CATÓLICA PORTO BUSINESS SCHOOL

PORTO

Master in Management

Course: Regression and Data Analysis

Professor: Pedro Duarte Silva

Logistic Regression and Log-linear Analysis

“Taiwan-Customers Default Payments”

Inês Amorim nº 355419069

José Miguel Teixeira nº 355419031

SangHoon Han nº 357919153

11 May 2020

Index

Index of Figures

Index of Tables

Introduction

In the context of the course Regression and Data Analysis, we pretend to develop a study of association and interdependence of qualitative variables using logistic regression and log-linear models about the database “Taiwan customers default payments”. The database contains information on default payments, demographic factors (gender, age, marriage), history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. Having the goal of applying the knowledge gained during this course, this report will include insights about the models we used, how we select them, and how we screen the variables, and the analysis we did in order to predict the default payment for the next month, that is our dependent/response variable.

Context

Nowadays the use of credit cards in Taiwan is widespread throughout the entire population, with 25 million active credit cards in 2015 (for a population of 23.4 million at the time). However, that was not always the case. Beginning in 1990, the Taiwanese government allowed the formation of new banks, and the credit card and cash card business became the biggest “cash cow” for these new banks. Some banks even allowed young people (mainly university students) to apply for credit cards, and many cardholders became the so-called "credit card slaves," a term referring to those who could only pay the minimum balance on their card debt every month.

Back in 2005, credit card issuers in Taiwan faced a cash and credit card debt crisis. In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit cards for consumption purposes, and accumulated heavy cash and credit card debts. This crisis caused a blow to consumer financial confidence and presented a big challenge for both banks and cardholders, and so this assignment is aimed at the case of customers default payments in Taiwan.

Database and variables

The database “Taiwan-Customers default payments” gives us information regarding the default payment of credit card clients in Taiwan from April 2005 to September 2005, with 4991 observations and 17 variables. Before proceeding to the

descriptive and graphical analysis of our data, we needed to prepare the database, so we eliminated the first column of the "Row" (ID# of each client, from 1 to 5000) and we confirmed that we didn't have any missing values. Note that the currency used is the New Taiwan Dollar (NT$ or TWD), where 1 TWD is roughly 0.03EUR.

The response/dependent variable is named "Default" as refers to the default payment for the next month, it's binary and when it is equal to 1 (Yes) means that the client will be in default for the payment next month. The 16 explanatory/predictor variables are:

1. LIMIT_BAL (Amount of the given credit, including individual consumer credit and his/her family (supplementary) credit, from 10.000 to 1.000.000);
2. Gender (1=male, 2=female);
3. MARRIAGE (referring to the marital status, where 1=married, 2=single, 3=others);
4. AGE (customers age in years, from 21 to 75);
5. BILL_AMT1 (Amount of bill statement in September 2005);
6. BILL_AMT2 (Amount of bill statement in August 2005);
7. BILL_AMT3 (Amount of bill statement in July 2005);
8. BILL_AMT4 (Amount of bill statement in June 2005);
9. BILL_AMT5 (Amount of bill statement in May 2005);
10. BILL_AMT6 (Amount of bill statement in April 2005);
11. PAY_AMT1 (Amount paid in September 2005);
12. PAY_AMT2 (Amount paid in August 2005);
13. PAY_AMT3 (Amount paid in July 2005);
14. PAY_AMT4 (Amount paid in June 2005);
15. PAY_AMT5 (Amount paid in May 2005);
16. PAY_AMT6 (Amount paid in April 2005).

Firstly, we performed a descriptive statistical analysis of the numeric variables, as seen in Table 1, and frequency analysis for the categorical variables Gender and Marriage (Figure 1).

