Small Business Loan in North Carolina Statistical Analysis

Heather Qiu, Jenny Shen, Aditya John, Isha Singh

2023-01-04

I. Abstract

U.S. Small Business Administration (SBA) is responsible for promoting and assisting small businesses in the U.S. credit market. The dataset is obtained from Li, Mickel, and Taylor (2018), "Should This Loan be Approved or Denied?" The file contains various data points about companies that have applied for loans, such as the number of employees, the locality type, and the loan term, amongst others. Our goal with this analysis is to determine if these data points can be used to determine the likelihood of loan defaults and what impact the amount of loan the SBA approves for small businesses in the state of North Carolina.

For this study, we investigate the impact of loan duration, the presence of a revolving line of credit, and the employee count on the loan amount approved by the SBA. We find that all have a significant relationship with the amount covered by the SBA loan. The loan term and the number of employees have a positive relationship with the approved amount. On the other hand, the presence of a revolving line of credit has a negative relationship with the amount covered by the SBA. The final model has an R^2 of 0.4031, meaning that 40.31% of the variability in the dataset could be explained by our model.

Additionally, we investigate multiple factors that might affect the good standing of a loan. The variables (1) the loan term, (2) the number of employees, (3) the number of jobs created, (4) whether the business is in urban or rural areas, (5) the revolving line of credit status, (6) whether the loan is part of the LowDoc Loan program, (7) whether the loan was disbursed during recession years between 2007 and 2009, and (8) SBA's guaranteed amount of approved, are found to have a significant relationship with the status of loan default at the $\alpha = 0.05$ significance level. Of the above, we find that the most significant factor is the loan term, with the intuition being that the longer the loan term, the more likely the company is to default on the loan. We also observe that the second most impactful driver of loan defaulting is whether the loan was distributed during the 2008 recession. Last, we identify that the whether the company is a new or existing business and the number of jobs retained by the company rank last and second last, respectively, in terms of the effect. Our model has an accuracy rate of 79.72%.

II. Introduction

The default rate in the banking industry is a significant indicator for evaluating the economic situation. Small businesses, more specifically, have higher probabilities of default due to their limited size and budget. Moreover, the proportion of bank debt to total debt in small businesses is roughly twice that of large firms (Ou, 2005). Therefore, it is crucial to evaluate the default risk of small businesses for economic stability purposes.

SBA plays the role of an insurance provider to banks by guaranteeing a portion of the loan. In the case that a loan goes into default, SBA then covers the amount that they guarantee. Despite the safety net provided by the SBA, banks will still incur losses when a loan defaults because not all dollar amount is covered by SBA. Therefore, our analysis aims to analyze the probability that a loan is to default and identify the key factors that impact the loan amount of SBA guaranteed.

For our first research question, we are interested in how the aforementioned specific factors can be used to determine the amount covered at the SBA level. We hope to understand the impact that these factors have on the loan and whether they can be used to quantify the amount that the SBA should cover.

Our second research question is more open-ended. We aim to determine what are the various factors that have significant relationships with the status of loan default. We are also interested in measuring the relative impact size among all predictors in scope.

