**Analysis of Amount of the Given Credit for Credit Card Clients**

***Analysis of Education Level of Credit Card Clients (Individual)***

**STAT 448, Final Project**

***Wenke Huang, Group 8***

**1 Introduction:**

**1.1 Data Set Description:**

The data set we chose describes the customers’ default payment information in Taiwan. There are 30,000 observations and 24 variables in the raw data set. The first column is Identification Number. Of course, the original data set will be read into SAS and cleaned before we fit our Discriminant Analysis model.

1. X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2. X2: Gender (1 = male; 2 = female).
3. X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
4. X4: Marital status (1 = married; 2 = single; 3 = others).
5. X5: Age (year).
6. X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; …; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 0 = payment delay for less than one month; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
7. X12 - X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September,2005; X13 = amount of bill statement in August, 2005; …; X17 = amount of bill statement in April, 2005.
8. X18 - X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19= amount paid in August, 2005; …; X23 = amount paid in April, 2005.

The original data set can be obtained from the following link:

<http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

**1.2 Project Goal**

Our goal is to find out whether we can use an individual’s personal information and credit card usage historical information to predict his or her education level. Most of the time, when applying for a credit card from a credit card company in Taiwan, we will always be asked to provide the information about our education level. This information will be considered as one of the metrics of risk management of the credit card company. However, customers will not be asked to show their diploma when answering this question, which means the credit card company cannot confirm whether the customer tell the truth or not.

The Discriminant Analysis will be attempted in this project. And we know that our model may not work perfectly in our first attempt. Further modification and model re-fitting will be applied based on the results of the previous attempts. If we can build a model which works successfully, the credit card company can use this model to predict and detect whether the future customers lie on their application form or to find out the potential customers.

**2 Preparation**

**2.1 Data Cleaning**

The observations which contain values against the above description are considered as the invalid ones and removed from the data set. Based on the definition of Discriminant Analysis which is the method we are about to use, all of the predictors should be continuous variables, and at the same time we only focus on the Education Level here. As a result, some of the constraints can be totally ignored, which can save us some time. In general, the constrains based on the data set description which will be applied for our data cleaning is shown here.

1. Repayments in these six month must be an integer ranging from -1 to 9;
2. Education Level must be an integer ranging from 1 to 4;
3. Marriage Level must be an integer ranging from 1 to 3;
4. Sex Level must be either 1 or 2, luckily there is no observation against it;
5. There should be no missing values within the any valid observation.

What’s more, some more additional constraints which are not provided by the data set description are also applied. These constraints were proposed based on the real world view.

1. The bill amounts of each month must be less than their corresponding given credit;
2. Any individual must keep using their credit card in these six continuous months in a normal way, which means that none of the bill or payment amount is less than or equal to 0.

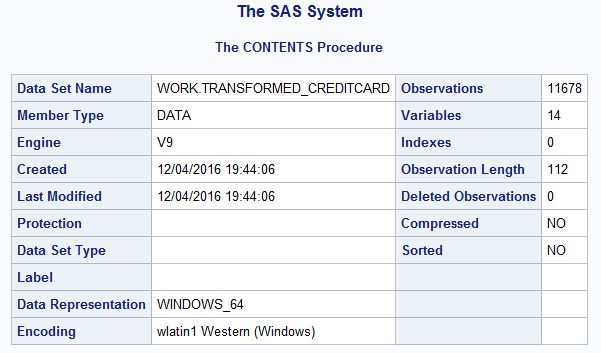
The data set after cleaning contains 11,678 observations and 24 variables.

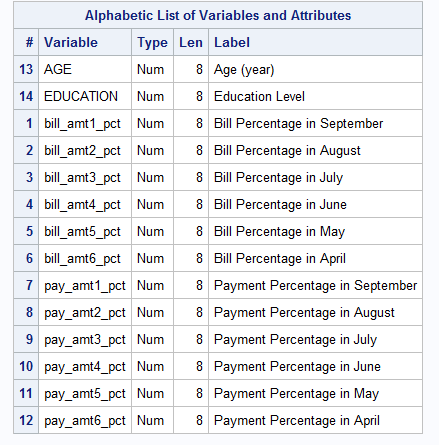
**2.2 Transformation**

We realize that different individuals will have different given credits. For example, the given credit of a day job employee should be OK, but the given credit of Terry Gou who is the third richest businessman in Taiwan will be a totally different story. Hence, if we just treat the amounts of bill and payment as our predictors, those records which stand for rich people will totally influence the whole model. However, it is a little subjective if we just set a line and remove the observations with given credits greater than the line.