Variables:	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**IQR**	**Std**	**Coef. Variation (CV)**
LIMIT_BAL	10.000	50.000	140.000	165.640	230.000	1.000.000	180.000	130.442,1	0,7875
AGE	21	28	34	35.36	41	75	13	9,265295	0,2620
BILL_AMT1	-14.386	3.212	21.494	50.275	62.781	964.511	59.568	74.884,01	1,4894

BILL_AMT2	-30.000	2.945	20.610	48.312	60.547	983.931	57.601,5	72.368,51	1,4979
BILL_AMT3	-15.000	2.442	19.538	45.405	56.593	578.971	54.251	68.180,97	1,5016
BILL_AMT4	-170.000	1.807	18.009	40.822	49.309	891.586	47.502	64.059	1,5692
BILL_AMT5	-28.335	1.494	17.363	39.584	49.166	927.171	47.671,5	61.655,59	1,5576
BILL_AMT6	-339.603	976	15.874	38.002	48.035	961.664	47.059	61.518,28	1,6188
PAY_AMT1	0	1.000	2.108	5.556	5.000	368.199	4.000	14.812,11	2,6661
PAY_AMT2	0	639	2.000	5.433	4.998	344.261	4.359	16.294,09	2,9993
PAY_AMT3	0	219	1.402	4.602	4.000	896.040	3.781	18.014,61	3,9149
PAY_AMT4	0	232	1.500	4.745	4.000	497.000	3.768	15.140,04	3,1910
PAY_AMT5	0	208	1.500	4.765	4.000	332.000	3.792	14.862,62	3,1193
PAY_AMT6	0	0	1.319	5.276	4.000	528.666	4.000	19.763,54	3,7456

Table 1: Descriptive statistics of the numeric variables

The amount of given credit ranges from 10.000 to 1.000.000 new Taiwan dollars, and the average age of the credit card customers is 35 years. To measure the dispersion of the numeric variables, we calculated their standard deviation (Std), the interquartile range (IQR=Q3-Q1) that considers the absolute dispersion, and the coefficient of variation (CV) that is a relative measure for dispersion, showing the deviation as a percentage of the mean. Note that the mean and the standard deviation are very sensitive to outliers, so usually the median is better for location when there are outliers, which is the case. Variables with a lower CV are less dispersed than variables with a higher CV, and when the CV is above 1 it means that the standard deviation of a variable is higher than its mean.

Gender	
Male (=1)	**Female (=2)**
2.133 (42,7%)	2.858 (57,3%)

Marriage (Marital status)		
Married (=1)	**Single (=2)**	**Others (=3)**
2.214 (44,4%)	2.712 (54,3%)	65 (1,3%)

Figure 1: Frequency distribution of variables Gender and Marriage

Both genders are well represented on the database in terms of proportion, and it covers the different marital status, with a slightly higher percentage for single persons, followed by the married ones. There are 3.884 clients with no default payments on their credit cards (that represents around 77% of total clients) and a remaining of 1.107 clients that do have default payments for the next month (when the variable “Default” is equal to 1). In figure 2 we can see how this variable is “divided” according to the marital status of the clients.

Figure 2: Default payment for next month according to the marriage status

In addition, we also did a correlation analysis of the numeric variables to see how they relate to each other. By look over the graphic (Figure 3), we can conclude that the variables representing the amount of bill statement (BILL_AMT) from April 2005 (BILL_AMT6) to September 2005 (BILL_AMT1) are strongly and positively correlated, meaning that the amount of bill statement of one month is related with the amount of the previous month, which makes sense.

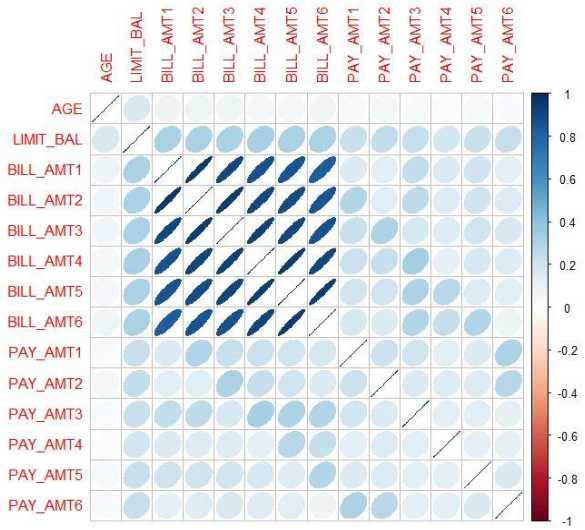

Figure 3: Correlation graphic of the numeric variables

Logistic Regression Analysis

Model Development

Logistic regression is used to form prediction models, and in this case, we want to predict if the credit card clients will have default payment for next month or not, which is our dependent binary variable. We want to know the impact of the predictor variables

in the default payment, so our first model with logistic regression includes all the explanatory variables, where model 1 is:

Default = β0 + β1 AGE + β2 LIMIT_BAL + β3 BILL_AMT1 + β4 BILL_AMT2 + β5 BILL_AMT3 + β6 BILL_AMT4 + β7 BILL_AMT5 + β8 BILL_AMT6 + β9 PAY_AMT1 + β10 PAY_AMT2 + β11 PAY_AMT3+ β12 PAY_AMT4 + β13 PAY_AMT5 + β14 PAY_AMT6 + β15 Gender + β16 MARRIAGE

In model 1, the variables that are statistically significant are LIMIT_BAL, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT4 and MARRIAGE2 (when the marital status goes from married to single).

We did a stepwise search for the second model based on the Akaike information criterion (AIC) that tries to find the model that predicts the best. Doing so, the second model dropped the variables AGE, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT3, PAY_AMT5, PAY_AMT6 and Gender. The AIC went from 5120,949 (model 1) to 5109 (model 2). For model 3 we dropped the variable BILL_AMT2 from model 2 and the AIC is 5109,3. The model 4 (variable MARRIAGE dropped) is: *Default = β0 + β1 LIMIT_BAL + β2 BILL_AMT1 + β3 BILL_AMT3 + β4 PAY_AMT1 + β5 PAY_AMT2 + β6 PAY_AMT4* ; and the AIC is equal to 5121,9.

Then, and although this is an informal measure of diagnostic, we calculated the variance inflation factor (VIF) of model 3 to see if the predictors maintain their typical relation (no problems of multicollinearity). To remember, model 3 is:

Default = β0 + β1 LIMIT_BAL + β2 BILL_AMT1 + β3 BILL_AMT3 + β4 PAY_AMT1 + β5 PAY_AMT2 + β6 PAY_AMT4 + β7 MARRIAGE

According to VIF, the variables BILL_AMT1 and BILL_AMT3 are correlated with each other. In model 5 we consider the model 3 without the BILL_AMT1, and model 6 is equal to the model 3 without BILL_AMT3. Finally, model 7 is an update of model 3 without both BILL_AMT1 and BILL_AMT3. To decide our final model, we will compare the 7 different models according to their AIC.

	AIC
Model 1	5120,949
Model 2	5109,036
Model 3	5109,309
Model 4	5121,946
Model 5	5115,543
Model 6	5132,431
Model 7	5150,812

Table 2: Akaike Information Criterion (AIC) of the 7 logistic regression models

A lower value for AIC among all models indicates the model with a better fit , so based on that, and as we can see in table 2, the model that predicts better the default payment (lower AIC) is the model 2, followed close by model 3 and by model 5.

In order to try to identify the best and final model, we performed an ANOVA chi-square test (goodness of fit test) with model 1 (initial model with all predictor variables), model 2, model 3 and model 5. The p-value is well below 0.01 (we reject the null hypothesis) and statistically significant for model 5, meaning that model 5 is the model that predicts better and it would be our final model.

Interpretation of the results

The final logistic regression model (model 5) is:

Default = β0 + β1 LIMIT_BAL + β2 BILL_AMT3 + β3 PAY_AMT1 + β4 PAY_AMT2 + β5 PAY_AMT4 + β6 MARRIAGE

In the first place, it is important to note that all the predictor variables are statistically significant and with p-values very close to 0, with the exception of MARRIAGE3, and this makes sense because other marital status besides being married or single only represents 1,3% of our database. When the variable MARRIAGE changes from being a married person (=1) to being single (=2), the log odds of the default payment next month decrease by 0.2851.

The estimated coefficient of BILL_AMT3 is 3.889e-06, which means that for 1 unit increase in the amount of bill statement from July, the log odds of being in default payment for next month increase by 0.00000389 (ceteris paribus). The estimated coefficient of LIMIT_BAL is -1.626e-06, which means that for 1 unit increase in the amount of given credit in TWD, the log odds (probability) of being in default payment for next month decrease by 0.00000163 (and the other explanatory variables remain constant).

