is calculated for each group. The means of the dichotomous is then plotted against the 2 forms of the binned continuous variable (1st and 2nd plot in the left column of the panel).

LOWESS SMOOTHING

Another method of visualizing the relationship between dichotomous and continuous variables is through smoothed scatterplots. The LOWESS smoothing is helpful as a graphical assessment of linearity of the original form of the continuous variable in the logit. It is also used as a tool for identifying extreme observations that could influence the assessment of linearity. This method requires computing a smoothed value for the response variable for each subject that is a weighted average of the values of the outcome variable over all subjects. This is done in SAS using PROC LOESS. The smoothed values are plotted against the log-odds in the 3rd plot in the left column of the panel.

COMPARATIVE HISTOGRAMS

The 1st and 2nd plot in the right column are comparative histograms of the continuous variable by the dichotomous variable. These two plots facilitate the visual comparison of the distributions of the two groups (1 and 0) through comparing side-by-side histograms. Each plot contains summary statistics of the continuous variable by the group (Number of observations, Mean, and standard deviation).

TWO-SAMPLE T-TEST AND SIDE-BY-SIDE BOXPLOTS

The 3rd and final plot in the second column shows side-by-side boxplots of the continuous variable by the dichotomous variable a long with two-sample two sided t test results (t-statistic and p-value). The null hypothesis is that the mean of the continuous variable is not different across the values of the dichotomous variable. The two-sample t –test is an alternative analysis which is nearly equivalent at the univariable level to fitting a univariable logistic regression model, and it is maybe preferred in an applied setting.

POST DISCRETIZATION

COLLAPSING SIMILAR BINS

Both ordinal forms of the continuous variable undergo a diagnostic process. To easily explain the process, we will refer to the equal size bins and user-defined bins as "groups" since both forms of the variable go through the same process. After the mean of the dichotomous variable is calculated for each group, the dichotomous variable mean within a group and the mean within the consecutive group are tested using T-tests for a significant statistical difference. If the two groups are significantly different, no action is required and we move to the next consecutive groups. If there is no statistical difference, the two groups are collapsed into the higher one, and the next t test is performed using the new collapsed groups. The t-tests need to be less conservative so a default alpha of 0.15 is used (The %DISCIT macro allows you change the value of alpha using Alpha= parameter). The purpose behind collapsing groups is to improve the relationship between the transformed form of the variable and the dependent variable by making it more linear; hence improving its predictive power. The t-tests are optional in the DISCIT macro and are printed by specifying ttest in the options= parameter. Also, by specifying outtest in the options= parameter, the macro will output a dataset *"TTestResults"* with the results of the ttests.

Although the t-tests approach aids the decision making process of whether to collapse the groups or not, it is sample size dependent. For example, in case of a very large data set the t tests detect tiny (and sometimes impractical) differences. Alpha level that is more conservative than 0.15 is recommended. In a case of a small dataset, the power of the tests is low and a more liberal alpha value is recommended (0.15-0.25). In practice, besides the p-values of the t-tests, the analyst should consider (1)the graphical representation of the relationship, along with (2) the value of t-statistic when determining the "practicality" of the differences and making a decision whether to collapse two groups or not.

After the groups are collapsed, the odds and log odds are calculated for each of the two collapsed ordinal forms of the variable. Because the predictors are linear in the log of the odds, it is often helpful to transform the continuous variables to create a more linear relationship. The transformations are optional in the %DISCIT macro by specifying Transformit in the options= parameter.

WOE AND INFORMATION VALUE

Another common transformation used mostly in the credit and financial industries when building binary logistic models to predict the risk of default is the weight of evidence transformation (WoE). WoE value is a widely used measure of the "strength" of a grouping for separating good and bad risk. By specifying WOE keyword in the options= parameter, the %DISCIT macro will calculate the weight of evidence for the two ordinal forms of the variable before collapsing and output them in a dataset "*WOEResults*" in the work library. Along with the WoE, the macro will print a list of the information values (IV) arranged in descending order for the all variables that entered the macro. IV is a number that attempts to quantify the predictive power of the independent variable in capturing the dichotomous variable.

EXAMPLES AND DEMONSTRATIONS

EXAMPLE 1: USING A SMALL DATA SET AND ONE CONTINOUS VARIABLE

In this example, using the %DISCIT macro, we will evaluate the relationship between the age in years (AGE/continuous variable) and presence or absence of evidence of significant Coronary Heart Disease (CHD/dichotomous variable) for 100 subjects in a hypothetical study of risk factors for heart disease.

