Analyzing Credit Limits and Predicting Credit Card Defaults

Cem Tener
Bradley A. Hartlaub
STAT 206
May 9, 2022

Introduction:

With the improvement of technology, the finance and banking businesses started to grow rapidly. As cash lost its importance in everyday transactions, money started to become an arbitrary number on computer and phone screens. Banks used this opportunity to leverage their assets in the hope of clients not withdrawing cash from their accounts. As banks and depository institutions were leveraging up their assets, they were also trying to figure out how much cash they need to store in their vaults and calculate how many of their clients will miss a payment or default on their loans. As long as the bank's clients are making their payments and have enough cash their clients would want to withdraw their deposit, banks would survive and continue making money. Therefore, banks have 2 very important questions to answer. First, how much credit should they give to a client, so that individuals make their payments. Second, how likely a client is to default.

In this project I am going to analyze the factors that affect an individual's credit card limit by running Anova and Ancova models. This would be an important question to answer for any banking institution because banks should have a structure on how much credit they give out. They would like to give enough but not too much credit to each individual so that they can pay back their borrowings and not risk the bank. Also, I will try to calculate the probability of someone defaulting on their credit card payment. At the same time, banks should also calculate an expectation on how many of their clients will default on their payments. If they do, then they might run into cash problems when many people default on their payments. I will build a logistic regression model to predict how likely an individual is to default on their credit card according to their characteristics.

Data:

I gathered data for a Taiwanese bank. The following are the descriptions of the variables that will be used in this research. I have a random sample which satisfies the independence condition of Anova, Ancova and logistic regression and the randomness condition of logistic regression.

Limit = Amount of limit an individual has in New Taiwanese dollars.
Default = An indicator variable which gets 0 if an individual has not defaulted and 1 if they had.
Education = Categorical variable with 3 levels. 3 Categories are; 1 is "High School" , 2 is "College", and 3 is "Graduate School"
Age Categories: Categorical variable with 3 levels. 3 Categories are Young, Mid Age, and Elder; Young is ages until 30, Mid Age is ages between 30 and 65, and elder is ages older than 65.
Billing= I used 6 different billing variables which are Billing_1 up to Billing_6. Each variable shows the amount of Taiwanese dollars spent 1 to 6 months prior to data collection.

Pay= Categorical variable that shows how many months an individual is late to pay its credit card payment. Pay can get -1 which indicates that it is not a late payment. There were 6 Pay variables from Pay_0 to Pay_5 showing repayment status of that month's payment 1 to 6 months prior to data collection.

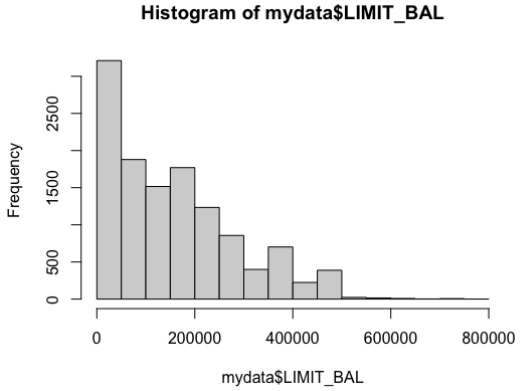
Part A: Evaluating Credit Card Limits-Factors Affecting Credit Card Limits

A.1) Do people who have defaulted on their credit card get lower credit amounts?

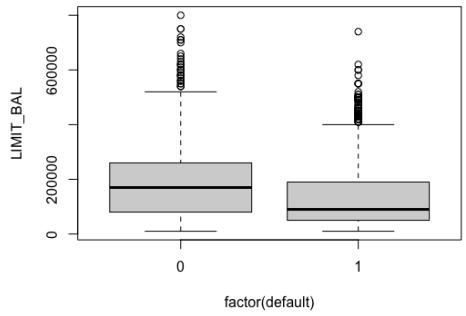
To begin exploring the dynamics of credit card limits, I would like to analyze if defaulting affects an individual's credit amount. To explore this question, I ran a t-test for the two groups, one for those who have defaulted and one for those who have not.