In what regards the PAY_AMT variable, the ones present and significant to our final model are referring to the months of September (PAY_AMT1), August (PAY_AMT2) and June (PAY_AMT4). For a 1 unit increase in the amount paid in August with credit card, the log odds of being in default payment for next month decreases by 0.00003876. The variables PAY_AMT1 and PAY_AMT2 have higher mean than the other PAY_AMT variables, which makes sense because people tend to go on vacations and spend more in credit cards in August and September, by booking hotels, airplane

tickets and so one. The database tells us that they tend to pay more on these months, and so the probability of being in default payment for next month gets reduced.

In order to measure the performance of our final model, we used AUC (Area under the curve) ROC (Receiver Operating Characteristics) curve, as we can see in figure 4.

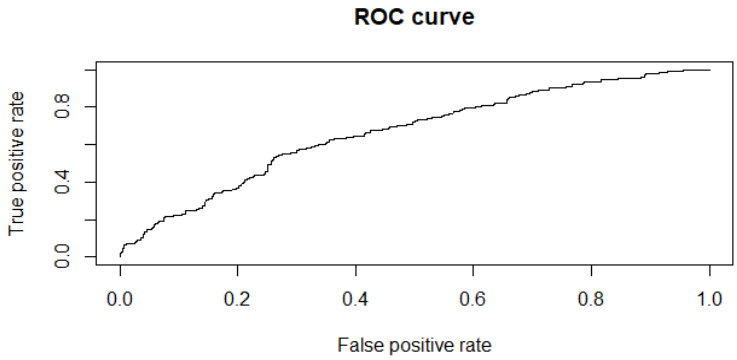

Figure 4: ROC Curve

ROC is a probability curve and AUC represents the degree of separability. It tells how good is the capability of the model in distinguishing between classes. Higher the AUC (ideally close to 1), better the model is at predicting, and in our case, at predicting if the payment for next month regarding the credit card will be in default or not. In general, an AUC of 0.5 suggests no discrimination (i.e., poor ability to distinguish between default payment or not based on the test). The AUC for our final model is equal to 0.6648843, which is considered acceptable discrimination (close to 0.7).

Log-linear Models

Log-linear analysis is a technique used in statistics to examine the relationship between more than two strategic variables. This one, when using Logistic Regression, does not attempt to adjust the models, but to test a model fit quality test, assuming that it is well adjusted from the beginning. The main goal of Log-Linear Analysis is, above all, to study the dependency relationships between categorical variables or, in the terminology of the software used, factor-type variables.

In order to perform an analysis with a greater degree of clarity, we considered three of the variables of the initial database. These ones were, obviously, the three categorical ones: *MARRIAGE*, *Gender* and *Default*. Subsequently, given that the initial database does not present the desirable format, we converted it into a contingency table.

Contingency Table

In the first place, and before proceeding with the construction of the model, we performed an independence test to the contingency table. The test resulted in a p-value of 0.01302, being this one very close to 0, meaning the we can reject the null hypothesis of existing perfect independence between the variables. This being said we can verify there is some sort of dependence between the **marriage status,** the **gender,** and the **default payment for next month**.

SELECTION OF THE MODEL

In the Log-Linear Analysis, the principle of parsimony must be respected, and, as such, we should follow the model that is the simplest and, simultaneously, that best explains the variance of the frequencies in the contingency table. In order to achieve this goal, it is necessary to calculate models iteratively.

Independence Model

This is the simplest model, which presents a p-value of 0.01326 and a Chisq of 17.72761. In addition to that, the values of the residuals of the independence model were evaluated too, that is, the differences between the expected values and the observed values. From Figure 5, it is concluded that all absolute variations less than 2 are considered "noise" and are, therefore, random, and non-significant variations. On one hand, the visual representation of the model expresses that the number of married male customers with no default payment next month, is lower than expected. On the other hand, the number of male married customers with "Yes" as an answer to default payment next month is higher than expected.

