36-402 DA Exam 2

Clara Ye (zixuany)

April 29, 2022

Introduction

The Small Business Administration is interested in knowing how to efficiently advertise loan programs and distribute loans to create more jobs. To support their decision-making, we investigate two questions regarding the relationship among number of jobs created, loan amount, and types of businesses. The first question is whether there is diminishing returns for increasing loan amount, that is, whether the relationship between number of jobs created and loan amount is linear. If there is diminishing returns, then distributing several small loans would more effectively create jobs than making a single big loan. The second research question is to find the types of businesses that can create the most jobs per dollar loaned. Knowing this information would allow the SBA to target loan programs to those businesses more specifically.

We found that the relationship between the number of jobs created and loan amount is indeed non-linear, such that increasing loan amount has diminishing returns beyond around 1×10^6 USD. In addition, after controlling for loan amount and several other industry variables, we found that the industries with the highest expected number of jobs created are accommodation and food, and mining and gas (2).

Exploratory Data Analysis

Our response variable is Jobs Created, which is the number of jobs the business expects to create with the loan money. Its mean is 1.388 and its median is 0, and as displayed in Figure 1, it is heavily positively skewed (2). Other key variables involved in the research questions are: Loan Amount, the total amount of money loaned in dollars; Industry Category, the category of industry the business is in; Franchise, whether the business is a

franchise (Y) or independent (N); Area, whether the business is in an urban area (Urban) or a rural area (Rural); New Business, whether the business is new (New) or already existed (Existing) (1).

The distribution of Loan Amount is shown in Figure 1, and we can see that it is also heavily positively skewed, with mean 132,000 and median 539,00. The distribution of Industry Category is displayed in Figure 2. Retail and trade is the most common category ($n = 2769$), followed by professional services ($n = 1944$) and other ($n = 1754$). Public administration, utilities, and mining and gas are the least represented categories. The distributions of the other categorical variables are summarized in Figure 3. The majority of businesses are independent, in urban areas, and already existed.

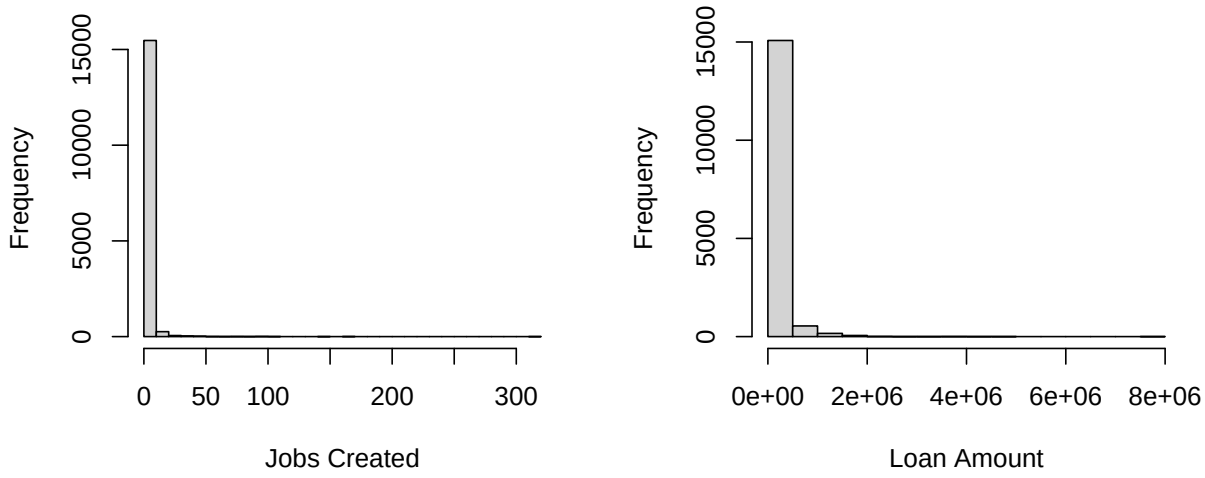

Figure 1: Histograms of Continuous Key Variables

Figure 4 shows the relationship between Jobs Created and Loan Amount, where the plot on the right is the same plot as the one on the left zoomed to where most of the data points are clustered. There is a weak, positive correlation between Jobs Created and Loan Amount ($r = 0.122, p < .001$), but it is difficult to see from the scatterplot whether the relationship is linear (3). Therefore, we may need to perform some tests to determine if a linear model or a non-linear model is more desirable.