III. Methods

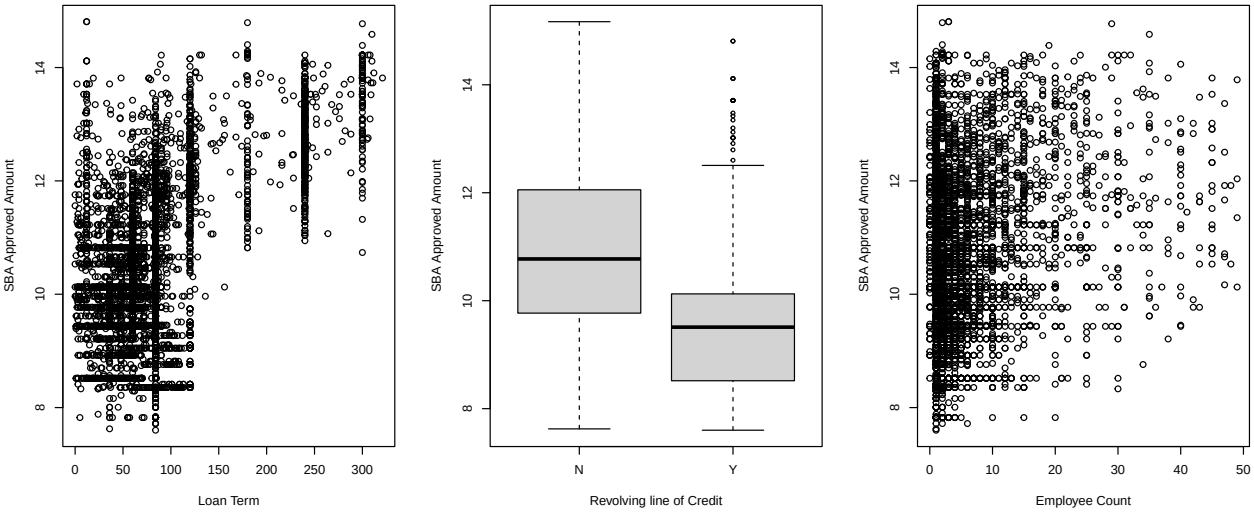


Figure 1: Exploratory Data Analysis of SBA Loans

Data

The initial dataset includes over 14,000 observations for small business loans in North Carolina. We first removed loan records with missing values. Additionally, there are non-binary values in binary predictors. Under the variable that determines whether the loan is applied through the low document program, for instance, we found more values than the expected Yes and No answers. Without additional information from the data source, we removed loan records with non-binary values in binary covariates. We also subset variables based on our research interests using apriori variable selection. Because the number of predictors is limited compared to the total observations, additional variable selection is not necessary. We also collapsed the disbursement date of loans into two main categories and named the variable, Recession. The years between 2017 and 2019 are labeled as 'yes' and 'no' for the rest. After trimming the dataset, the final data includes 5,448 observations. At last, we limit the analysis to only companies with less than 50 employees for the first research question.

With regard to the EDA for our first research question, we can see that the loan term has a positive relationship with the SBA approval amount. We also notice a clear difference corresponding to whether the business has a revolving line of credit. Companies with existing revolving lines of credit receive lower amounts guaranteed by the SBA when compared to businesses without revolving lines of credit.

With regard to the EDA for our second research question, we are interested in exploring the relationships among several predictors and the status of loan default. The descriptive statistics of the dataset (Table 0. Descriptive Statistics of SBA Loan) show a comprehensive overview of the data set bifurcated using loan status. Among the 5,448 loan records, one-fourth of loans defaulted, while the rest was paid in full. Among all borrowers, over 60% are existing small businesses. On average, these businesses create two jobs and retain five jobs. Roughly four-fifths of these companies operate in cities across the country, while the remaining fifth runs in rural areas. About 65% of the small businesses do not have a revolving line of credits at the time of application; the vast majority (98.5%) applied through the low documentation program, designed to streamline the application process for borrowers.

Table 0: Descriptive Statistics of SBA Loan

	Not Defaulted	Defaulted	Overall
	(N=4051)	(N=1397)	(N=5448)
Term			
Mean (SD)	98.5 (63.5)	55.7 (33.0)	87.5 (60.2)
Median [Min, Max]	84.0 [1.00, 321]	53.0 [0, 300]	84.0 [0, 321]
Employee Count			
Mean (SD)	8.43 (30.9)	5.31 (8.16)	7.63 (27.0)
Median [Min, Max]	3.00 [0, 1630]	3.00 [0, 95.0]	3.00 [0, 1630]
New or Existing Business			
Existing Business	2560 (63.2%)	918 (65.7%)	3478 (63.8%)
New Business	1491 (36.8%)	479 (34.3%)	1970 (36.2%)
Job Created			
Mean (SD)	2.39 (9.28)	2.39 (17.9)	2.39 (12.1)
Median [Min, Max]	0 [0, 451]	0 [0, 451]	0 [0, 451]
Job Retained			
Mean (SD)	4.98 (12.6)	4.13 (6.33)	4.76 (11.4)
Median [Min, Max]	2.00 [0, 275]	2.00 [0, 73.0]	2.00 [0, 275]
Urban or Rural Location			
Urban	2974 (73.4%)	1224 (87.6%)	4198 (77.1%)
Rural	1077 (26.6%)	173 (12.4%)	1250 (22.9%)
Revolving Line of Credit			
N	2544 (62.8%)	988 (70.7%)	3532 (64.8%)
Y	1507 (37.2%)	409 (29.3%)	1916 (35.2%)
Low Document Program			
N	3978 (98.2%)	1389 (99.4%)	5367 (98.5%)
Y	73 (1.8%)	8 (0.6%)	81 (1.5%)
Recession			
N	3147 (77.7%)	758 (54.3%)	3905 (71.7%)
Y	904 (22.3%)	639 (45.7%)	1543 (28.3%)
SBA Approved Amount			
Mean (SD)	143000 (283000)	55900 (129000)	121000 (255000)
Median [Min, Max]	32000 [2000, 3870000]	18700 [2500, 1500000]	25000 [2000, 3870000]

Model

For our first research questions, we are interested in assessing the impact of 3 variables: (1) Loan Term, (2) Revolving Line of Credit, and (3) Employee count. Our dependent variable is the amount approved by the SBA, which is a continuous variable. Given the nature of our response variable with more than one independent variable, we fit a multi-linear regression model on the dataset. Linear regression helps quantify the relationship between the dependent and independent variables.

For the second research question, we performed apriori variable selection to prune the list of variables and retained the ones that are useful in modeling the dependent variable based on domain knowledge. In addition, we excluded the variables loan charged-off amount and outstanding gross amount, as they are multivariates of the dependent variable. Here is the final list of variables we will include: (1) the loan term, (2) the number of employees, (3) the new or existing business, (4) the number of jobs created, (5) the number of jobs retained, (6) whether the business is in urban or rural areas, (7) the revolving line of credit status, (8) whether the loan is part of the LowDoc Loan program, (9) whether the loan was disbursed during recession years, and (10) SBA's guaranteed amount of approved loan. We chose to fit a multiple logistic regression model because there are several predictors, and our dependent variable is a binary categorical variable. Logistic regression estimates the probability of an event occurring (i.e., loan defaulting versus not) based on a given set of

independent variables. In logistic regression, a logit transformation is applied to the odds, which is the probability of success divided by the probability of failure. In the case of the small business loan, the preferred outcome is when a loan gets paid in full.

