Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/25/20230482

Different Factors the Banks Can Use to Evaluate and Screen the Customers' Credits

Yuang Li[1,a,*]

[1]*Accounting school, Shanghai University of International Business and Economics, Shanghai, 201620, China*
a. 2423171941@qq.com
**corresponding author*

Abstract: The global economy has become increasingly volatile due to the facts such as wars and epidemics. Therefore, banks may suffer even higher risks from credit card default. This may cause a vicious circle. This paper aims to help the banks evaluate and screen the customer's credits based on different factors. Statistical methods such as ANOVA analysis or Binary logistic regression are used to study and assess the significance of other variables. The results demonstrate that several factors, such as given credit or marital status, are significant to the credit card default, and education may be considered the most critical factor.

Keywords: customers' credits, credit card, the global economy

1. Introduction

There was an outbreak of epidemics in 2020 [1], while in 2022, Russia-Ukraine Conflict suddenly happened. These are all black swan events that have an extremely high impact on the global economy [2, 3]. For example, due to the epidemics, many people were sick and could not go to work. And because of the war, insufficient resources supply led the inflation. As a result, many people may have a shortage of money. Eventually, this problem will be passed on to the bank. The risk of credit card defaults is becoming higher and higher. This will cause the bank to fall into a shortage of capital cash flow, exacerbating the possibility of a financial crisis and trapping it in a vicious cycle. Therefore, at particular times, the bank should check the customers' credit even carefully to avoid credit card default as much as possible.

In this work, we use statistical methods to find the significance or importance of different variables for credit card default. Also, we compare and analyze other variables to find the essential factors the bank should consider when the bank evaluates and screen the customers' credits.

2. Description of Data

The data for this paper, credit card customer default, is sourced from Kaggle. The data set has employed a binary variable, default payment (Yes = 1, No = 0), as the response variable and 23 variables as explanatory variables, including records of 30000 customers. The description of all features can be seen as follows (Table 1).

DOI: 10.54254/2754-1169/25/20230482

Table1: Features description.

Variables	Description
X1	Amount of the given credit
X2	Gender (1 = male; 2 = female)
X3	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
X4	Marital status (1 = married; 2 = single; 3 = others)
X5	Age (year)
X6-X11	History of past payments. Six months of records are collected (April to September). X6 = the repayment status in September; X7 = August; . . .; X11 = April. -1,1,2,3 show the time the customer delays paying, and -1 means to pay on time.
X12-X17	Amount of bill statement. X12 = amount of bill statement in September; X13 = August; . . .; X17 = April.
X18-X23	Amount of last payment. X18 = amount paid in September; X19 = August; . . .; X23 = April.
Default	Default. Payment. Next.month: Default payment (1=yes, 0=no)

3. Methods

3.1. Chi-square Analysis

In the statistical test, Chi-square is often used to analyze the relationship between groups of certain data types, such as satisfaction and age. In other words, it is often used to recognize whether discrete variables are significantly correlated. We should first determine whether the p-value (Pearson Chi-square) shows significance. The values are primarily compared with 0.05. If it shows effectiveness, it means that the two groups of variables are significantly correlated. Also, you will get the table. The rows usually represent the response variables, while the columns generally represent the explanatory variables. Then, you can analyze and get results based on these two things [4]. Here, we give the null hypothesis: factors (Discrete variables) are insignificant to the credit card default.

3.2. Normality Tests & T-tests & Non-parametric Test

Normality is the basic premise of many analysis methods, especially continuous variables. Other analysis methods should be chosen if the normality trait is not satisfied. If it is normality, we use t-tests; if it is not, we use non-parametric tests.

The t-test is often used to analyze the continuous variables or quantitative data and the relationship between the qualitative data and them. The non-parametric test is nearly the same. The only difference is that it is only used when the data does not pass the normality tests [5]. Here, we give the null hypothesis: factors (continuous variables) are insignificant to the credit card default.