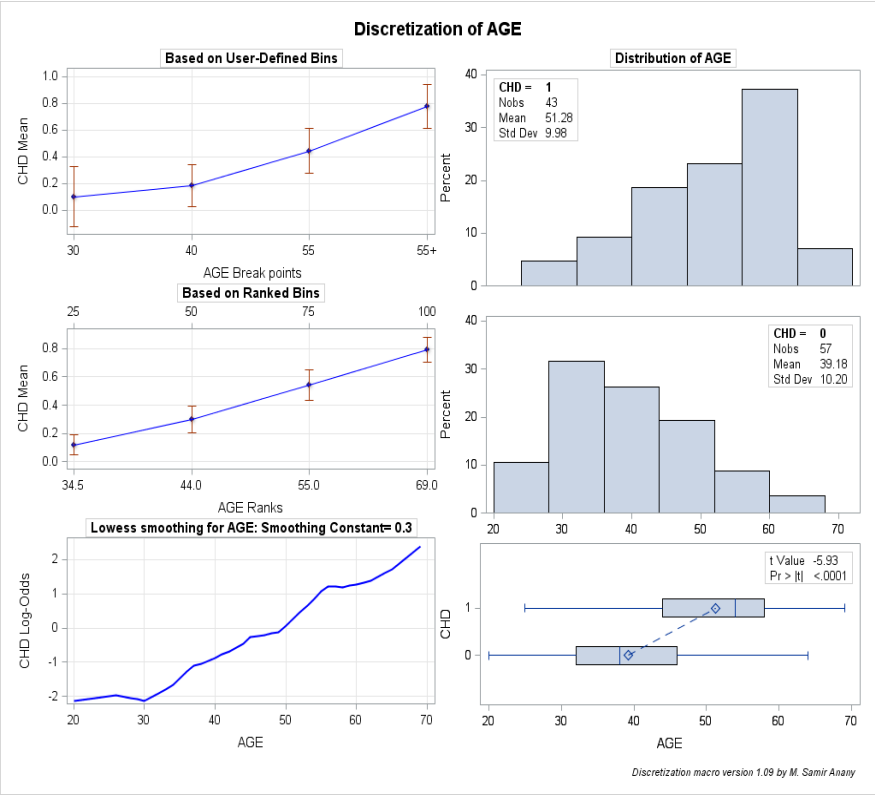

Figure 1. Graphical Panel of Age

We can see, figure 1, in the three plots in the first column that as age increases, the risk of getting CHD increases. The comparative histograms show a left skewed distribution of Age for subjects with CHD with an average age of 51 and right skewed distribution of Age for subjects without CHD with an average age of 39; the older you get the higher the risk of having CHD. The side by side boxplots confirms the previous conclusion with a significant t-statistic at alpha 0.05. We would definitely include Age as a candidate variable in a subsequent logistic regression model.

The macro invocation for example 1 is as follows:

```
%DISCIT(InDS=chdage, /*Input dataset*/
        dep=CHD,     /*Binary dependent variable*/
        vars=AGE,    /*list of continuous variables*/
        Breakpts=30  40  55, /*list of custom binning points*/
        SmoothConst=0.3,     /*LOWESS smoothing constant*/
        Groups=4,          /*Number of equal size bins*/
        Alpha=0.15);       /*alpha level for collapsing t-tests*/
```

EXAMPLE 2: USING A LARGE DATA SET AND MORE THAN ONE CONTINOUS VARIABLE

In this example, the dichotomous variable, DELQ_ID, indicates whether a customer is low (0) or high (1) credit risk. The data consists of 1.25 million customers (observations) and has more than 300 variables. Only two continuous variables are entered into the macro in this example for demonstration purposes; TADB and RBAL. TADB is a ratio of total debt burden to average debt burden, while RBAL is the total balance on open revolving trade.