The summary statistics of Jobs Created in each Industry Category is presented in Table 1. The industry category with the highest mean Jobs Created is accommodation and food, and the gap with the second highest mean is remarkable. Agriculture and public administration are the industry categories with the lowest mean Jobs Created, and the gap with the third lowest mean is also quite considerable. They are hence good candidates for the industry categories with a significant association with the number of jobs created (4).

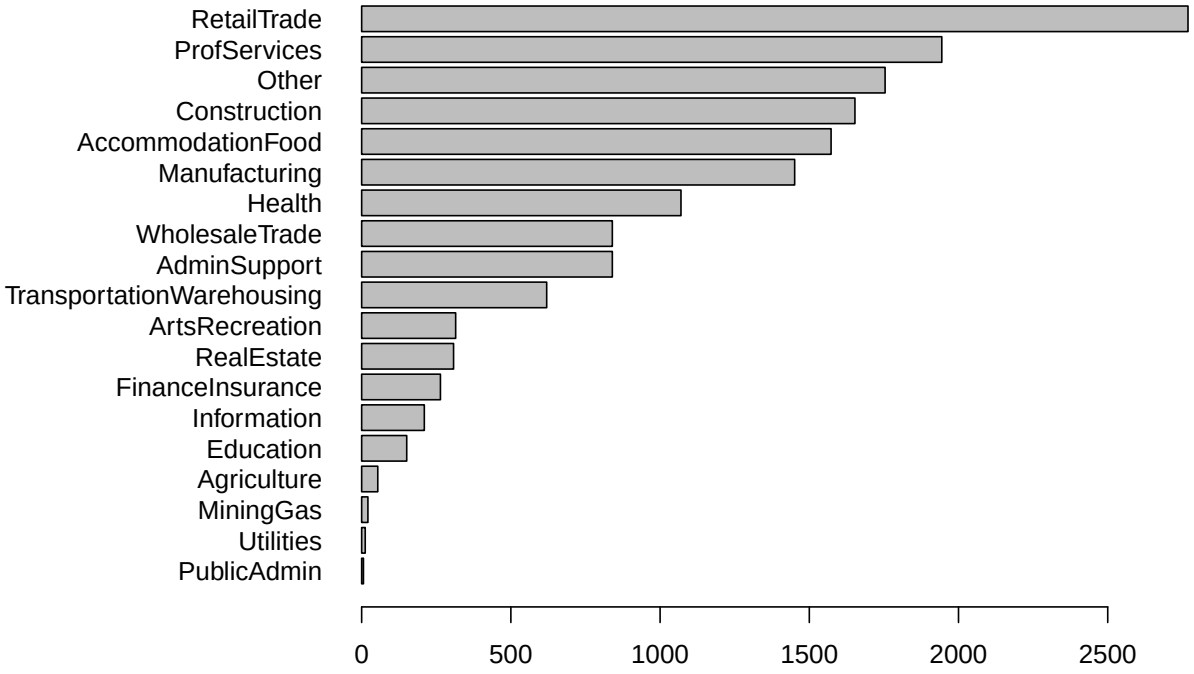

Figure 2: Bar Plot of Frequencies of Each Industry Category

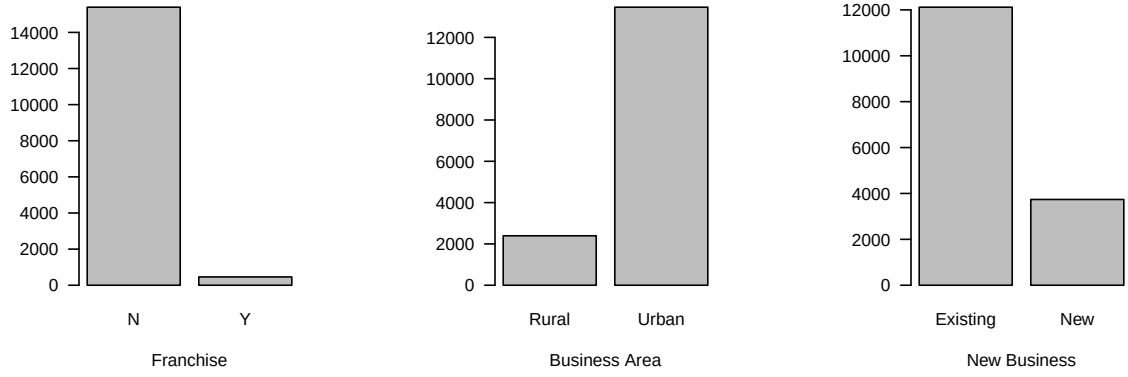

Figure 3: Bar Plots of Frequencies of Other Categorical Key Variables

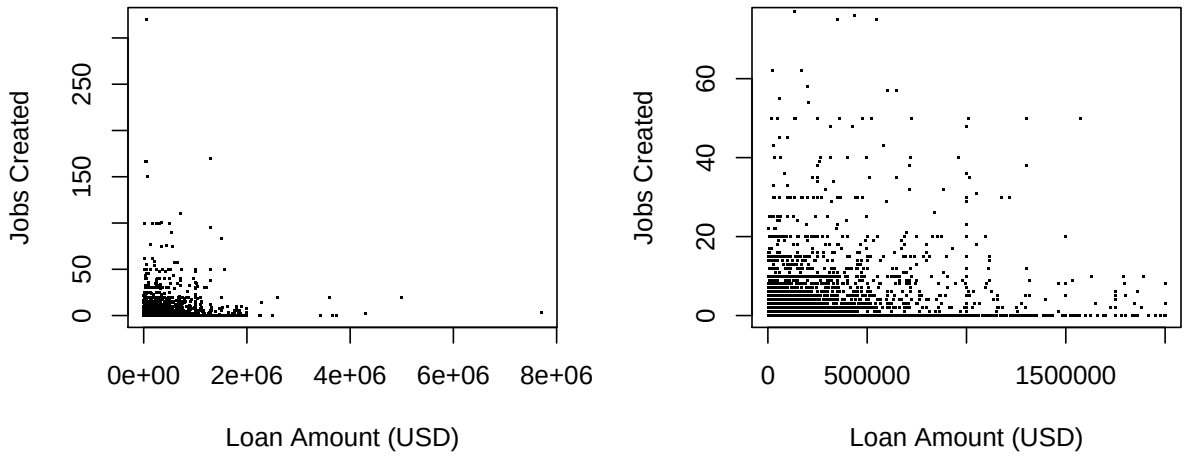

Figure 4: Scatterplot of Number of Jobs Created and Loan Amount

Table 1: Summary Statistics of Number of Jobs Created for Each Industry

Industry	Min.	1st Qu.	Median	Mean
AccommodationFood	0	0.0	3.100	95
MiningGas	0	0.0	1.950	12
AdminSupport	0	0.0	1.870	150
ArtsRecreation	0	0.0	1.810	50
Health	0	0.0	1.760	320
Manufacturing	0	0.0	1.570	170
ProfServices	0	0.0	1.230	167
RealEstate	0	0.0	1.220	100
Education	0	0.0	1.050	18
Construction	0	0.0	1.010	110
RetailTrade	0	0.0	1.010	83
TransportationWarehousing	0	0.0	0.998	90
WholesaleTrade	0	0.0	0.994	40
Information	0	0.0	0.905	19
Utilities	0	0.5	0.833	2
Other	0	0.0	0.826	57
FinanceInsurance	0	0.0	0.803	22
PublicAdmin	0	0.0	0.167	1
Agriculture	0	0.0	0.130	2

Table 2 shows the summary statistics of Jobs Created for Franchise, Business Area, and New Business. Franchise businesses have higher mean Jobs Created than independent businesses. Businesses in urban and rural areas have similar mean Jobs Created. New businesses have higher mean Jobs Created than existing businesses.