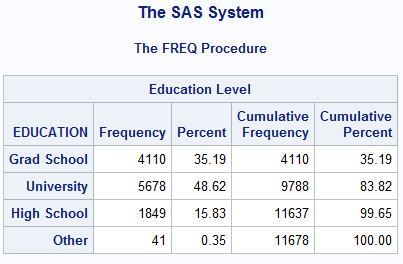
As a consequence, we here would like to focus on the percentage of the given credit that an individual would like to spend and pay back. The new variables were created based on the original variables. New Variables = Old Variables / Given Credit. All of the 12 variables which include 6 payment history and 6 bill statement records are transformed here, and the 12 old ones will be dropped after the transformation is done. What’s more, variable AGE will also be contained without being transformed.

The final data set we are about to use contains 11,678 observations, 13 predictor variables and 1 response variable EDUCATION.





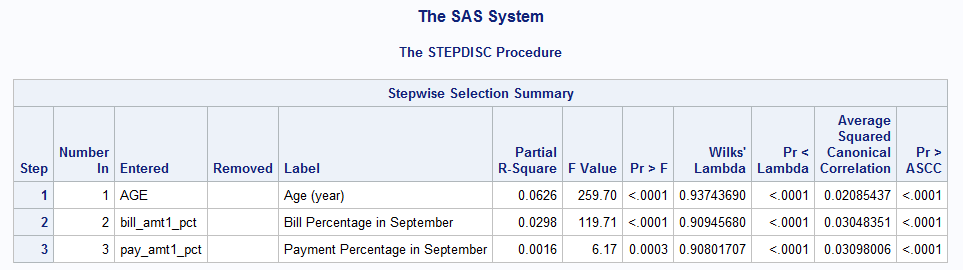
The frequency table of different EDUCATION groups is shown below and may help us have a better understanding about the response which we would like to focus on in this case. We can read that all of the three groups Grad School, University and High School contain thousands of observations while group of Other contains only 41 points.



**3 Model Fitting – First Attempt**

**3.1 Variable Selection**

First, “proc stepdisc” procedure will be applied to find out the best predictors we will use in our final model. The criterions for both staying in and entry into model are set to be 0.05 here.

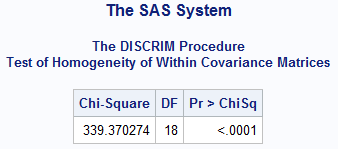


In the output, it is clear that three variables which are AGE, bill\_amt1and pay\_amt1 were selected into our final model.

**3.2 Test of Homogeneity of within Covariance Matrix**

Here, we only focus on two kinds of Discriminant Analyses, Linear Discriminant Analysis and Quadratic Discriminant Analysis. The difference between these two kinds of analyses is that whether the covariance matrixes of different groups are equal to each other or not. If they are all equal to each other, we should use Linear Discriminant Analysis. If they are different from each other, we should and must use Quadratic Discriminant Analysis.

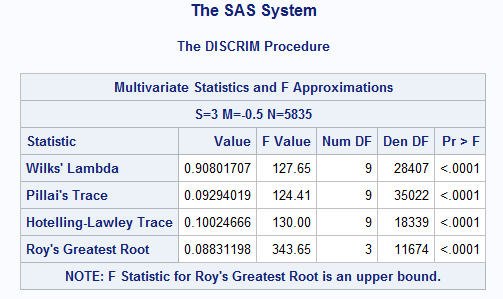
In order to find out which kind of Discriminant Function should be applied here, Test of Homogeneity of within Covariance Matrix is performed here.



Based on the table, we read the Chi-Square is 339.370274 which is pretty large. And its corresponding p-value is less than 0.0001 which is definitely less than 0.05. Since the Chi-Square value is significant at the 0.05 level, the within covariance matrices will be used in the Quadratic Discriminant Function.

By default, SAS will tell us whether Chi-Square value is significant at the 0.1 level. In this case, it does not matter because our Chi-Square value is significant at both the 0.05 and 0.1 levels.