Model Assessment

For both models, the following will be performed to ensure goodness of fit and performance.

1. Significance Evaluation:
   For each predictor, we will examine the p-values. A predictor is considered significant at the $\alpha = 0.05$ significance level. Additionally, we will also use t-values to assess the impact size of each predictor on the outcome variable.
2. Outliers, Influential Points, and High-Leverage Points Identification:
   Using Cook's distance, we will determine the presence of outliers, influential points, and high-leverage points in the model. If any is identified, we will fit two models, one with and one without outliers, and present our findings.
3. Validity of Assumptions:
   Diagnostic plots and binned residual plots will be used to show that the underlying assumptions about the data concerning the model are not violated.
4. Presence of Multicollinearity:
   Models will be tested for multicollinearity using VIF values. Any variables that exhibit this behavior will be removed.
5. Performance Assessment:
   To evaluate the overall model performance, we utilize the R^2 metric and the confusion matrix's accuracy metric.

IV. Results

First Research Question: How do the requested loan duration, revolving line of credit, and employee count of a small business impact the amount of the approved SBA loan in North Carolina for businesses with less than 50 employees?

Table 1: Linear Regression Model Summary Statistics

	Estimate	SE	t value	Pr(>|t|)	CI 2.5 %	CI 97.5 %
Intercept	9.518	0.035	268.136	<.001	9.448	9.588
Term	0.011	0	38.486	<.001	0.01	0.011
Revolving line of Credit						
Yes	-0.861	0.034	-25.21	<.001	-0.928	-0.794
No	-	-	-	-	-	-
Employee Count	0.051	0.002	24.679	<.001	0.047	0.055

[1] SD:Standard Error; CI: Confidence Interval
[2] Adjusted R-squared value: 40.28%

Since the p-value for all the predictors, the loan term, the number of employees, and the revolving line of credit are smaller than 0.05, they are statistically significant in predicting the approved amount by SBA. Based on the absolute t-values of three independent variables, the loan term has the highest absolute t-value (38.49), which demonstrates the duration has the greatest impact on the loan approved amount. The other two predictors present similar absolute t-values, with 25.21 for the revolving line of credit and 24.68 for employee count.

The highlight is that the loan term and the number of employees positively correlate to the guaranteed loan amount by the SBA. For every additional month in the loan term, the average log of the SBA approved amount, measured in dollars, increases by approximately 0.01; for every worker employed by the small businesses, the average log of the SBA loan coverage increases by about 0.05, holding all else constant. On the other hand, the presence of a revolving line of credit negatively correlates to the amount covered by the SBA. For companies with existing revolving lines of credit, the average log of SBA's approved loan amount in USD decreases by approximately 0.86 compared to businesses without revolving lines of credit, holding all else constant.

Hence, we infer that there is a positive correlation between the length of the loan term and the SBA's approved loan amount. SBA seems more inclined to back up loans with extra dollars for small businesses with more employees. However, SBA tends to be less willing to lend more money to companies with existing revolving lines of credit.

There are no significant outliers, high-leverage points, or influential points identified in the linear regression model. We do not find any underlying linear regression assumptions to be violated in the data, thereby requiring no additional transformation. There is also no evidence for multicollinearity based on VIF values.

Second Research Question: What are the different factors that impact the defaulting of a loan in North Carolina?

Table 2: Logistic Regression Model Summary Statistics

	Estimate	SE	z value	Pr(>|t|)	Odds Ratio	CI 2.5 %	CI 97.5 %
Intercept	1.568	0.111	14.135	<.001	4.796	3.865	5.971
Term	-0.032	0.001	-24.245	<.001	0.969	0.966	0.971
Employee Count	-0.02	0.006	-3.514	<.001	0.98	0.969	0.991
New or Existing Business							
New Business	-0.045	0.078	-0.58	0.56	0.956	0.82	1.114
Existing Business	-	-	-	-	-	-	-
Job Created	0.006	0.003	2.1	<.05	1.006	1	1.012
Job Retained	-0.005	0.007	-0.775	0.43	0.995	0.982	1.008
Urban or Rural Location							
Rural	-1.17	0.1	-11.64	<.001	0.31	0.254	0.377
Urban	-	-	-	-	-	-	-
Revolving Line of Credit							
Yes	-0.998	0.08	-12.5	<.001	0.369	0.315	0.431
No	-	-	-	-	-	-	-
Low Document Program							
Yes	-1.309	0.389	-3.365	<.001	0.27	0.117	0.547
No	-	-	-	-	-	-	-
Recession							
Yes	1.133	0.077	14.746	<.001	3.104	2.671	3.61
No	-	-	-	-	-	-	-
SBA Approved Amount	0	0	-2.271	<.05	1	1	1

[1] SD:Standard Error; CI: Confidence Interval

The following predictors are statistically significant because their p-values are smaller than 0.05: (1) the loan term, (2) the number of employees, (3) the number of jobs created, (4) whether the business is in urban or rural areas, (5) the revolving line of credit status, (6) whether the loan is part of the LowDoc Loan program, (7) whether the loan was disbursed during recession years between 2007 and 2009, and (8) SBA's guaranteed amount of approved. By observing the absolute z-values, we identify the four most impactful drivers of loan

defaulting: the loan term (24.25), whether the loan was distributed during the 2008 financial crisis (14.75), the revolving line of credit status (12.5), and whether the business locates in urban or rural settings (11.64).