Table 2: Summary Statistics of Number of Jobs Created for Other Categorical Variables

Variable	Value	Min.	Median	Mean	Max.
Franchise	N	0	0	1.32	320
Franchise	Y	0	0	3.67	101
UrbanRural	Rural	0	0	1.48	90
UrbanRural	Urban	0	0	1.37	320
NewBusiness	Existing	0	0	1.06	110
NewBusiness	New	0	0	2.44	320

Modeling & Diagnostics

To investigate the relationship between business types and job creation, we first fit a linear model that predicts Jobs Created with Loan Amount, Area, New Business, Industry Category, and Franchise, and so the response distribution would be Gaussian (1). We would refer to it as Model 1 in the rest of this report. To examine whether the relationship between Jobs Created and Loan Amount is linear, we fit an additive model where Loan Amount enters non-linearly as splines with at most 5 degrees of freedom, chosen automatically by `mgcv`, and all other variables remain the same as in Model 1 (2). We refer to this model as Model 2.

We investigate the residuals to check our model assumptions before proceeding to further analysis. Figure 5 displays the diagnostics for the two models. In the residuals vs. fitted plots for both models, there is a trend where the residuals are more positive for smaller fitted values and more negative for larger fitted values. For Model 1, the residuals also seem to have greater variance for smaller fitted values and smaller variance for larger fitted values. The variance is mostly constant across the fitted values for Model 2. In the normal Q-Q plots for both models, the sample quantiles deviate considerably from the theoretical quantiles, so the residuals do not seem to follow the normal distribution.

Therefore, the assumption that the residuals are independent and identically distributed with $N(0, \sigma^2)$ is clearly violated, and so any inference under this assumption, such as the construction of confidence intervals, can be unreliable (3).