**3.3 MANOVA**



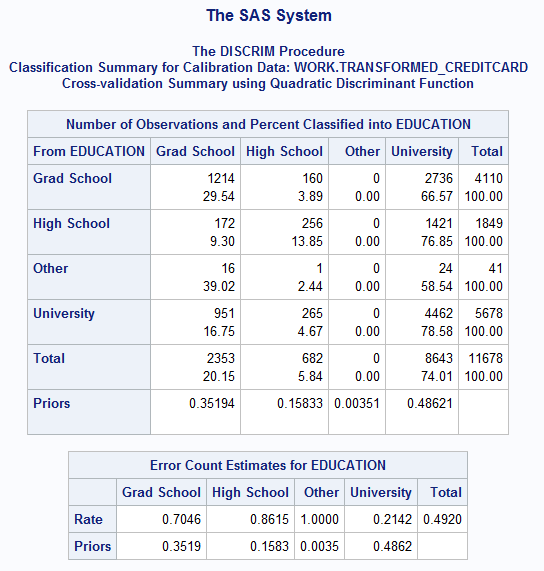
We can read from the above MANOVA table that the values of four statistics Wilks’ Lambda, Pillai’s Trace, Hotelling-Lawley Trace and Roy’s Greatest Root are 127.65, 124.41, 130.00 and 343.65, respectively. And their corresponding p-values are all less than 0.0001 which are absolutely less than 0.05. It shows that all of them are statistically significant, which tells us that the possibility of discriminating between different education levels based on predictors which we have selected are high.

**3.4 Model Fitting**

Leave-one-out cross-validation is used here. In cross-validation, we only treat only part of the data set as our training set to fit the model, the rest points will be considered as our testing set to test the model which we have fitted. Most of the time, we just divide the original training data set into k groups evenly or almost evenly and we would like to fit the model k times. Each group will be treated as testing data set one by one while the rest corresponding k-1 groups will be used to fit our model. After k times of fitting, the mean value of the k testing errors will be considered as our cross-validation error. Leave-one-out cross-validation means that every time we only treat one different point of the whole data set as our testing data set and the rest observations will be used to train our model.

Basically speaking, I think cross-validation is an effective way in our model fitting.

Apart from the cross-validation, the “priors proportional” statement is also applied here. This statement means that we set the prior probabilities proportion to the sample size. Basically we have no idea the real prior probabilities proportion, but we know that if we set them to be equal, it is definitely preposterous. Here, our data set is pretty large, so we may think that the prior probabilities proportion is equal to the sample size approximately.



The result of cross-validation is shown in the above two tables. We can realize that the model prediction is not satisfactory at all. The misclassification error rates for Grad School, High School, Other and University are 0.7046, 0.8615, 1.0000 and 0.2142, respectively. It tells us that about 70% of Grad School points, 86% of High School points and 100% of Other points were misclassified into a different group. If we re-read the contingency table again, we may find that, 66.57% of Grad School points, 76.85% of High School points and 58.54% of Other points were misclassified into the group of University, which is not good at all.

**3.5 Conclusion of First Attempt**

We may think that this attempt is not satisfactory at all. And this model which we just fit may not be considered as a useful one even though the misclassification error rate of University is 21.42% which is not bad. Further model may be attempted to see if we can still try to figure out a way to find the relationship between them.

**4 Model Fitting – Second Attempt**

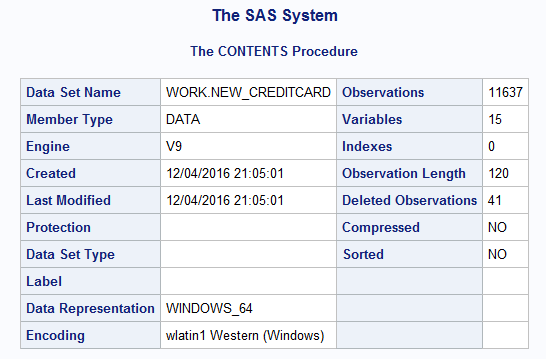
**4.1 Data Set Re-generation**

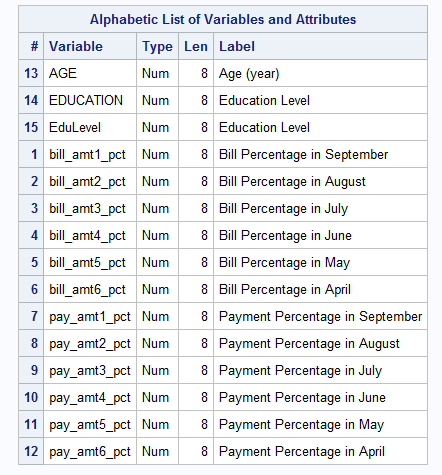
We can re-take a look at the frequency table and cross-validation result contingency table. We may notice that the group of Other contains only 41 observations which is way less than observation numbers of other groups and it takes only 0.3511% of the whole training data set. Especially in the cross-validation result contingency table, we may also find that none of the 41 points were assigned into the group of Other. It is totally understandable, because “Other” is not a real exited education level in the real world. It contains observations just because the customers would not like to answer this question because of his or her own habits or some other reasons which we do not know.