Ho: Mean(Default) = Mean(No Default) vs H1: Mean(Default) $\neq$ Mean(No Default)

With t = 26.331, df = 10168, p-value < 2.2e-16<.05, I can conclude that there is a difference in the mean credit card limit of people who have defaulted and have not defaulted.



However, looking at the histogram, we can observe that there is no normality in my data. Therefore, it would be best to use nonparametric tests and compare the test results to make sure that our analysis results are correct. I used the Wilcox test to test if there is a difference in the medians between two groups. With a W = 21928208, p-value < 2.2e-16 <.05, Wilcox test indicated that there is a significant difference between the medians. This evidence supports what we have concluded with the parametric t-test. Looking at the boxplot below, we can observe that people who have not defaulted on their credit cards have a higher limit. (1 is coded as being defaulted and 0 as not defaulted on a credit card)



A.2) Does your credit card limit increase as your level of education increases?

Recent salary datas show that higher education is rewarded by higher salaries. Therefore, banks might also give higher credits to individuals with higher education. I would like to test a one-sided hypothesis that the mean credit card limit increases as an individual gets more educated. To test this hypothesis, I build up a contrast in linear(increasing) weights;

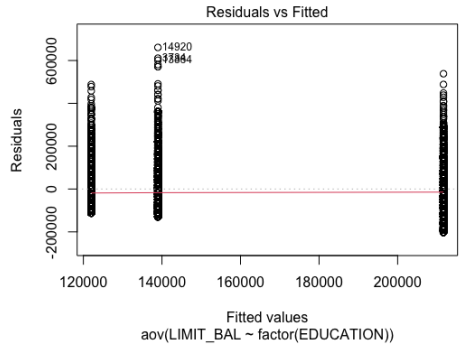
C1*Mean(High School) + C2*Mean(College) + C3*Mean(Graduate School),
where C1+C2+C3=0
(-.5,0,.5)

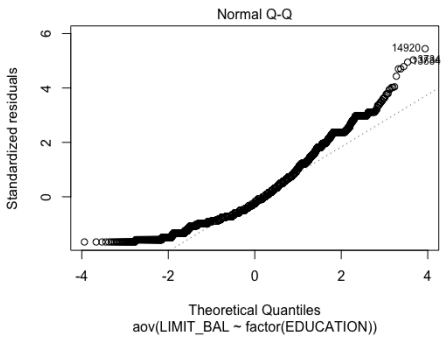
Ho: $\sum$ C1*Mean(High School) + C2*Mean(College) + C3*Mean(Graduate School)=0

vs

H1: $\sum$ C1*Mean(High School) + C2*Mean(College) + C3*Mean(Graduate School)$\neq$ 0

After testing my contrast, I got an estimate of 89669.04 and with t-value=27.76772 and p-value = 1.275712e-164<.05, I can reject H0 and conclude that credit card limits increase as an individual's level of education increases. This also follows the logic that individuals with higher education are more likely to make more money who would have higher credit card limits.





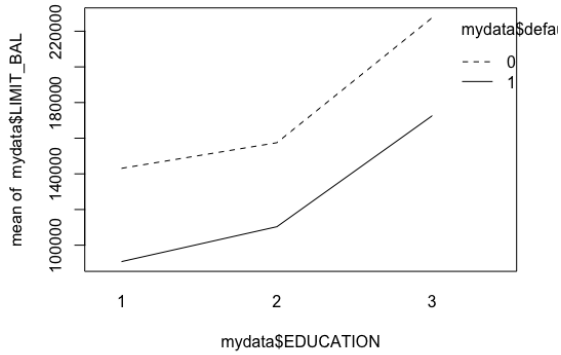
Unfortunately, the residual plots show that our conditions are not met. Therefore, I used a nonparametric test to make sure my conclusion is correct.
I used a one-sided Jonckheere-Terpstra test to see whether credit card limits increase as individuals get more educated.