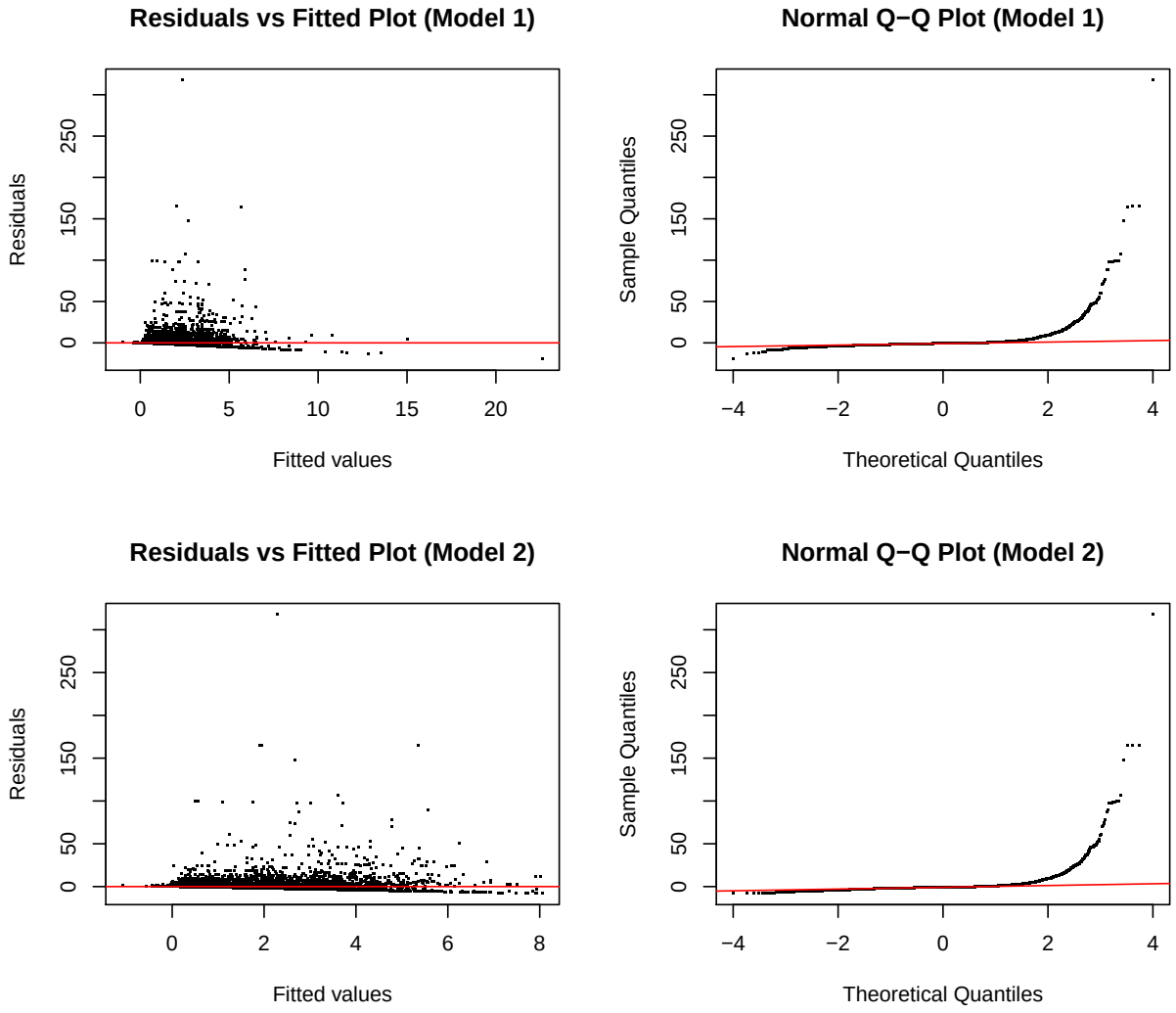

Figure 5: Diagnostic Plots for Model 1 and Model 2

To investigate on the first research question, we perform an F-test to examine whether making the term with Loan Amount non-linear is necessary. Since the smoothing parameter was chosen automatically by `mgcv`, we need to bootstrap the distribution of the test statistic under the null hypothesis and obtain the p-value using this distribution (4). To do this, we use a parametric bootstrap in which for each predictor X_i, we draw $Y^*_{b,i}$ from $N(\mu_i, \sigma^2)$ where μ_i is the predicted value by Model 1 and σ is the standard deviation of the residuals by Model 1. We refit Model 1 and Model 2 using the bootstrapped data and

perform the F-test on them. Finally, the bootstrap p-value is computed as the proportion of the bootstrapped F-statistics that are greater than the observed F-statistic.

We then examine the second research question by constructing confidence intervals for the coefficients of Industry Category. Again, since the smoothing parameter was chosen automatically by `mgcv`, we need to bootstrap the confidence intervals. According to the diagnostics, all the model assumptions are violated, including the assumption that the residuals are independent from the predictors. Therefore, we need to construct pivotal confidence intervals by sampling cases (5).

Results

The observed F-statistic is 19.9 and the bootstrap p-value is given by $P(F \geq F_{\text{obs}} \text{when } H_0 \text{is true})$, which is 0. Therefore, we reject the null hypothesis that Model 1 is correct, and we conclude that Model 2, which includes Loan Amount as a non-linear term, is the better model (1). We would use Model 2 for the rest of our analysis.

Figure 6 shows the partial response plot for Loan Amount, and we can see that the curve indeed has considerable curvature. When Loan Amount is smaller than 1×10^6, there is a positive relationship that is quite steep and approximately linear. Beyond this threshold, the relationship first becomes slightly negative, and then goes up slightly again, and finally falls back into a declining trend.

The 95% pivotal confidence intervals for the association between Industry Category and Jobs Created are presented in Table 3 (2). Note that the reference level is accommodation and food. We can see that most of the confidence intervals do not cover 0 and have negative upper bound, suggesting that businesses in most of the industry categories expect to create fewer jobs than those in accommodation and food (3). The only exception is mining and gas, whose 95% confidence interval covers 0, suggesting that the number of jobs created by these businesses are no different from that of the businesses in accommodation and food. Therefore, accommodation and food as well as gasoline and gas appear to have the highest number of jobs created; on the other hand, public administration appears to have the lowest number of jobs created (3).