3.3. One-way ANOVA

One-way ANOVA is used to analyze whether the qualitative things affect the quantitative stuff. In other words, it examines variance, focusing on the changes in different conflicts [6]. Here, we give the null hypothesis: factors are insignificant to the credit card default.

DOI: 10.54254/2754-1169/25/20230482

3.4. Binary Logit Regression

Binary logit regression is often used to analyze and study the relationship between the effect of X on Y, where X is usually quantitative data (if X is a fixed class of data, a dummy (dummy) variable setting is generally required). Y is a dichotomous selected class of data (the number of Y must be only 0 and 1). Before the analysis, the fitness of the model should be found. The cut value is often used to balance recall and precision. Through the study, the relationship between several predictor variables and the response variables can be known. Also, the magnitude of the correlation can be found by regression coefficients [7].

3.5. Factor Analysis

Factor analysis is used when there are a lot of different variables. The variables can be divided into groups through factor analysis, often called reducing the dimensions. In other words, it is an analysis method to condense information into a few keyword descriptions [8]. Each group has similar characteristics. They can be analyzed why they are in the same group and must have some underlying relationships between them. Factors are often extracted based on the Scree Plot or the root value greater than 1.

4. Results

In Chi-Square Tests (Table 2), the Asymptotic significance(2-sided) of Pearson Chi-Square is 0.000, which is less than 0.05, so we should reject the null hypothesis. This indicates that education is significant to credit card default.

Through the cross table (Table 3), about 20% of people with graduate degrees (1) in their group will default their credits, followed by people with university degrees (2), which is nearly 23%. Meanwhile, the people with high school degrees account for almost 25% of their group. These figures tell us that people with a higher education status have less possibility of having a default payment.

Table 2: Chi-Square tests education.

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	163.217[a]	6	.000
Likelihood Ratio	184.708	6	.000
Linear-by-Linear Association	23.529	1	.000
N of Valid Cases	30000		
a. 1 cells (7.1%) have an expected count of less than 5. The minimum expected count is 3.10.			

Table 3: Crosstab education.

		EDUCATION							Total
		0	1	2	3	4	5	6	
Default.payment.next.month	0	14	8549	10700	3680	116	262	43	23364
	1	0	2036	3330	1237	7	18	8	6636
Total		14	10585	14030	4917	123	280	51	30000

DOI: 10.54254/2754-1169/25/20230482

In Chi-Square Tests (Table 4), the Asymptotic significance(2-sided) of Pearson Chi-Square is 0.000, which is less than 0.05, so we should reject the null hypothesis. This indicates that the marriage is significant to the credit card default.

Interestingly, through the cross table (Table 5), about 23% of people who are married (1) in their group will default their credits, while nearly 20% of single people will default their credits. This is contrary to our prediction that when people are married, they are less likely to default on the credits.

Table 4: Chi-Square tests marriage.

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	35.662[a]	3	.000
Likelihood Ratio	36.609	3	.000
Linear-by-Linear Association	17.771	1	.000
N of Valid Cases	30000		
a. 0 cells (0.0%) have an expected count of less than 5. The minimum expected count is 11.94.			

Table 5: Crosstab marriage.

		MARRIAGE				Total
		0	1	2	3	
Default.payment.next.month	0	49	10453	12623	239	23364
	1	5	3206	3341	84	6636
Total		54	13659	15964	323	30000

In Chi-Square Tests (Table 6), the Asymptotic significance(2-sided) of Pearson Chi-Square is 0.000, which is less than 0.05, so we should reject the null hypothesis. This indicates that the history of past payments is significant to the credit card default.

Table 6: Chi-Square tests history of past payment.

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	5365.965[a]	10	.000
Likelihood Ratio	4563.158	10	.000
Linear-by-Linear Association	3164.623	1	.000
N of Valid Cases	30000		
a. 3 cells (13.6%) have an expected count of less than 5. The minimum expected count is 1.99.			

As shown in Table 7, the figures 0.000 of the Kolmogorov-Smirnov, less than 0.05, mean that all the four continuous variables LIMIT_BAL, AGE, BILL_AMT1, and PAY_AMT1 are not significant, so we should reject the null hypothesis.