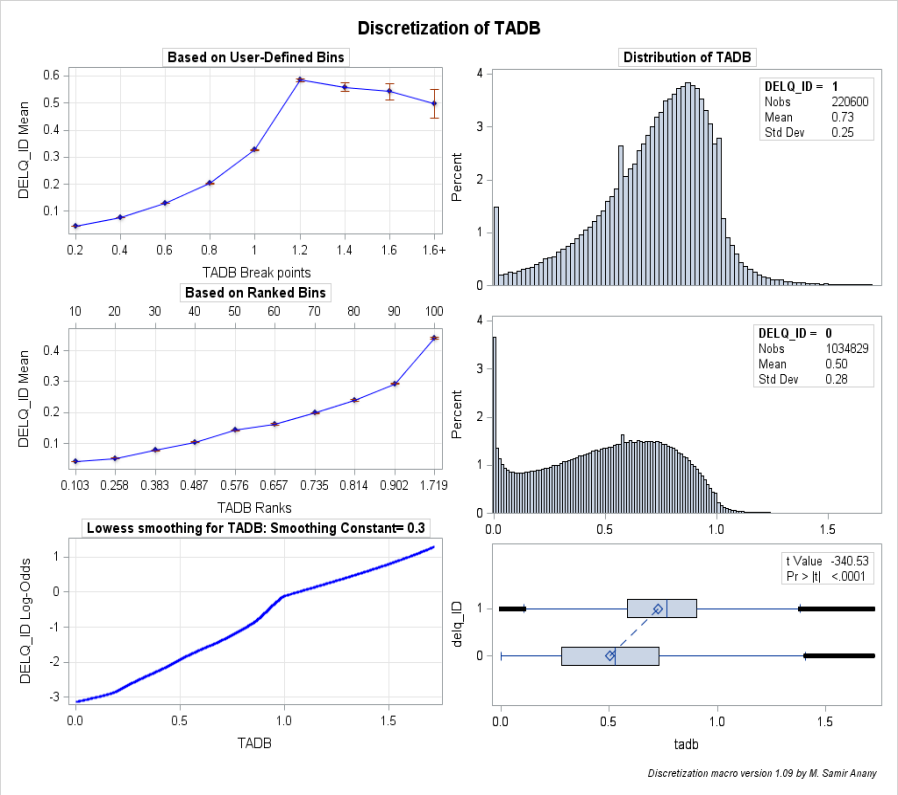

Figure 2. Graphical Panel of TADB

Clearly high risk customers tend to have higher ratio of total debt burden to average debt burden (Figure 2). As the ratio (TADB) increases, the proportion of high risk customers increases consistently. The ranked bins and lowess smoothing plots confirm the consistency of the relationship. The user-defined bins plot show an inflection point at 1.2 where the relationship starts to decrease. The post t-tests for ordinal form of TADB based on custom categories in Table 1, however, show that the risk is not significantly different for ratios above 1.2 and should be grouped together. The 2-sample t-test results (t-statistic -340.53 and p-value <0.001) indicates that TADB is highly significant.

t-Tests for TADB Custom categories

Category	Tested Against Category	Category Frequency	Against Category Frequency	Mean	Against Mean	t-Statistic	p-value
1	2	200995	194671	0.04453	0.07614	-41.92	<.0001
2	3	194671	270139	0.07614	0.12993	-58.65	<.0001
3	4	270139	317285	0.12993	0.20196	-73.79	<.0001
4	5	317285	237351	0.20196	0.32592	-105.85	<.0001
5	6	237351	29750	0.32592	0.58514	-89.40	<.0001
6	7	29750	3829	0.58514	0.55785	3.22	0.0013
7	8	3829	1081	0.55785	0.54209	0.92	0.3574
8	9	4910	328	0.55438	0.49695	2.02	0.0429

Table 1. Post t-tests for ordinal form of TADB based on custom categories

The graphical panel of RBAL in figure 3 shows that the proportion of high risk customers does not change that much over the values of RBAL. The comparative histograms show that RBAL has almost the same distribution for high and

low risk customers. The p-value of the t-test is significant but given the large sample size we ought to look at the t-statistic rather than the p-value which tells us that the significance is not that strong. Also, the ranked bins and lowess smoothing show that above 1000 the proportion does not change that much.