Ho: Median(High School) = Median(College) = Median(Graduate School)
vs
H1: Median(High School) ≤ Median(College) ≤ Median(Graduate School), at least one is a strict <

After running the test, the nonparametric Jonckheere-Terpstra test gave a parallel conclusion with my contrast. With J statistic = 491.5 and p-value= .009<.05, I was able reject Ho and conclude that at least one of the medians increased.

A.3) Testing Defaults and Education in Two Way Anova

After finding out statistical differences in credit card limits between defaulting or not and getting higher education, it is appropriate to run a two way Anova test to see how the results are affected. First, I looked at the following interaction plot to see if interaction is present.

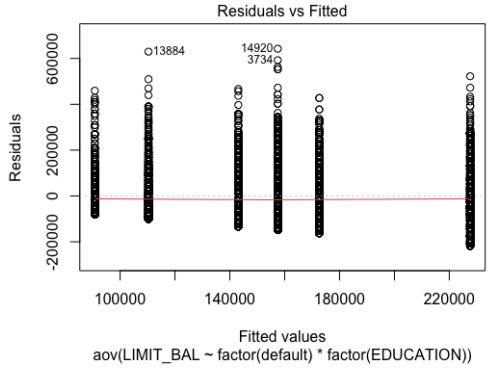


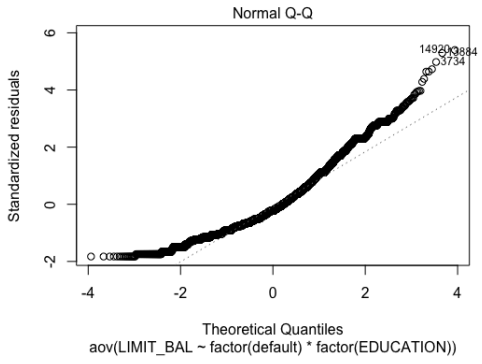
Looking at the interaction plot, both default groups react the same as education levels differ. To statistically justify interaction is not present, I built up the following two way Anova Model:

$Y= \mu + \tau_j + \beta_i + \epsilon \sim(0,\sigma^2)$, where j= 0,1(defaulted or not) and i = 1,2,3(High school, College, Graduate School)

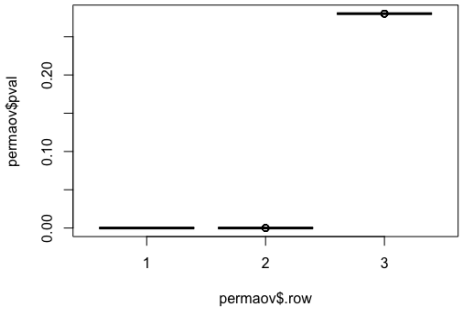
	DF	Sum Square	Mean Square	F-value	P-value
Default	1	9717577433625	9717577433625	685.936	<2e-16
Education	2	153891074741	76945537370 9	543.137	<2e-16
Default*Education	2	3605646051	18028230259	1.273	.28
Residual	12229	173246797870252	14166881828		

Two-Way Anova results justify that there is no significant interaction between Default and Education. Further, I got the same results with individual One-Way Anova tests. However, looking at the residual plot below, my Two-Way Anova conditions are not met. Therefore, I ran a permutation test and looked at the Anova results in 1000 different samples(without replacement)





Permutation Test for Two-Way Anova Results



The numbers below the box plots correspond to the row numbers on the Anova table. Number 1 corresponds to Default, 2 corresponds to Education, and 3 corresponds to the interaction term of Education and Default. The boxplot shows that out of 1000 randomized samples, there were no observations with a p-value greater than .05 for Education and Default. Conversely, there were no observations with a p-value less than .05 for the interaction between Education and Default. Therefore, I can conclude that there is a difference in credit card limits between Default and Education factors and no interaction is present between these two factors.