DOI: 10.54254/2754-1169/25/20230482

Since it is not normally distributed, the nonparametric tests are done (Table 8). The figures, 0.000, 0.000,0.000, less than 0.05, show that the factors LIMIT_BAL, BILL_AMT1, and PAY_AMT1 are significant to the credit card default, and we should reject the null hypothesis. However, figures of the age were 0.373, exceeding 0.05, showing no significance between AGE and credit card default, so we should retain the null hypothesis.

Table 7: Tests of normality.

	Default.payment.next.month	Kolmogorov-Smirnov[a]		
		Statistic	df	Sig.
LIMIT_BAL	0	.102	23364	.000
	1	.158	6636	.000
AGE	0	.095	23364	.000
	1	.095	6636	.000
BILL_AMT1	0	.233	23364	.000
	1	.250	6636	.000
PAY_AMT1	0	.363	23364	.000
	1	.361	6636	.000
a. Lilliefors Significance Correction				

Table 8: Tests summary.

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of LIMIT_BAL is the same across categories of default.payment.next.month.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
2	The distribution of AGE is the same across categories of default.payment.next.month.	Independent-Samples Mann-Whitney U Test	.373	Retain the null hypothesis.
3	The distribution of BILL_AMT1 is the same across categories of default.payment.next.month.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
4	The distribution of PAY_AMT1 is the same across categories of default.payment.next.month.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .050.				

DOI: 10.54254/2754-1169/25/20230482

Through the ANOVA analysis (Table 9), the figure 0.000, which is less than 0.05, means that the education status is significant to the LIMIT_BAL. These results are also supported by the sum of squares between and within groups. Also, based on Figure 1, we know that with a higher education status, people will have a higher LIMIT_BAL.

Table 9: ANOVA.

LIMIT_BAL					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	36380981205725.420	6	6063496867620.903	388.068	.000
Within Groups	468635854803311.100	29993	15624840956.334		
Total	505016836009036.560	29999			

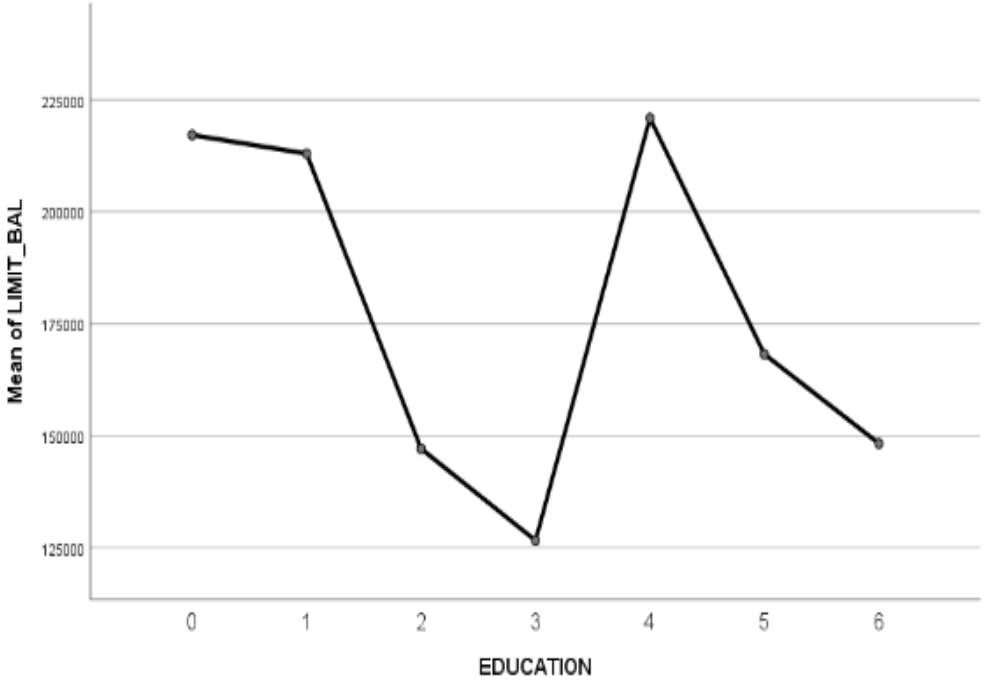

Figure 1: A higher education status, people will have a higher LIMIT_BAL.