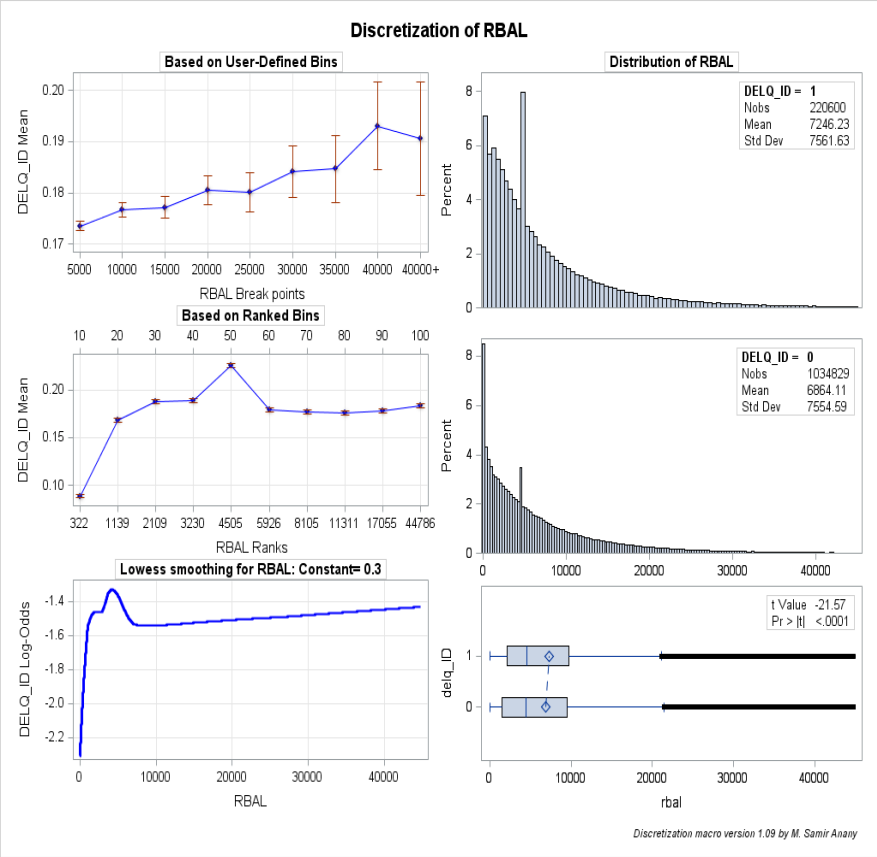

Figure 3. Graphical Panel of RBAL 1

Table 2 shows the information value summary for the variables entered in the macro (TADB and RBAL). IVs for both ordinal forms of TADB (0.3 +: Strong) indicates that they are strong predictors of DELQ_ID, while the IV of the ranked ordinal form of RBAL (0.02-0.1: Weak) shows that it is a weak predictor. The IV of the customized ordinal form of RBAL (<0.02: Unpredictive) indicates that it is unpredictive of DELQ_ID.

Information Value Summary

Variable	Discrtization Type	Information Value
TADB	Ranked	0.66552
TADB	Customized	0.65993
RBAL	Ranked	0.06202
RBAL	Customized	0.00047

Arranged in descending order

Table 2. IV summary for TADB and RBAL 1

The macro invocation for example 2 is as follows:

```
%DISCIT(InDS=CPR,
        dep=delq_ID,
        vars= RBAL TADB,
        Breakpts=  5000 10000 15000 20000 25000 30000 35000 40000 |
                   0.2 0.4 0.6 0.8 1 1.2 1.4 1.6,
        SmoothConst=0.3,
        Groups=10,
        Alpha=0.001,
        Options=Printtest transformit ttest WOE,
        Out= Tansformed_Data)
```

IMPLICATIONS

There are some points that the analyst needs to be aware of when using the %DISCIT macro. If the graphs show nonlinear relationship, it is usually hard to guess the function that will linearize the relationship. Another important point that should be highlighted is that all the relationships are bivariate; hence they do not take into account the confounding effect of other variables in the model. Also, when creating user-defined bins for a very skewed variable, most of the observations will be captured by one bin leaving the remaining bins to only capture a few outlying observations. On the other hand, by creating equal size bins (1/nth of the sample size), the range of each bin can vary greatly.

CONCLUSION

Prescreening variables for modeling is a long and exhausting process that involves a lot of coding. The %DISCIT macro makes this process shorter and much easier for analysts by accommodating more than one variable in a single macro call and let analysts devote more time to the modeling stage. %DISCIT macro is available upon request by contacting the author via email.

REFERENCES

- Maggie Zhang, Shaolin Chen, Suzanne C. Rain. “Evaluating Continuous Variable Transformations in Logistic Regression”. *SAS Conference Proceedings: Midwest SAS User Group 2004*
- David W. Hosmer, JR. and Stanley Lemeshow (2013) Applied Logistic Regression: Third Edition, Wiley
- Paul D. Allison (2012) Logistic Regression Using SAS: Theory and Application, Second Edition, SAS Institute Inc
- SAS Institute, Inc. SAS/STAT User's Guide, Version 9.3 (online at http://support.sas.com/), SAS Institute Inc
- Moez Hababou, Alec Y. Cheng, Ray Falk and Bridgeport. “Variable Selection in the Credit Card Industry”. *SAS Conference Proceedings: Northeast SAS User Group 2006*

Contact Information

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

Name: Mohamed Samir Anany, MSc
Enterprise: Videa (Cox Media Group subsidiary)
Address: 3390 Peachtree Road NE
City, State ZIP: Atlanta, GA 30326
E-mail: Samir.anany@videa.tv