The highlight is that the odds of borrowers defaulting on loans increase during the 2008 financial crisis and when businesses operate in urban areas. For instance, the odds of loans folding in North Carolina during the recession rise by approximately 210% compared to the non-recession period; the odds of small businesses in rural areas defaulting on a loan decrease by roughly 69% compared to urban areas, holding all else constant.

On the other hand, the odds of businesses defaulting on loans decrease with each additional month in loan term and if the companies have existing lines of credit. For every extra month of the loan term, the odds of small businesses in North Carolina defaulting on a loan decrease by about 3%; the odds of loan folding for small businesses with existing revolving lines of credit decrease by approximately 63% compared to an establishment with no revolving line of credit, holding all else constant.

Hence, we infer that small businesses in urban areas with no revolving lines of credit and shorter loan terms are more likely to default on their loans during the recession period.

We are not concerned about multicollinearity issues in the logistic regression analysis because all the VIF values are below two. From the binned residuals versus predicted probabilities (Figure 4 in the appendix), we confirm the assumptions of logistic regression since all residuals are distributed randomly within the range. Though a few points lie outside the confidence intervals, there is no systematic pattern in the binned residual plots. As a result, no transformation is performed on variables, and the model demonstrates an acceptable model fit. There are no significant outliers, high-leverage points, or influential points identified in the logistic regression model.

The confusion matrix is shown below. The model's accuracy using 0.5 as the cutoff for predicting the small business loan default rate is 0.7972. The ROC curve shows the percentage of true positives predicted by the model as the prediction probability cutoff lowers from 1 to 0. The sensitivity is 0.4216, and the specificity is 0.9267. In this case, the area under the curve is 0.833, which is relatively high and shows a reasonably good model performance.

	Not Defaulted	Defaulted
Not Defaulted	3754	808
Defaulted	297	589

Table 3: model confusion matrix

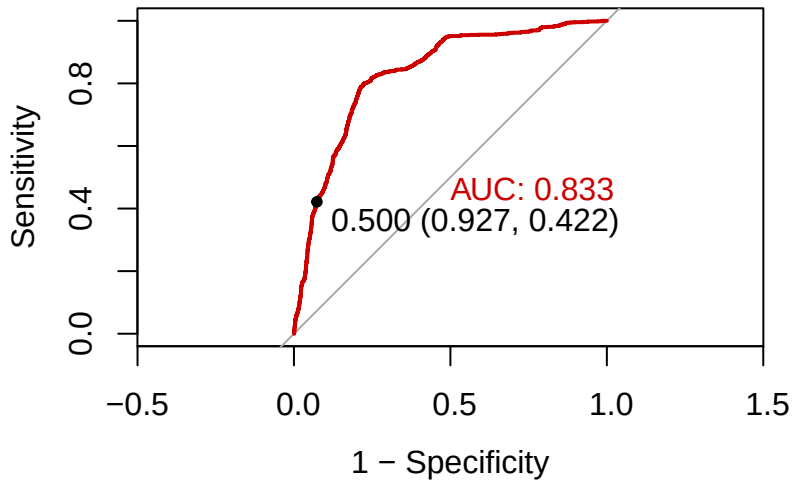


Figure 2: Logistic Regression Receiver Operating Characteristic (ROC) Curve

V. Conclusion

Key Takeaways

For our first research question, all in-scope variables share a significant relationship with the amount of loan covered by the SBA. We find that the employee count and the loan term positively correlate to the amount guaranteed by the SBA. The intuition is that more employees and longer loan terms both point to an increase in the loan amount requested by the company. Larger businesses tend to provide more jobs and require additional capital investments than smaller ones. Similarly, we would expect the requested loan amount to be greater as the loan terms increase. Therefore, when companies request loans with higher dollar values, the amount covered by the SBA also increases. For example, a 10% coverage on a one-million-dollar loan is higher than the loan coverage of 100,000 dollars at the same percentage. On the other side, we see that the amount the SBA covers decreases if the business has a revolving line of credit.

For our second research question, amongst the variables considered via apriori variable selection, we find eight of them significant at the $\alpha = 0.05$ significance level: (1) the loan term, (2) the number of employees, (3) the number of jobs created, (4) whether the business is in urban or rural areas, (5) the revolving line of credit, (6) whether the loan is part of the LowDoc Loan program, (7) whether the loan was disbursed during recession years between 2007 and 2009, and (8) SBA's guaranteed amount of approved. The loan term has the greatest impact on the probability of loan folding. Considering a loan of 20 years, as opposed to a 3-year loan, we expect the loan with a shorter term at a higher risk of default. One likely reason is that the shorter timeframe gives small businesses less time to manage their cash flow and pay back on time. Additionally, the 2008 recession is a crucial driver of the loan default status. According to Glennon and Nigro (2005), small businesses are more sensitive to local and industry time-varying economic trends. A market slowdown increases the likelihood of default. This explanation aligns with our research finding that the economic downturn has a comparably large negative impact on the default status of small businesses. We suspect that the reduced customer demand and decreased revenue during the recession are the leading causes.