A.4) Assessing Credit Card Limit with Ancova

I continued my analysis by adding the billing amount on a monthly basis to my model. One caveat in my project is that I could not find income data of these individuals. Therefore, I am replacing income with these individuals' monthly billing amounts. An individual would spend more money as their salary increases. Therefore, billing amount data should also reflect a similar implication in my model. Billing 1 shows the billing amount of an individual 6 months before the data has been collected and Billing 5 shows the billing amount of an individual 2 months before the data has been collected. I also added variable Age Categories to test how credit card limits change in different age groups. I ran the model for the log of credit card limit to satisfy OLS conditions. In my Ancova model, Billing 1 and Billing 5 are my covariates and Education, Default and Age Categories are my factor variables. After taking out the insignificant terms, I built the following model:

Credit Card Limit=$\beta_0 + \beta_1$Billing1+ β_2Billing5+ β_3Education+ β_4Default+ β_4Age Categories
$\epsilon \sim (0, \sigma^2)$

Anova Table for Ancova

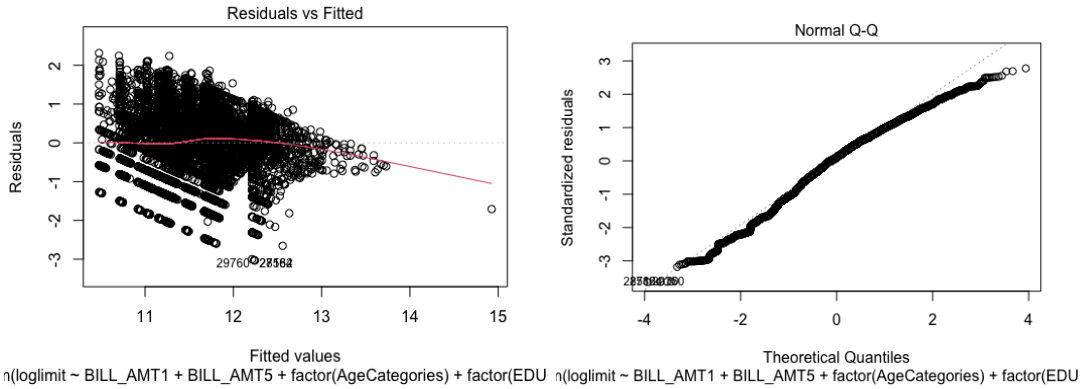
	DF	Sum Square	Mean Square	F-value	P-value
Billing 1	1	186.8	186.84	268.38	< 2.2e-16
Billing 5	1	74.9	74.85	107.52	< 2.2e-16
Age Categories	2	749.5	374.76	538.31	< 2.2e-16
Education	2	1224.9	612.47	879.76	< 2.2e-16
Default	1	514	514.01	738.34	< 2.2e-16
Residual	12227	8512.1	.7		< 2.2e-16

Ancova Summary Table

	Estimate	Standard Error	t-value	P-value
Intercept	10.9166360671	0.0240801135	453.347	< 2e-16
Billing 1	0.0000012559	0.0000002465	5.094	0.0000003556642
Billing 5	0.0000024754	0.0000002742	.028	< 2e-16
Age Category (2)	0.5401232865	0.0164701581	32.794	< 2e-16
Age Category (3)	0.8561753980	0.1281657973	6.680	0.0000000000249
Education(College)	0.2334468050	0.0220521337	10.586	< 2e-16
Education(Graduate)	0.7539787913	0.0223845562	33.683	< 2e-16
Default(Yes)	-0.4384358354	0.0161353420	-27.172	< 2e-16

The fitted estimates of the Ancova model can seem very odd because they are very small values. However, this is because I did the log transformation. For this reason my data points are "squeezed" because of the nature of the log function. My Ancova model has an adjusted R-square of .2438 which is not very satisfying. Even though I had put many explanatory variables into my model, it is still weak to explain the variability in the data. The Ancova model suggests that monthly billing amount is positively associated with an individual's credit card limit. The model suggests that individuals who spend more would have higher limits. Further, only people who have higher income can spend higher amounts. Therefore, the Ancova model draws a conclusion on the implication of the billing variables. Secondly, the model concludes that there is a difference of credit card limits between age groups. From the summary table, we

can conclude that the age categories are associated with the credit card limit. Since the coefficient of Age Category for Elder is greater than the coefficient for Mid-Age which has a positive coefficient, the Ancova model suggests that as individuals get older, their credit card limit increases. Finally, we get the same conclusion as with the Two-Way Anova for Education and Default variables. Similar to the Age Categories variable, beta coefficients have the same orientation and suggest an increase in the credit card limit as an individual's education level increases. Additionally, people who default on their credit cards tend to have lower limits because the beta coefficient is negative.