In binary logit regression, all twenty-three factors are put into the model, and the cut value is 0.220 (Table 10) to balance the precision and recall results. As the table 11 shows, factors LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_5, BILL_AMT1, BILL_AMT3, PAY_AMT1, PAY_AMT2, PAY_AMT4, PAY_AMT5 are all significant to the credit card default. LIMIT_BAL, SEX, EDUCATION, MARRIAGE, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT4, PAY_AMT5 are significantly and negatively correlated with credit card default while the AGE, PAY_0, PAY_2, PAY_3, PAY_5, BILL_AMT3 are significantly and positively correlated with credit card default.

DOI: 10.54254/2754-1169/25/20230482

Table 10: Regression.

Variables in the Equation							
Step 17[p]	LIMIT_BAL	.000	.000	28.044	1	.000	1.000
	SEX	-.108	.031	12.350	1	.000	.898
	EDUCATION	-.103	.021	24.012	1	.000	.903
	MARRIAGE	-.155	.032	24.002	1	.000	.856
	AGE	.008	.002	17.791	1	.000	1.008
	PAY_0	.582	.018	1086.368	1	.000	1.789
	PAY_2	.082	.020	16.435	1	.000	1.085
	PAY_3	.085	.020	17.606	1	.000	1.089
	PAY_5	.052	.018	8.819	1	.003	1.054
	BILL_AMT1	.000	.000	38.378	1	.000	1.000
	BILL_AMT3	.000	.000	17.002	1	.000	1.000
	PAY_AMT1	.000	.000	38.513	1	.000	1.000
	PAY_AMT2	.000	.000	33.121	1	.000	1.000
	PAY_AMT4	.000	.000	5.334	1	.021	1.000
	PAY_AMT5	.000	.000	5.146	1	.023	1.000
	Constant	-.690	.119	33.857	1	.000	.501
a. Variable(s) entered on step 1: PAY_0.							
b. Variable(s) entered on step 2: LIMIT_BAL.							
c. Variable(s) entered on step 3: PAY_3.							
d. Variable(s) entered on step 4: BILL_AMT1.							
e. Variable(s) entered on step 5: PAY_AMT1.							
f. Variable(s) entered on step 6: MARRIAGE.							
g. Variable(s) entered on step 7: EDUCATION.							
h. Variable(s) entered on step 8: AGE.							
i. Variable(s) entered on step 9: BILL_AMT2.							
j. Variable(s) entered on step 10: PAY_2.							
k. Variable(s) entered on step 11: PAY_AMT2.							
l. Variable(s) entered on step 12: SEX.							
m. Variable(s) entered on step 13: PAY_5.							
n. Variable(s) entered on step 14: PAY_AMT5.							
o. Variable(s) entered on step 15: PAY_AMT4.							
p. Variable(s) entered on step 16: BILL_AMT3.							

As the Scree Plot shows (Figure 2), there is a steep down between factor 4 and factor 5, so we use five elements to extract from all the 23 factors.

DOI: 10.54254/2754-1169/25/20230482

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy figure is 0.799, exceeding 0.5, while Bartlett's Test of Sphericity is 0.000, which is lower than 0.5 (Table 11). These two figures show a correlation between variables, and the factors analysis is meaningful.

Then we again put the five factors into the Binary logistic regression to see the new model's compatibility. We still use 0.22 as the cut value (Table 12), using the forward method: conditional. However, according to the model summary and Nagelkerke R Square (Table 13), this model can only explain 14.1 of the credit card default.