Limitations

Our primary limitation comes from a lack of understanding of the dataset collection process. Therefore, though we find the presence of outliers in the data, we cannot verify whether they are valid observations. In addition, we have restricted the acceptable values for categorical variables by including only loan records with Yes or No answers. Again, this primarily arose because we do not have additional insights into the data collection process, making the interpretation difficult in the context of our research question.

We have only considered three variables for our first research questions. Hence, including more variables

in the dataset can lead to a more holistic understanding. The proportion of variation of the current linear regression fit will likely improve once more predictors are used. By limiting our inference to three variables, we are unable to assess other significant relationships that may impact our outcome variable. We may also find our currently considered variables insignificant once we include additional variables.

In assessing the accuracy of our logistic regression model, we find that the true positive rate is relatively low at 0.42, although the specificity is quite high (0.93). The sensitivity indicates that about 42% of defaulted loans are correctly classified, while the specificity shows that 93% of loans paid in full are accurately labeled. We believe the relatively lower sensitivity can be justified by the imbalanced data set where defaulted loans represent only about 26% of loans in North Carolina.

Future Work

Our first proposed future work is expanding the research to the entirety of the U.S. rather than restricting only to the state of North Carolina. The heat map (Figure 3) below shows the default rate of loans in small businesses by states in the continental United States. The warm color indicates a high loan default rate of up to 0.38. It appears that the southwest and the southeast of the United States have relatively high default rates. In contrast, the colder color represents the low default rates, primarily depicting the situation in the northern region of the states. As such, additional insights can be gained by evaluating the loan default rate across states.

Similarly, we can incorporate industries of small businesses in our analysis. Different industries face their unique challenges, not accounted for in the current analysis. Laws, regulations, and the black swan effects also vary across sectors. Thereby, we might be able to better model these risks into our understanding of potential loan defaults by segmenting across different industries.

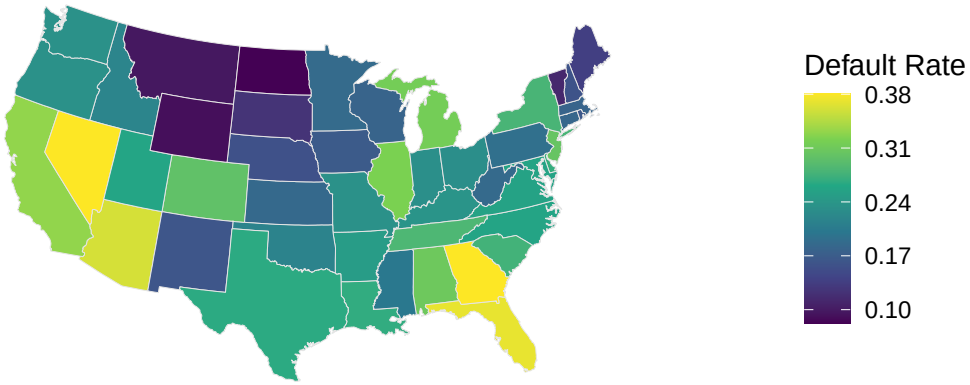


Figure 3: Default Rate of Loan by States in Continental US

Appendix

Table 4: Code Book

Variable	Description
LoanNr_ChkDgt	Identifier Primary Key
Name	Borrower Name
City	Borrower City
State	Borrower State
Zip	Borrower Zip Code
Bank	Bank Name
BankState	Bank State
NAICS	North American Industry Classification System Code
ApprovalDate	Date SBA Commitment Issued
ApprovalFY	Fiscal Year of Commitment
Term	Loan Term in Months
NoEmp	Number of Business Employees
NewExist	1 = Existing Business, 2 = New Business, 0 = Undefined
CreateJob	Number of Jobs Created
RetainedJob	Number of Jobs Retained
FranchiseCode	Franchise Code, (00000 or 00001) = No Franchise
UrbanRural	1 = Urban, 2 = Rural, 0 = Undefined
RevLineCr	Revolving Line of Credit: Y = Yes, N = No
LowDoc	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	The date when a loan is declared to be in default
DisbursementDate	Disbursement Date
DisbursementGross	Amount Disbursed
BalanceGross	Gross Amount Outstanding
MIS_Status	Loan Status Charged off = CHGOFF, Paid in Full =PIF
ChgOffPrinGr	Charged-off Amount
GrAppv	Gross Amount of Loan Approved by Bank
SBA_Appv	SBA's Guaranteed Amount of Approved Loan