The residual plots show violation of OLS conditions and show that my Ancova coefficients are biased. Therefore, I am running a rank based nonparametric regression to see if I can make similar conclusions.

Rank Based Nonparametric Regression Summary Table

	Estimate	Standard Error	t-value	P-value
Intercept	55248.601238	2713.988564	20.3570	< 2e-16
Billing 1	0.083942	0.027936	3.0047	0.002664
Billing 5	0.417439	0.031068	13.4365	< 2e-16
Age Category (2)	62889.731790	1866.303524	33.6975	< 2e-16
Age Category (3)	105119.854657	14523.010500	7.2382	0.000000000000482

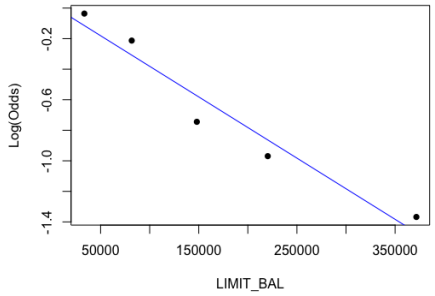
Education(College)	24751.398762	2498.820869	9.9052	< 2e-16
Education(Graduate)	91471.300546	2536.489076	36.0622	< 2e-16
Default(Yes)	-50825.520497	1828.364093	-27.7984	< 2e-16

Looking at the summary table for the nonparametric regression, we can observe that the coefficient estimates are more intuitive because we do not have to run for the log of the credit card limit or make any transformation because we do not have any conditions to perform nonparametric regression. In comparison to the Ancova model, there are no changes in the signs of the coefficient estimates which concluded that the relationship of my variables are still the same and it is safe to conclude the associations explained in the Ancova model.

Part B: Probability of Default

B.1) Assessing the Probability of Default with Simple Logistic Regression

I would like to assess the probability of default as individuals get higher limits. To answer this question, I started off my analysis by building a simple logistic regression with the amount of limit individuals have as the explanatory variable.



Before building up my logistic regression, I built an empirical logit plot to visualize the linear pattern between credit card limit and the log of odds. I divided my sample into 5 groups to combine limit balances that are similar to each other. My plot shows that there is a linear pattern and satisfies the linearity condition of logistic regression.

Now, I built a logistic regression with the following logit and probability forms:

$\log(\pi/1-\pi) = \beta_o + \beta_1 * \text{Limit}$ and $\pi = e^{\beta 0 + \beta 1 * Limit} / (1 + e^{\beta 0 + \beta 1 * Limit})$

From here, I can estimate the probabilities based on credit card limits. The table below shows the probabilities of defaulting when the limit balances in quantile 1, mean, median, and quantile 3.

Amount of Limit	Probability
Quantile 1 = 50000	0.4566469
Median = 140000	0.3667806
Mean = 165241.2	0.3428902
Quantile 3 = 230000	0.2853131

Looking at the chart above, we can observe that people with higher credit limits tend to default less than those with lower limits. The results in this research follow a chain of intuitive connections. First, from part A, individuals with higher spending, which implies they have higher income, have higher credit limits. Second, logistic regression shows that those of individuals with higher limits are less likely to default. It is interesting that individuals with 50000 limit are almost 50% likely to default. This is a very high probability of an individual defaulting. Banks should consider this fact in their internal risk models. Lastly, combining the conclusions from part A and B, individuals with higher spending/income have higher limits; therefore, they are less likely to default. This is what I was expecting before starting this research because higher income individuals are better in managing their wealth and cash flows. Therefore, they are more likely to pay their credit card bills on time.