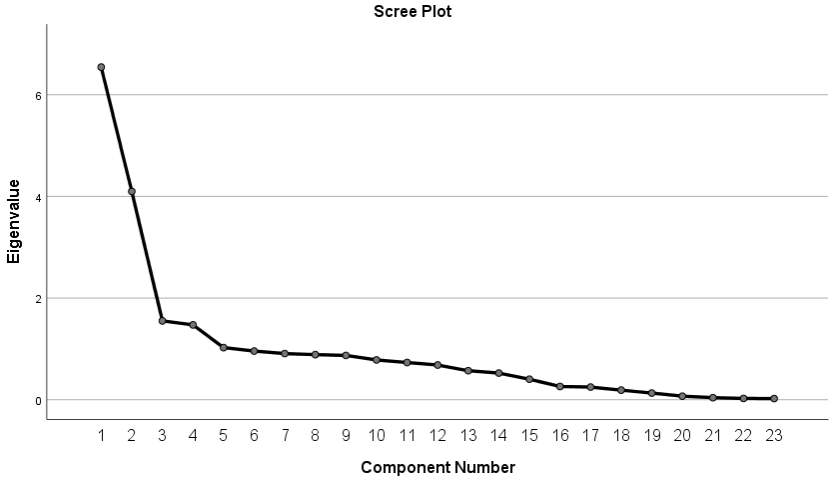

Figure 2: Scree plot.

Table 11: KMO and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.799
Bartlett's Test of Sphericity	Approx. Chi-Square	570911.655
	df	253
	Sig.	.000

Table 12: Classification table.

			Predicted		
	Observed		default.payment.next.month		Percentage Correct
			0	1	
Step 1	default.payment.next.month	0	14454	8910	61.9
		1	2396	4240	63.9
	Overall Percentage				62.3
Step 2	default.payment.next.month	0	13890	9474	59.5
		1	2247	4389	66.1
	Overall Percentage				60.9
Step 3	default.payment.next.month	0	14150	9214	60.6
		1	2254	4382	66.0
	Overall Percentage				61.8
Step 4	default.payment.next.month	0	14397	8967	61.6

DOI: 10.54254/2754-1169/25/20230482

		1	2284	4352	65.6
	Overall Percentage				62.5
Step 5	default.payment.next.month	0	14593	8771	62.5
		1	2320	4316	65.0
	Overall Percentage				63.0
a. The cut value is .220					

Table 13: Model summary.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29365.201[a]	.075	.115
2	28908.907[b]	.089	.136
3	28865.691[b]	.090	.138
4	28835.006[b]	.091	.140
5	28820.311[b]	.092	.141
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.			
b. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.			

5. Discussion

Through the Chi-square Test, education, marriage, and history of past payments are significant to the credit card default.

Unsurprisingly, education is significantly positively correlated with credit card defaults. The more educated a person is, the more knowledge they have learned. They can manage their wealth better and can anticipate the risks that may occur in the future so that they can make the proper decisions currently. They will not cause excessive consumption and are less likely to default on their credit cards.

Also, when people get married, it means that there is a relationship that people can support each other. It indicates they will have more money to lower default risks [9]. However, two people mean that they will have more expenditure. For example, they will buy unnecessary gifts for each other. Often, when people get married, they will have a child. Raising a child may be a considerable expenditure since they consume things instead of earning back money. It seems reasonable that when people are married, they are more likely to default on the credits.

History of past payments is significant to the credit card default because it reflects how long you delay paying your bill. The bank can evaluate your faith through this data and assess whether lending money to you is a high risk [10].

As the results show, LIMIT_BAL, AGE, BILL_AMT1, and PAY_AMT1 do not pass the normality test, so nonparametric tests are done instead of the t-test. Based on the nonparametric tests LIMIT_BAL, BILL_AMT1, and PAY_AMT1 are significant, while AGE is not. LIMIT_BAL means that the bank gives a person how much they can credit. A higher LIMIT_BAL indicates that you can buy many things in advance. It can put more pressure on the loan when the date comes. A higher BILL_AMT1 often results from a higher LIMIT_BAL. A high bill amount also means pressure and risk. PAY_AMT1 is similar to the history of past payments. But instead, the bank

DOI: 10.54254/2754-1169/25/20230482

evaluates your faith based on your financial performance last degree. The possible reason why Age is not significant is that when people are young, they may get an excellent job because they can stay up to finish the task, or they may not get a good job because they lack experience. When they are old, they may still work in a company with tremendous expertise or are fired since they cannot do much work. Their salary has a high degree of uncertainty, delaying whether they will have a credit card default.