Figure 4. Assumptions Check for the Multiple Linear Regression Model

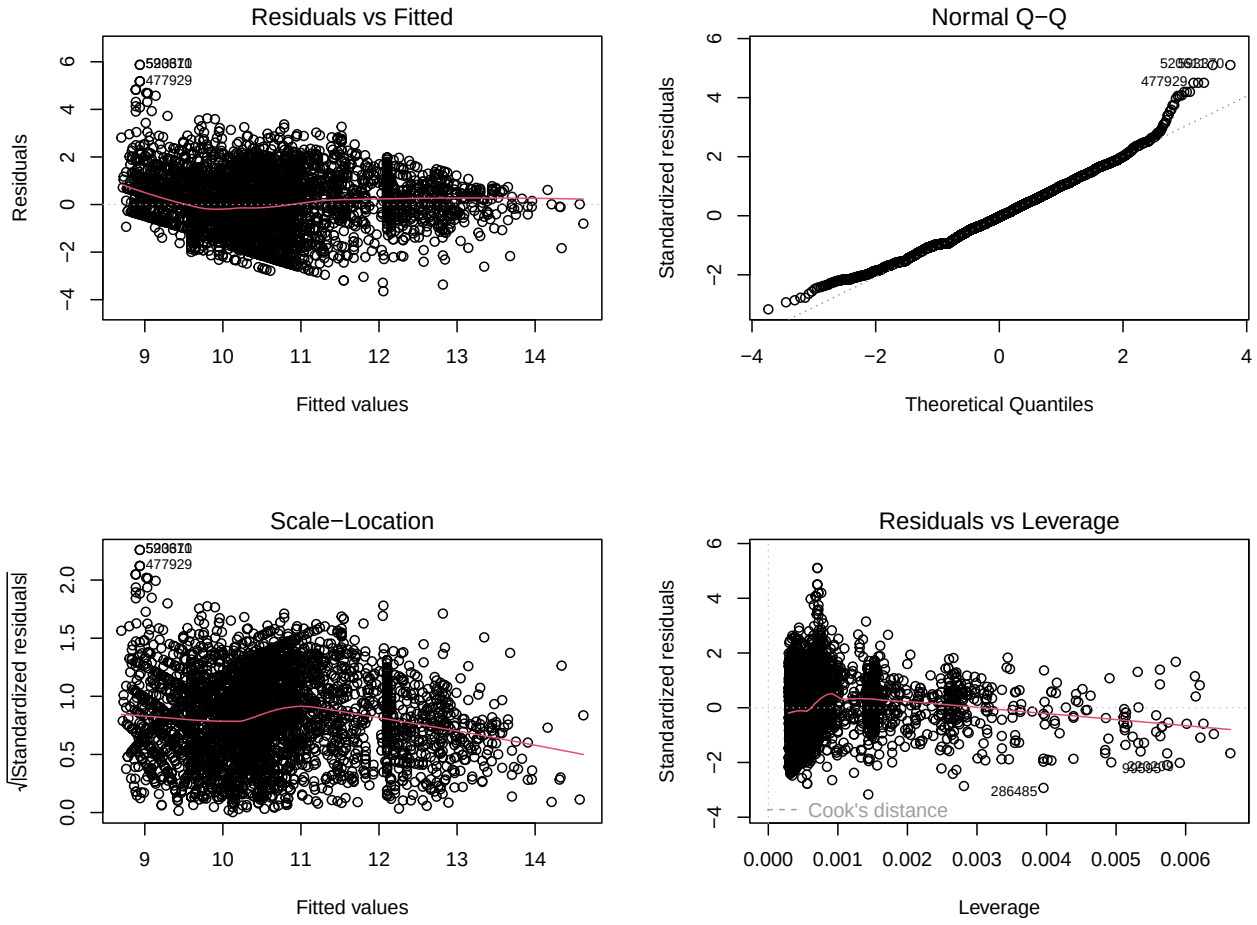


Figure 5. Assumptions Check for the Logistic Regression Model

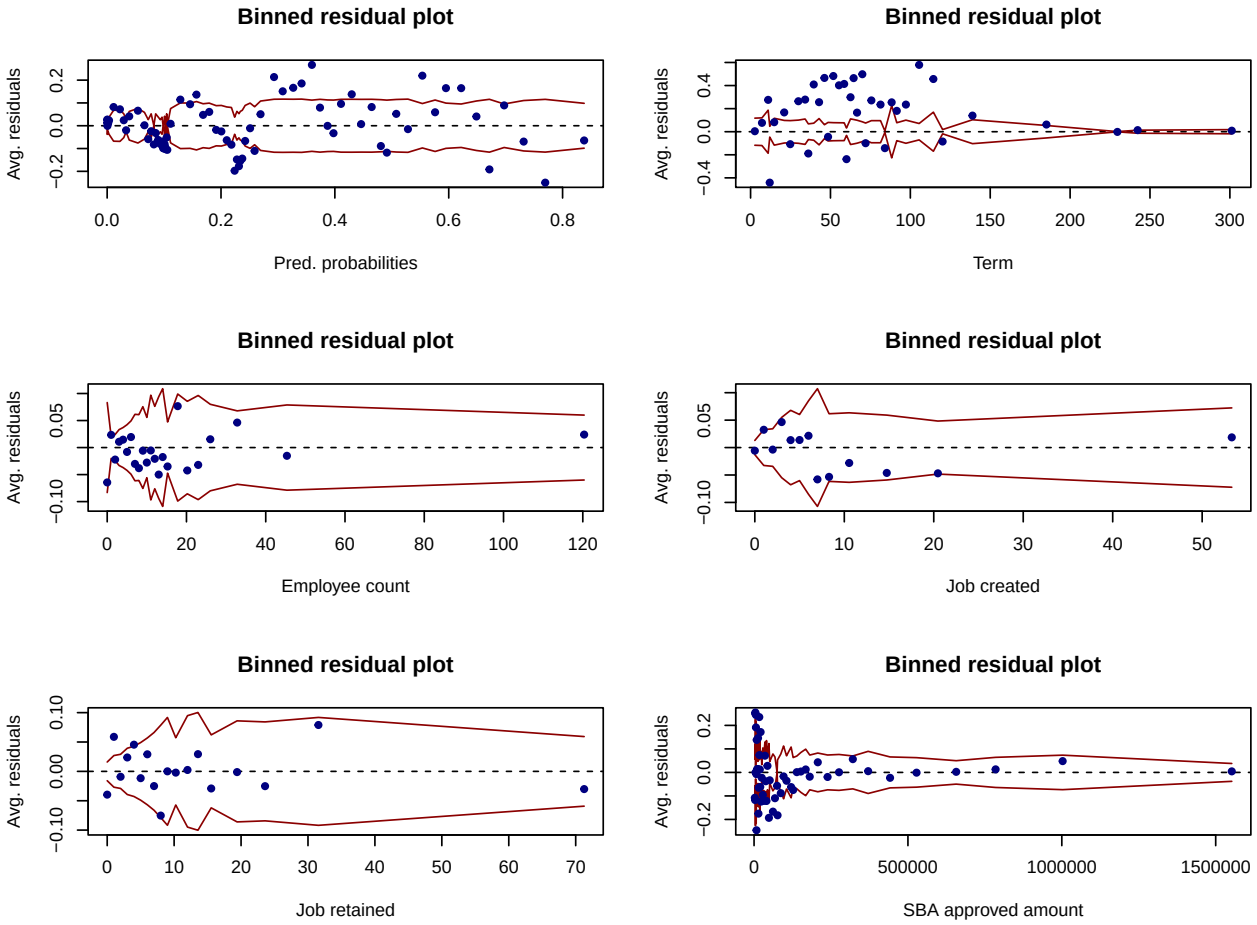


Figure 6. Cook's Distance for the Multiple Linear Regression Model

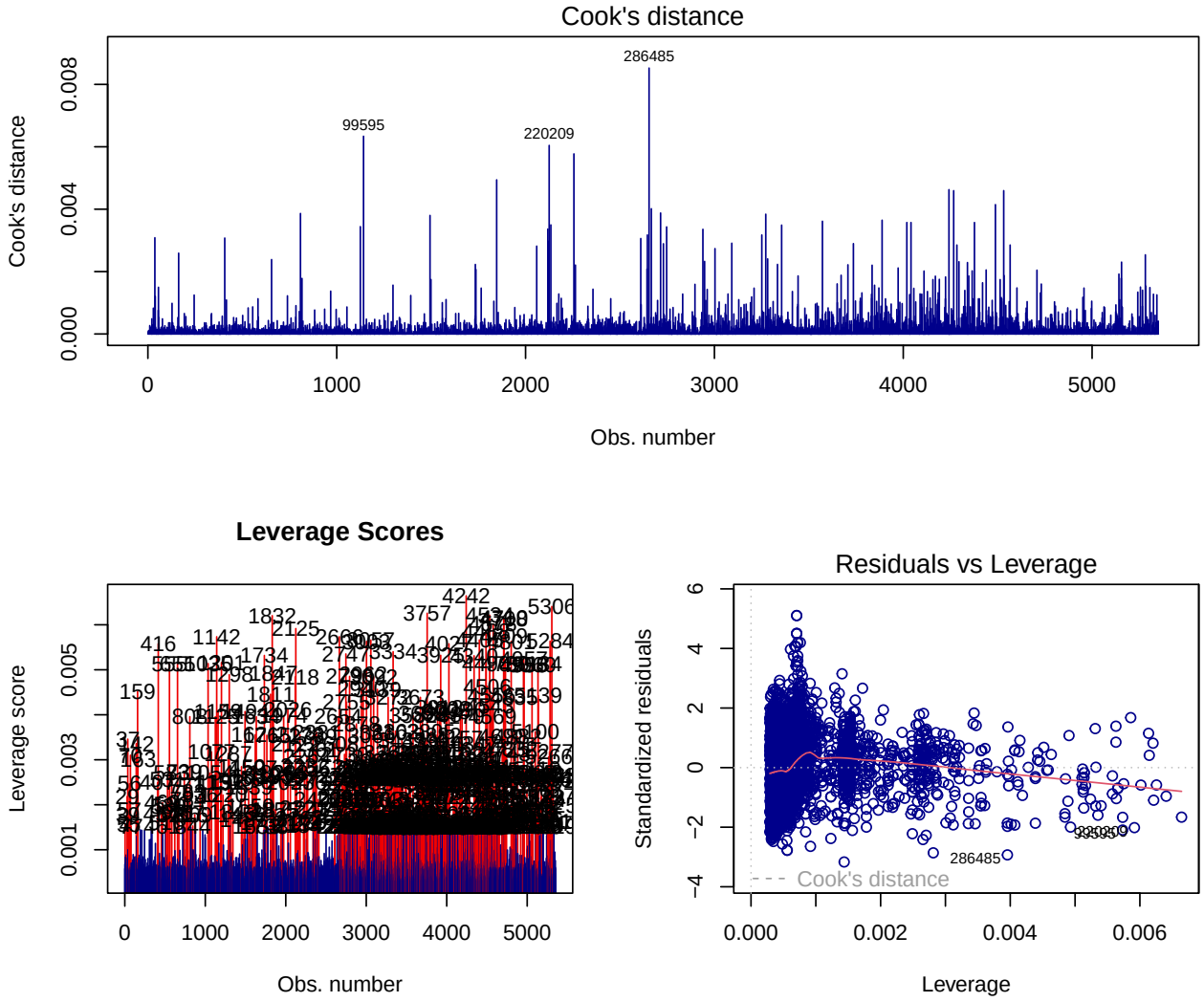