B.2) Multiple Logistic Regression

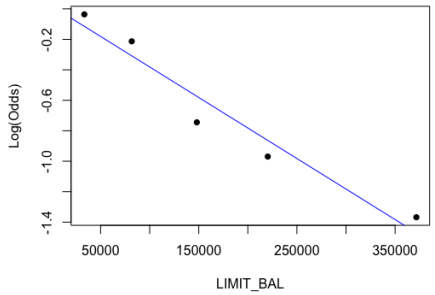
I wanted to improve my simple model by controlling other variables that would affect individuals defaulting on their credit card payments. I used the data for how long it took him/her to pay back his/her credit(Pay_0, 0 stands for 1 month prior data collection). I initially had data for up to 6 months prior to data collection. However, only the Pay_0 variable was significant in my model. Also, I added Age Categories data in expectation that younger individuals would be more careful not missing a credit card payment because they would want to build a good credit score. I finalized my model by backward elimination, taking out the least significant term in the model until all my variables are significant. I built up the following models in the logit and probability forms:

$\log(\pi/1-\pi) = \beta_o + \beta_1 *$ Limit $+ \beta_2 *$Pay0 $+ \beta_3 *$Age Categories

and

$$\pi = e^{\beta o + \beta 1 * Limit + \beta 2 * Pay0 + \beta 3 * Age\ Categories} / (1 + e^{\beta o + \beta 1 * Limit + \beta 2 * Pay0 + \beta 3 * Age\ Categories})$$

H0: $\beta 2 = \beta 3 = 0$ vs Ha : some βi does not equal 0 for $i > 2$



I only have one continuous variable in my model. Therefore, I am only checking linearity for the limit variable. Looking at the empirical logit plot, we can observe that the linearity condition is met.

After running a Nested G-test to compare the multiple logistic regression with the simple one in B.1, I got a G value of 584.9 and a p-value of 0. Therefore, I can conclude that I was successful in improving my logistic regression model.

Multiple Logistic Regression Prediction Table

Limit	Age Category	Pay_0	Probability
50000	Young	Never missed	0.2121834
50000	Elder	Never Missed	0.3174922
50000	Elder	7 months	0.9012623
165241.2	Young	7 months	0.754371
165241.2	Young	Never Missed	0.1472338
165241.2	Elder	Never missed	0.2759096
230000	Young	2 months	0.6421061
230000	Mid Age	Never missed	0.1655754
230000	Elder	7 months	0.8363127

Looking at the table above, we can see that individuals who have never missed their monthly payments have lower probability of defaulting regardless of their age and credit limit. We can observe that for older individuals who have a 50000 limit, the probability of the ones who have missed their payments for 7 months is 0.9012623 - 0.3174922 = 0.5837701 higher than those who have never missed their payments. Further, the discrepancy between not missing any payments and missing a payment for 2 months is huge. Therefore, my regression implies that

missing any payment makes your probability of defaulting in the future much higher. Again, my regression gives intuitive results because if an individual missed a payment, it implies that his/her finances are not doing good. That's why it is very likely that his/her finances will go worse and miss other payments.

Assessing the impact of Age Categories, group Mid Age was significant, but group Elder was not to my model. However, I decided to keep the Age Categories variable because it can explain 2(Young and Mid Age) out of 3 factors in the Age Categories variable. Looking at the probabilities for individuals with 165241.2 (mean) limit and who have never missed a payment before, the difference between the probability of individuals who are Elder and probability of individuals who are Young is 0.2759096-0.1472338 = 0.1286758. This is almost double the probability of the Young individuals' default. Looking at this result, I am much more confident in keeping the Age Categories variable even though its Elder factor is not significant.

Conclusion:

Everything all in together, my conclusions gave out statistical evidence as what I was expecting. I used nonparametric tests when necessary to strengthen my conclusions. From my research, I was able to conclude that Default, Education, and spending amounts were important factors that affect individuals credit limits. While Education and spending amounts are positively associated with credit limits, individuals who default have less credit limits. Further, increasing the limit and not defaulting on credit cards makes you a less risky client for the bank. However, individuals' probability of defaulting their cards increases as they grow older and falling behind on previous credit bills.
For both parts, I did not have data for the income of these individuals to test credit limits of these individuals. Even though spending is very correlated with income, a future researcher might want to go over the same analysis with the income data.