Here, we would like to delete all of the 41 points which stand for the group of Other EDUCATION level.

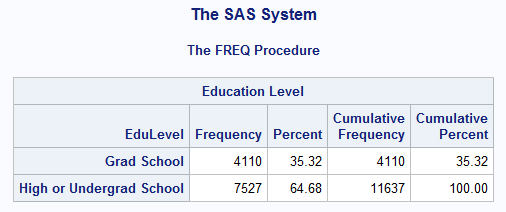
Here, we would like to generate a new categorical variable EduLevel to be our new response. If the EDUCATION is an integer either 2 or 3 which stands for University and High School, respectively, the new variable EduLevel is set to be 2 which stands for “High or Undergrad School” here. If the EDUCATION is an integer 1 which stands for Grad School, the EduLevel is set be the same value 1, the meaning of which is also the same.

The new data set we are about to use contains 11,637 observations, 13 predictor variables, 1 new response variable EduLevel and 1 previous response variable EDUCATION.



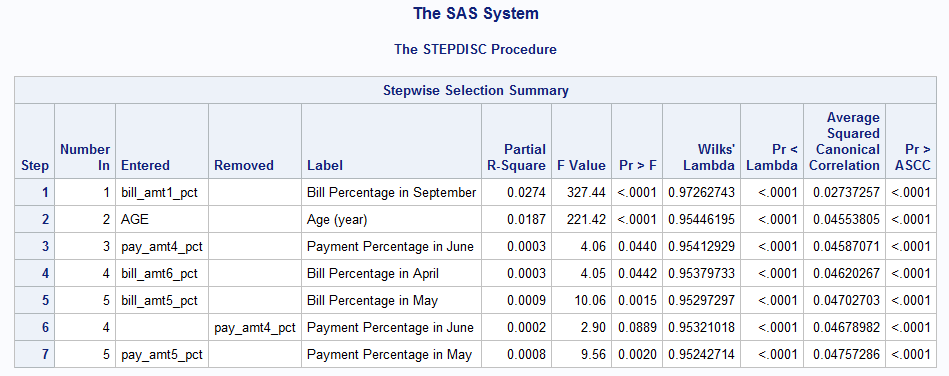


The frequency table of different EduLevel groups is shown below and may help us have a better understanding about the new response which we would like to focus on in our next attempt. We can read that the two groups Grad School and High or Undergrad School contain 4110 and 7527 observations, respectively.



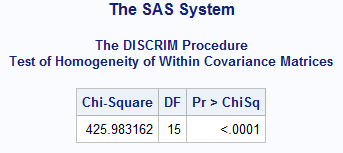
**4.2 Variable Selection**

Like previous part, “proc stepdisc” procedure will be applied to find out the best predictors we will use in our final model. The criterions for both staying in and entry into model are set to be 0.05 here, as well.



Here, the result of variable selection is little different compared with the previous part. We have 5 variables left in this part. They are, bill\_amt1\_pct, AGE, bill\_amt6\_pct, bill\_amt5\_pct, pay\_amt5\_pct. Variable AGE and bill\_amt1\_pct were included in both models of our first and second attempt.

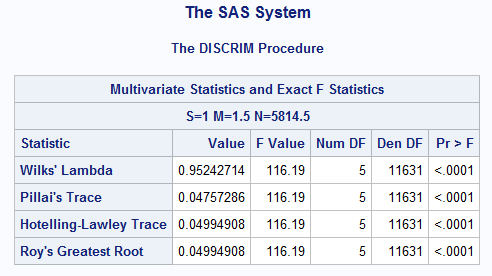
**4.3 Test of Homogeneity of within Covariance Matrix**



Like the previous part, we need to perform the Test of Homogeneity of within Covariance Matrix as well to determine what kind of Discriminant Function and Analysis we should use to fit our second model. We cannot just skip this part, because our data set has been updated before we fit this model.

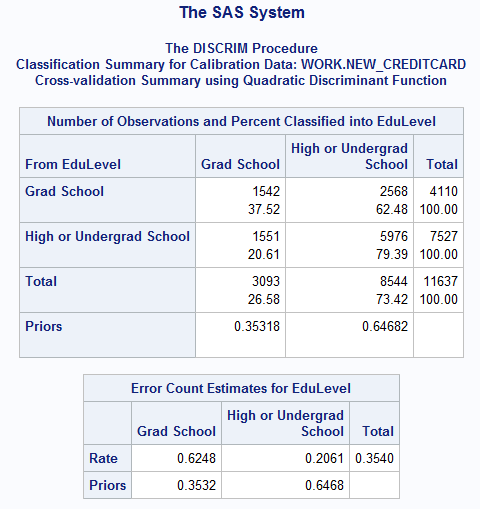
Based on the table, we read the Chi-Square is 425.983162 which is also pretty large. And its corresponding p-value is less than 0.0001 which is definitely less than 0.05. Since the Chi-Square value is significant at the 0.05 level, the within covariance matrices will be used in the Quadratic Discriminant Function.