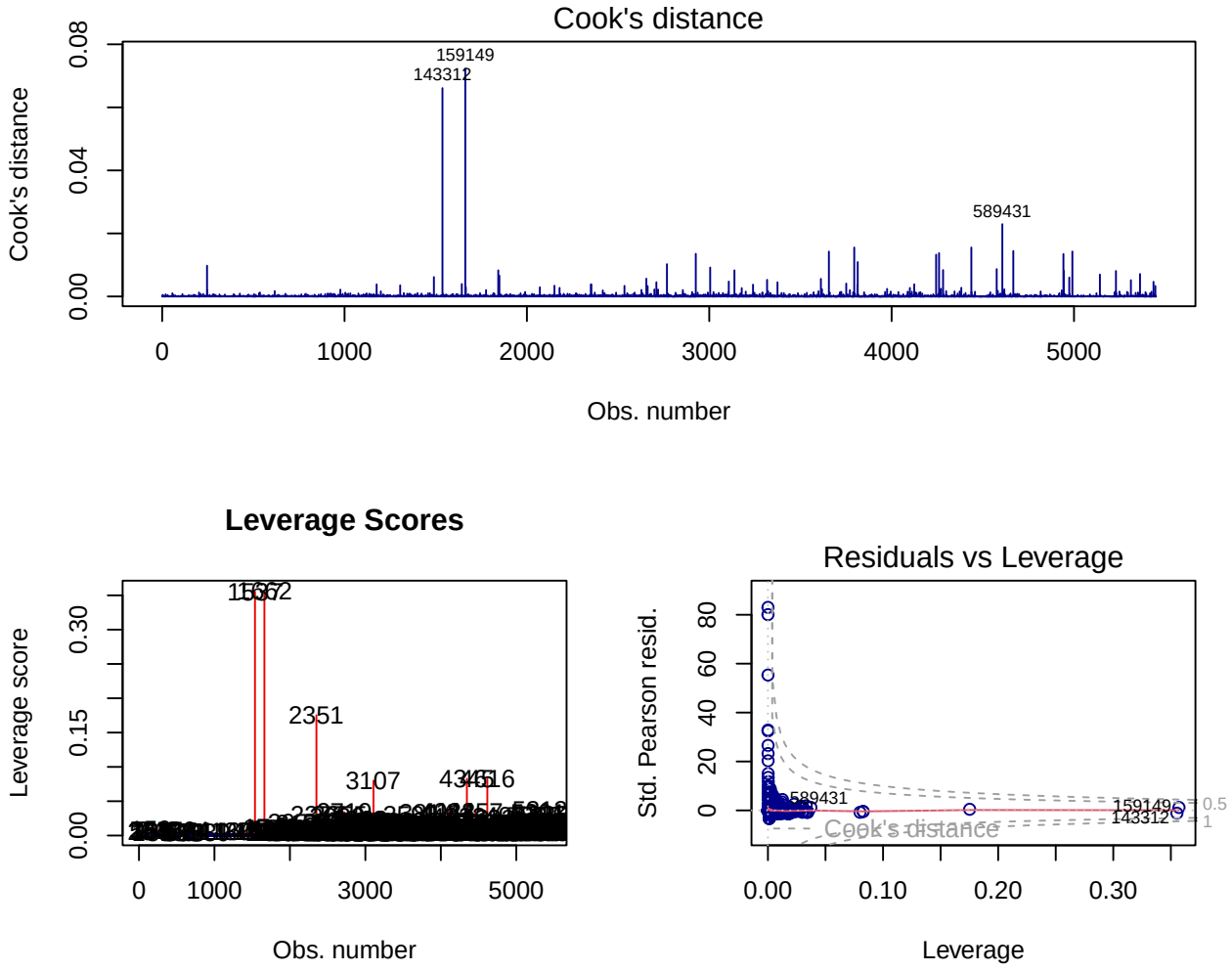


Figure 7. Cook's Distance for the Logistic Regression Model



References

Glennon, D., & Nigro, P. (2005). Measuring the default risk of small business loans: A survival analysis approach. *Journal of Money, Credit and Banking*, *37*(5), 923–947.

Li, M., Mickel, A.; & Taylor, S. (2018). "Should this loan be approved or denied?": A large dataset with class assignment guidelines. *Journal of Statistics Education*, *26*(1), 55-66. https://doi.org/10.1080/10691898.2018.1434342

Ou, C. (2005). *Banking consolidation and small business lending: A review of recent research working papers.* Office of Advocacy, U.S. Small Business Administration.