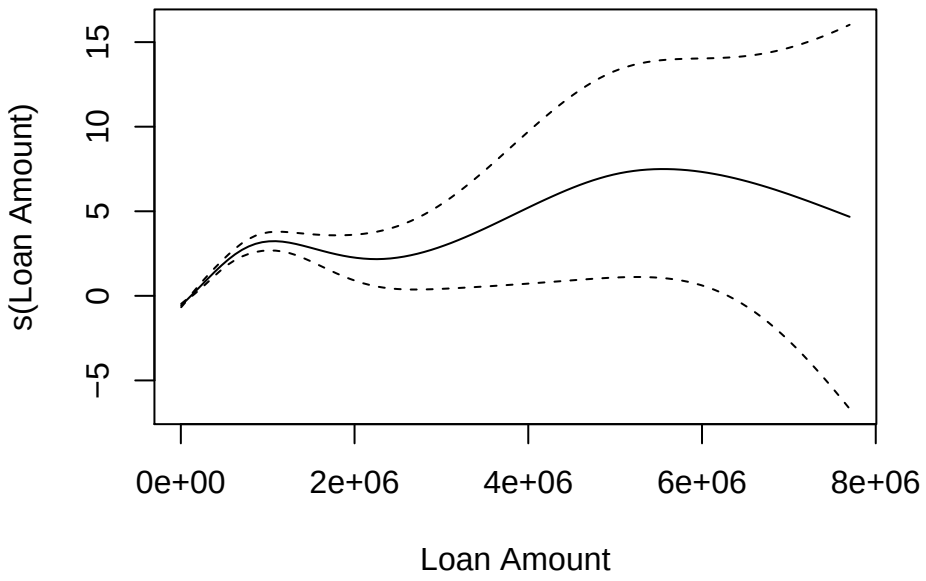

Figure 6: Partial Response Plot for Loan Amount

Table 3: 95% Pivotal Confidence Intervals for the Association between Industry and Jobs Created

	2.5 %	97.5 %
AdminSupport	-1.61	-0.199
Agriculture	-2.70	-1.890
ArtsRecreation	-1.88	-0.521
Construction	-2.02	-1.250
Education	-2.33	-1.290
FinanceInsurance	-2.24	-1.360
Health	-1.83	-0.479
Information	-2.31	-1.420
Manufacturing	-1.91	-1.030
MiningGas	-2.72	0.399
Other	-2.14	-1.420
ProfServices	-1.91	-0.985
PublicAdmin	-3.76	-2.670

	2.5 %	97.5 %
RealEstate	-2.45	-0.574
RetailTrade	-2.11	-1.380
TransportationWarehousing	-2.14	-1.170
Utilities	-2.78	-1.250
WholesaleTrade	-2.32	-1.500

Conclusions

In our analysis, we first fitted two models to evaluate whether the relationship between Jobs Created and Loan Amount is linear and how Industry Category associates Jobs Created. We found that the model where Loan Amount is included as a non-linear term performs significantly better than the model where Loan Amount is included as a linear term. Hence we conclude that the relationship between the number of jobs created and loan amount is not linear (1). In fact, diminishing returns occurs starting from a loan amount of 1×10^6, which is about the median Loan Amount of 5.25×10^4. Regarding the second research question, after controlling for Franchise, Business Area, and New Business, the industry categories with the highest expected number of jobs created are accommodation and food and mining and gas, and the industry category with the lowest expected number of jobs created is public administration. Therefore, targeting loan programs at businesses in accommodation and food or mining and gas is most likely to create jobs (1).

Our analysis has several limitations. First, there are likely other confounding variables that we did not control for, such as the number of employees the business had before receiving the loan. Intuitively, this could be a confounding variable since if the employee population is already large, then it is unlikely that the business would open up even more job positions after receiving the loan. We also did not examine the variable Default, which is whether the business failed to pay back the loan and which is presumably an important factor when the SBA makes loan decisions (2). If the business is unlikely to pay back the loan, then the SBA would probably not grant the loan regardless of the expected number of jobs to create, since the loan money would likely become a sunk cost. Finally, the point beyond which diminishing returns of increasing loan amount occurs was only estimated by visual inspection of the partial response plot. Using quantitative estimation techniques to determine a more accurate threshold may better help the SBA allocate the loan funds.