**4.4 MANOVA**



Here, we can read from the above MANOVA table that the values of four statistics Wilks’ Lambda, Pillai’s Trace, Hotelling-Lawley Trace and Roy’s Greatest Root are 116.19, 116.19, 116.19, 116.19, respectively. And their corresponding p-values are all less than 0.0001 which are absolutely less than 0.05. It shows that all of them are statistically significant, which tells us that the possibility of discriminating between different education levels based on predictors which we have selected are still high.

**4.5 Model Fitting**



Like previous part, both leave-one-out cross-validation and “priors proportional” statement are also performed here. In this case, we only have two groups here, Grad School and High or Undergrad School.

The result of cross-validation is shown in the above two tables. We can realize that the prediction of our second attempt is not satisfactory either. The misclassification error rates for Grad School is 0.6248 which is still high. At the same time, 20.61% of the points of the High or Undergrad School points were misclassified into the group of Grad School.

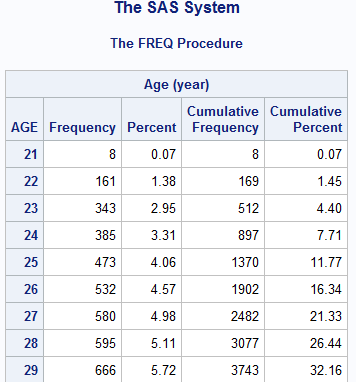
**4.6 Conclusion of Second Attempt**

We may think that our second attempt is not satisfactory, either. And this model which we just fit may not be considered as a useful. After the first two attempts, we can start thinking that maybe the education level of a customer has nothing to do with his personal information and credit card usage historical records.

We should stop here, and try to have some feedbacks or assumptions about why these models do not work well. It is because the limitation of the Discriminant Analysis, we did not select the most useful predictors, or some other possible reason.

**5 Possible Explanations and Recommendations**

**5.1 Categorical Predictor**



If we re-take a look at the variables which we used here, we may find an interesting one AGE. Basically we just consider AGE as a numerical continuous variable, because everyone is growing older and older every day or even every second. Part of the frequency table of variable AGE in our second attempt data set has been printed above, which tells us that all of the AGE values are integers here. Technically speaking, they are still a categorical variable instead of being a numerical one, which means that variable AGE is against the constraints of the Discriminant Analysis actually.

The frequency table of the data set of our first attempt tells us the same story that the variable AGE should be considered as a categorical variable technically. This data set only contains 41 more points than the data set of our second attempt. The frequency table does not change too much and will not be shown here.

Adding variable AGE may be a bad idea here, or some necessary and possible transformation methods can also be applied onto AGE, too.

**5.2 Data Set Historical Background**

We know that the data set contains the information about Taiwan which is a country of east hemisphere and with different culture. We all know that after World War II and China Civil War, the economy of Taiwan developed in a very fast speed. But there still are some different culture and habits which we ignored here. For example, credit card swiping is still not very popular over there. People in Taiwan still prefer to pay everything in cash rather than take their credit cards out of their wallets.

Generally speaking, we may and should collect some more information about Taiwan, and do some further analysis about their education, habits, or some other important and related area before we start further attempts.

**6 Conclusion**

In this case, we have tried to use Discriminant Analysis to find out if we could predict the customers’ education levels based on their ages and their credit card usage records. Unfortunately, it does not seem to be a plausible way. Because the results of both attempts are not satisfactory at all. In the first model, 70.46% points of Grad School, 86.15% of High School, all points of Other and 21.42% points of University were misclassified. Most of points which are from the first three groups were misclassified into the group of University. In the second model, 62.48% points of Grad School and 20.61% points of High and Undergrad School were misclassified. The misclassification errors were pretty large and not satisfactory at all.

The Discriminant Analysis which is Quadratic Discriminant Analysis in this case is not useful to help us predict the education level based on the corresponding individual’s age and credit card usage historical records. it gave us a really confusing result that education level has nothing to do with the information which is covered by the predictors we selected and mentioned above in both attempts. We can also conclude that it is not possible and plausible to predict the education level of an individual who is currently living in Taiwan just based on his or her age and credit card usage history record.