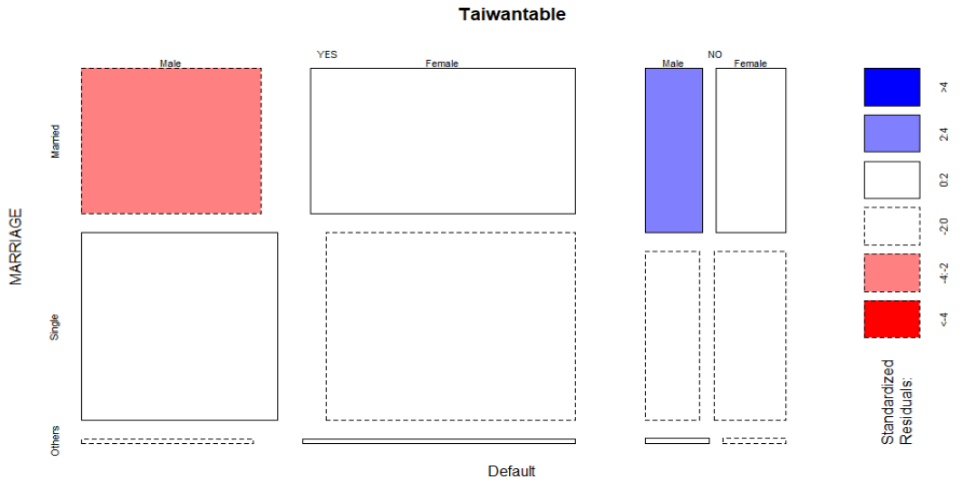

Figure 5: Residuals: Independence model (Mosaic Plot)

Saturated Model

The saturated model, as expected, has a p-value of 1 and, consequently, a Chisq of 0, being this the interaction that best explains the data, since it includes all the components of the model, however will be of little use statistically, as you have no data left to estimate variance.

Homogeneous Model

The model of homogeneous associations was also created, considering the dependence of variables two by two. This model has a p-value greater than 5% and, therefore, its use proves to be adequate.

Three Conditional Models

Finally, an interdependence test was performed on three models where two variables were conditionally independent of the third variable. One of them presented a proof value below 0.05 and therefore should not be considered and the other two present a p-value of over than 0.05, meaning that they could be considered.

Final Model

As the homogeneous model is the most parsimonious of those considered adequate and one of the simplest, this will be the final model to be adopted, despite we could use two of three conditional models:

```
> loglm(~ Default*MARRIAGE + Default*Gender + Gender*MARRIAGE,data=Taiwantable)
Call:
loglm(formula = ~Default * MARRIAGE + Default * Gender + Gender *
    MARRIAGE, data = Taiwantable)

Statistics:
                     X^2 df  P(> X^2)
Likelihood Ratio 2.831220  2 0.2427775
Pearson          2.833936  2 0.2424480
```

Figure 6: R output of homogeneous model

Conclusion

With this project it is intended to deepen our knowledge about logistic regression and log-linear models, applying what we have learned during the Regression and Data Analysis course. The database chosen “Taiwan-Customers default payments” allows us to take insights regarding the probability of the credit card customers being in default

payment next month or not, that is our dependent variable. We started our analysis with descriptive statistics of the variables, making an exploratory analysis of our database, followed by the logistic regression analysis and then the log-linear models.

Relatively to the logistic regression analysis, we conclude that the variables that have higher probability of affecting the default payment next month are the LIMIT_BAL (amount of the given credit), BILL_AMT3 (amount of bill statement in July 2005), PAY_AMT1, PAY_AMT2 and PAY_AMT4 referent to the amount of previous payments (September, August and June 2005, respectively), and finally the variable MARRIAGE. The gender and age of the credit card customers doesn't seem to affect the probability of them being in default with their credit card payments for the next month.

Regarding the Log-Linear Analysis, we conclude that there is a very low level of dependency between the variables *Default, MARRIAGE and Gender* and that the homogeneous model would be the considered as the most adequate model for our analysis.

Bibliography

- Jian, S. (2016), Analyzing Default Payments of Credit Card Clients in Taiwan, retrieved in 08/05/2020 from https://www.researchgate.net/publication/311714926_Analyzing_Default_Payments_of_Credit_Card_Clients_in_Taiwan
- Hsiung Chang, C., Chang, H., Tien, J. (2017), A Study on the Coping Strategy of Financial Supervisory Organization under Information Asymmetry: Case Study of Taiwan's Credit Card Market, Universal Journal of Management 5(9): 429-436
- Tien, Y. , Lin, L. (2015), Nearly 25 million active credit cards in Taiwan, Focus Taiwan English News, retrieved in 08/05/2020 from https://focustaiwan.tw/business/201511220009
- Wang, E. (n.d.), Taiwan's Credit Card Crisis, Seven Pillars Institute, retrieved in 08/05/2020 from https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/