The ANOVA analysis tells us that you can have a higher LIMIT_BAL if you have a higher education status. Higher education status means that you have more knowledge to manage your money, so the bank trust you can pay back when the date comes. On the other hand, a higher LIMIT_BAL may have a higher risk of credit card default. Consequently, the bank should find a balance between the LIMIT_BAL and the education status since they are conflicting factors. Other criteria or characteristics, such as marriage, should be included when the bank gives a person credit.

The first binary logistic regression model puts all factors into the model using the forward method: conditional. Before the analysis, we do the factor analysis to reduce the dimension and to find whether it can build a better model. Since many factors exist, factor analysis helps create a more accurate binary logistic regression model. Since there is a steep down between factor 4 and factor 5 in the Scree Plot shows, we use five elements instead of the characteristic root value greater than one as the judgment criterion. Since the model is significant, we use these five factors to build an accurate binary logistic regression model. To balance the recall and precision [11], we try several figures, such as 0.1 or 0.9, and in the end, 0.22 is used as the cut value. The percentage current for 0 and 1 is 62.5% and 65%, but the model's fitness is only 14.1%, far less than the first model's 67.2%, so we decide to use the first model.

As the figure shows, education is a more significant negative correlation to credit card default than marriage. A possible reason is that marriage is an external factor while education status is an internal factor. Marriage is more complex, and the bank could not evaluate it. For example, they could not know whether they have a good relationship or the condition of the other side. However, the education status can be examined more efficiently, such as a high school or graduate degree, so education is more significant to the marriage.

6. Conclusion

This paper aims to help banks evaluate and screen credit card customers. In this paper, whether people have credit card default is used to assess the user's credit. In the research, different methods in SPSS, such as ANOVA analysis or Binary logistic regression, are applied to determine whether other variables significantly affect credit card default and, when there is a conflict between different factors, which one should the bank should focus on more. The results show that variables including given credits, education status, marital status, the one-time payment, the billing statement, and the previous payment are all significant to the credit card default. And education seems to be more important among all the variables through Binary logistic regression. Therefore, when the bank needs to evaluate and screen the customers' credit, these variables may be considered as criteria, while the education level should be given more weight.

References

[1] *Gupta, M., Abdelmaksoud, A., Jafferany, M., Lotti, T., Sadoughifar, R., & Goldust, M. (2020). COVID-19 and economy. Dermatologic therapy, 33(4).*

[2] *Antipova, T. (2021). Coronavirus pandemic as black swan event. In Integrated Science in Digital Age 2020 (pp. 356-366). Springer International Publishing.*

[3] *Liadze, I., Macchiarelli, C., Mortimer-Lee, P., & Juanino, P. S. (2022). The economic costs of the Russia-Ukraine conflict.*

DOI: 10.54254/2754-1169/25/20230482

[4] Barceló, J. A. (2018). Chi-square analysis. The encyclopedia of archaeological sciences, 1-5.
[5] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.
[6] Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. Korean journal of anesthesiology, 70(1), 22-26.
[7] Harrell, Jr, F. E., & Harrell, F. E. (2015). Binary logistic regression. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, 219-274.
[8] Cureton, E. E., & D'Agostino, R. B. (2013). Factor analysis: An applied approach. Psychology press.
[9] White, L., & Rogers, S. J. (2000). Economic circumstances and family outcomes: A review of the 1990s. Journal of Marriage and Family, 62(4), 1035-1051.
[10] Eze, E. (1998). Lending credence to a borrowing analysis: Lone English-origin incorporations in Igbo discourse. International Journal of Bilingualism, 2(2), 183-201.
[11] King, J. E. (2008). Binary logistic regression. Best practices in quantitative methods, 358-384.